MACHINE LEARNING APPROACHES TO HUMAN MOVEMENT ANALYSIS
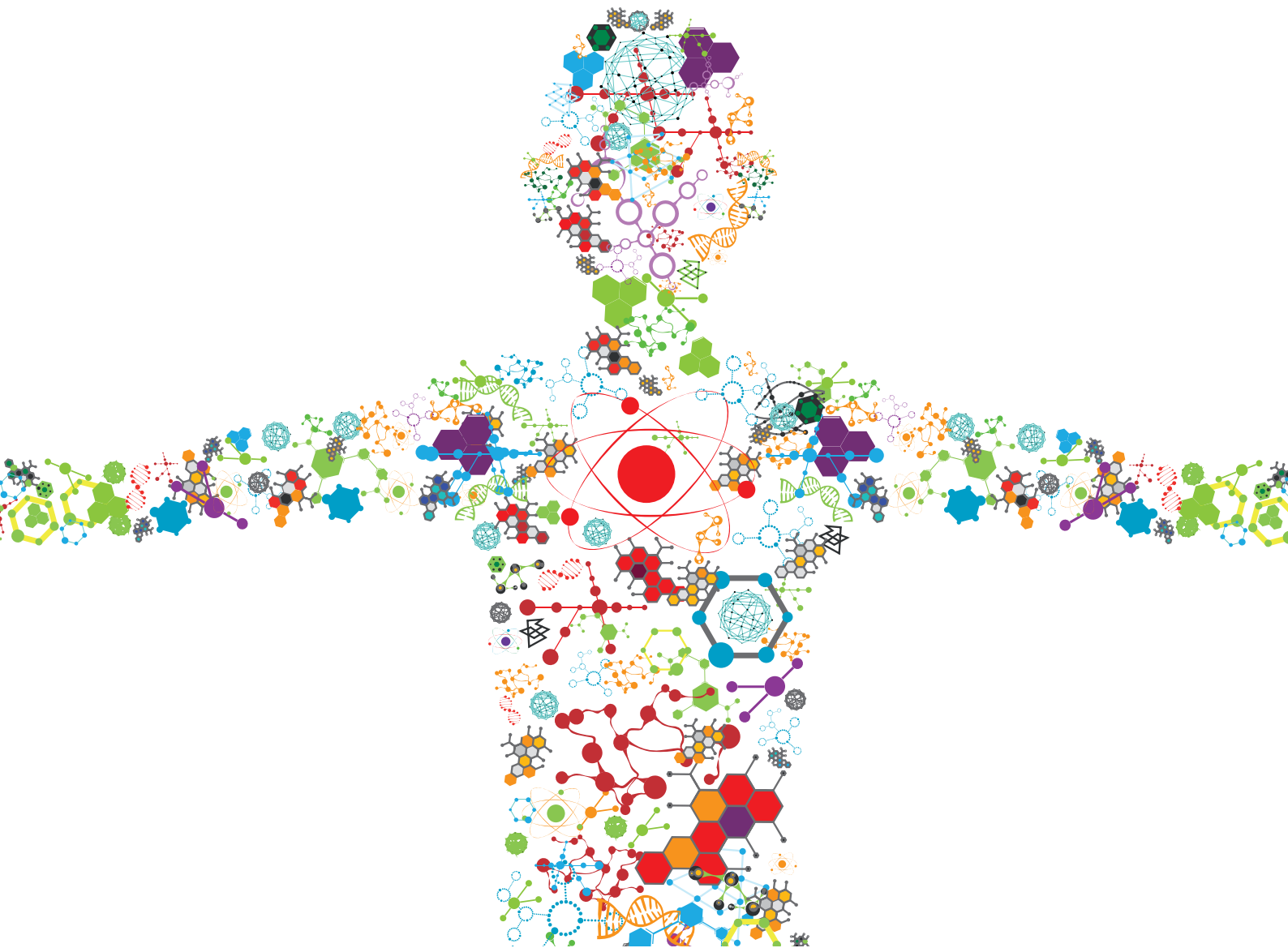
EDITED BY: Matteo Zago, Peter A. Federolf and Ana Francisca Rozin Kleiner

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MACHINE LEARNING APPROACHES TO HUMAN MOVEMENT ANALYSIS

Topic Editors:
Matteo Zago, Polytechnic of Milan, Italy
Peter A. Federolf, University of Innsbruck, Austria
Ana Francisca Rozin Kleiner, Federal University of São Carlos, Brazil

Table of Contents

Editorial: Machine Learning Approaches to Human Movement Analysis

Matteo Zago[1]*, Ana Francisca Rozin Kleiner[2] and Peter Andreas Federolf[3]

[1] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy,* [2] *MOVA4all, São Paulo, Brazil,*
[3] *Department of Sport Science, University of Innsbruck, Innsbruck, Austria*

Editorial on the Research Topic

Machine Learning Approaches to Human Movement Analysis

EMBRACING HUMAN MOVEMENT COMPLEXITY

Back in the mid-2000s, when Facebook was new and smartphones had still not become a major part of our everyday lives, researchers started to explore the use of Machine Learning in biomechanics. The road ahead appeared uncertain and people asked whether it was a new dawn or false hope? (Bartlett, 2006). Fifteen years later, we are in the middle of the era of data science, witnessing an unprecedented flourishing of techniques and applications. When large amounts of information can be collected and analyzed, the appeal and "unreasonable effectiveness of data" (Halevy et al., 2009) has found fertile ground in the study of complex biological and physical systems, human movement science among them.

The way we humans move, and the underlying cognitive control involved in this process is inherently complex, dynamic, multidimensional, and highly non-linear (Phinyomark et al., 2018). Machine Learning approaches enable us to embrace this complexity, working on three complementary tasks: predictive modeling, classification, and dimensionality reduction. With contributions from the five continents, the collection of papers in this Research Topic represents insightful viewpoints on the current landscape and potential new trends on the horizon.

ESTIMATION OF KINETICS AND KINEMATICS FROM WEARABLE SENSORS

A model is a summary of the best knowledge of a system at the time it is investigated, capturing essential aspects that are critical in answering the question at hand. Predicting modeling maps input data to a given output and can be used to anticipate future events with confidence. It is not surprising that the availability of large datasets obtained from wearable sensors has promoted considerable advancements in the estimation of quantities that traditionally required expensive laboratory setups, such as ground reaction forces and derived variables. This is one of the main technological trends in recent research. The study by Mundt et al. discusses motion capture systems and how they retrieve kinetic (and kinematics) data with less expert knowledge and without expensive equipment. As they state, motion capture systems are going to "increase the availability of motion analysis to a wider range of people." In other words, access to such systems enables a move toward pervasive healthcare systems (Zhou et al., 2020).

Exploiting the power and versatility of artificial (deep) neural networks as universal function approximators, the contributions by Stetter et al. and Derie et al. estimated external knee moments and vertical loading during various locomotion tasks, including running and using a minimal set of IMUs. Dorschky et al. worked on the same topic by augmenting a measured inertial sensor dataset with simulated data to demonstrate how to efficiently estimate sagittal plane angles, joint moments, and ground reaction forces. Complimentary work by De Brabandere et al. shows how the kinetics of the hip and knee can be estimated using smartphone embedded sensors.

ARTIFICIAL INTELLIGENCE ON THE MOVE

The automatic classification of athletic tasks based on motion data gathered in real-world conditions with inertial sensors is another expanding area of investigation, as evidenced in papers by Benson et al., Clouthier et al., and Preatoni et al. They use neural networks together with Support-Vector Machines and k-Nearest Neighbors algorithms. Pattern-recognition was also made possible by a combination of unsupervised Principal Component Analysis (PCA) and Gaussian Mixture Model or Linear Discriminant Analysis. Another contribution by Remedios et al. describes an alternative way to objectively identify movement phenotypes without the need to *a-priori* prescribe movement features. Ross et al. discriminate élite from novice athletes, devising inertial features that capture motion details related to competition level. The recognition of movement patterns was also a key focus in the study by Suda et al., which used ground reaction forces to recognize foot-ankle movement strategies in long-distance runners.

COMPUTER VISION

Computer Vision represents a parallel trend, involving a combination of deep neural networks and simpler classification algorithms. Guerra et al. propose an interesting application of automatic pose recognition (also used in Zago et al. for gait analysis) and classification to trigger an alarm in frail individuals. The study by Gregori et al. focused on rehabilitation by developing a deep-learning method to automatically evaluate grasping actions in people with upper limb prostheses. Background segmentation and shape classification allowed Monezi et al. to automatically detect the three-dimensional location of multiple players in a basketball court.

A STEP FORWARD

Gait analysis is undoubtedly a collector of the data revolution. Instrumented gait assessment is routinely used to evaluate an individual's quality of life, morbidity, and/or mortality. Here, data science is a powerful complement to traditional approaches when handling large, heterogeneous, and sometimes noisy data sets (Ferber et al., 2016). The integration of machine learning with biomechanics not only simplifies the assessment of several interdependent parameters (Khera and Kumar, 2020) but also

provides the opportunity for automated and unbiased analysis (Arac, 2020). Rethwilm et al. gained insights on trunk lean control in patients with Cerebral Palsy, combining PCA to binary logistic regression. In their technical paper, Burdack et al. explained how data filtering and unsupervised data reduction impacted gait classification based on ground reaction force data. Principal Component analysis is also the framework of analysis explored by Promsri and Federolf, who crafted a methodological paper explaining how to gain information about the coordinative structure of complex whole-body movements during balancing. In addition, De Roeck et al. focus on lower limb kinematics during deep squatting and in the forward lunge, devising a statistical model that predicts lower limb kinetics therein.

Zaroug et al. propose a method to predict lower limb kinematic trajectories during walking using long short-term memory (LSTM) neural networks. LSTM was also combined with convolutional neural networks by Yu et al. to predict pre-impact fall for older people. Notably, this approach was also implemented in a microcontroller unit featuring a working device. The practical translation of these techniques constitutes a crucial step that has not yet been undertaken that requires further exploration in future studies.

CLINICAL APPLICATIONS

Many papers in the Research Topic propose research-grade applications, and the effective combination of technology and data science will be topical in the near future. The seed of this trend is already visible. For instance, statistical shape modeling supported by logistic regression has clinical applications in automating the identification of surgically-relevant landmarks, as demonstrated by Cerveri et al.. de Araújo et al. showed how hand resting tremors could be used in the diagnosis of Parkinson's Disease. For patients in a similar condition, Lebel et al. worked on the prediction of motor performance based on visible symptomatology. A crucial issue is discussed in the work by Chia et al., which developed a decision support system based on gait kinematics, anthropometric characteristics, and physical examination and trained a system to learn the recommendations formulated by clinicians.

WHAT'S NEXT

While carefully avoiding falling into a simplistic (and potentially dangerous) "idolatry of data," we believe that the road is paved for rapid and inevitable (re)volutions. Assisting human decisions is among the most impactful advancements that data science and human movement science together can provide to medicine in the next decade (Jones et al., 2018). First data science can benefit education by supporting junior clinicians and potentially, later on, assisting in diagnosis and prognosis. In this journey, data is a powerful ally, but there is a need for machine learning to provide transparency and justifications of predictions (Halilaj et al., 2018; Horst et al., 2019). A framework to interpret deep learning features and the "magic inside the black box" is essential and significant efforts are currently being made toward

creating explainable artificial intelligence (Côté-Allard et al.). Furthermore, as anticipated by many experts (Ferber et al., 2016; Halilaj et al., 2018), a cultural shift toward data sharing is necessary to achieve the required general validity (and constant upgradability) that will bring these systems into clinical practice.

The constant growth of computational power and wearable sensor miniaturization will also open pathways to pervasive real-time applications, exploiting the wealth of data available "out in the wild," from marker-less motion capture to exercise monitoring and training assistance. To date, artificial intelligence does not simply provide new tools to study human motion.

Rather, the way we study human motion is evolving thanks to artificial intelligence. By "following the data" (Halevy et al., 2009), we are pursuing unexplored and fascinating avenues of knowledge.

AUTHOR CONTRIBUTIONS

MZ, AK, and PF contributed equally to the writing of this editorial. All authors contributed to the article and approved the submitted version.

REFERENCES

Arac, A. (2020). Machine learning for 3D kinematic analysis of movements in neurorehabilitation. *Curr. Neurol. Neurosci. Rep.* 20:29. doi: 10.1007/s11910-020-01049-z

Bartlett, R. (2006). Artificial intelligence in sports biomechanics: new dawn or false hope? *J. Sports Sci. Med.* 5, 474–479.

Ferber, R., Osis, S. T., Hicks, J. L., and Delp, S. L. (2016). Gait biomechanics in the era of data science. *J. Biomech.* 49, 3759–3761. doi: 10.1016/j.jbiomech.2016.10.033

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intell. Syst.* 24, 8–12. doi: 10.1109/MIS.2009.36

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. (2018). Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J. Biomech.* 81, 1–11. doi: 10.1016/j.jbiomech.2018.09.009

Horst, F., Slijepcevic, D., Lapuschkin, S., Raberger, A. M., Zeppelzauer, M., Samek, W., et al. (2019). On the understanding and interpretation of machine learning predictions in clinical gait analysis using explainable artificial intelligence. *arXiv.* 1912.0 7737.

Jones, L. D., Golan, D., Hanna, S. A., and Ramachandran, M. (2018). Artificial intelligence, machine learning and the evolution of healthcare:

a bright future or cause for concern? *Bone Joint Res.* 7, 223–225. doi: 10.1302/2046-3758.73.BJR-2017-0147.R1

Khera, P., and Kumar, N. (2020). Role of machine learning in gait analysis: a review. *J. Med. Eng. Technol.* 44, 441–467. doi: 10.1080/03091902.2020.1822940

Phinyomark, A., Petri, G., Ibáñez-Marcelo, E., Osis, S. T., and Ferber, R. (2018). Analysis of big data in gait biomechanics: current trends and future directions. *J. Med. Biol. Eng.* 38, 244–260. doi: 10.1007/s40846-017-0297-2

Zhou, L., Fischer, E., Tunca, C., Brahms, C. M., Ersoy, C., Granacher, U., et al. (2020). How we found our imu: guidelines to IMU selection and a comparison of seven IMUs for pervasive healthcare applications. *Sensors* 20, 1–28. doi: 10.3390/s20154090

On the Visuomotor Behavior of Amputees and Able-Bodied People During Grasping

Valentina Gregori[1,2]*, Matteo Cognolato[3,4], Gianluca Saetta[5], Manfredo Atzori[3], The MeganePro Consortium and Arjan Gijsberts[2]

[1] Department of Computer, Control, and Management Engineering, University of Rome La Sapienza, Rome, Italy, [2] VANDAL Laboratory, Istituto Italiano di Tecnologia, Genoa, Italy, [3] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland, [4] Rehabilitation Engineering Laboratory, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland, [5] Department of Neurology, University Hospital of Zurich, Zurich, Switzerland

Visual attention is often predictive for future actions in humans. In manipulation tasks, the eyes tend to fixate an object of interest even before the reach-to-grasp is initiated. Some recent studies have proposed to exploit this anticipatory gaze behavior to improve the control of dexterous upper limb prostheses. This requires a detailed understanding of visuomotor coordination to determine in which temporal window gaze may provide helpful information. In this paper, we verify and quantify the gaze and motor behavior of 14 transradial amputees who were asked to grasp and manipulate common household objects with their missing limb. For comparison, we also include data from 30 able-bodied subjects who executed the same protocol with their right arm. The dataset contains gaze, first person video, angular velocities of the head, and electromyography and accelerometry of the forearm. To analyze the large amount of video, we developed a procedure based on recent deep learning methods to automatically detect and segment all objects of interest. This allowed us to accurately determine the pixel distances between the gaze point, the target object, and the limb in each individual frame. Our analysis shows a clear coordination between the eyes and the limb in the reach-to-grasp phase, confirming that both intact and amputated subjects precede the grasp with their eyes by more than 500 ms. Furthermore, we note that the gaze behavior of amputees was remarkably similar to that of the able-bodied control group, despite their inability to physically manipulate the objects.

Keywords: visuomotor strategy, eye-hand coordination, upper-limb amputees, object segmentation, phantom limb movements, object tracking

1. INTRODUCTION

Humans interact continuously with objects in activities of daily living (ADLs). Vision and gaze play an important role during these interactions, not only to guide the activity itself but also in the initial planning phase. Gaze is thus said to be anticipatory and can be used to understand an individual's intentions even before they manifest themselves in the motor domain. Several studies have attempted to explore this proactivity to help disabled people, such as in a robot assistant scenario (Admoni and Srinivasa, 2016; Koochaki and Najafizadeh, 2018; Saran et al., 2018). Another compelling use-case is the control of dexterous upper-limb prostheses (Castellini and Sandini, 2006; Markovic et al., 2014, 2015; Gigli et al., 2018), where deciphering the grasp intent from myoelectric activations alone can be challenging. The integration of gaze and vision as contextual information could be helpful

especially during the initial transient phase of a movement. Executing this fusion successfully requires however a precise understanding of eye-hand coordination.

Gaze behavior has been studied extensively over the last decades. Early studies typically involved constrained settings, for instance by fixating the chin to avoid head movements or by limiting the field of view to a monitor (see Tatler et al., 2011, and references therein). Obviously, these findings may not be representative for unconstrained settings where free movement of the body is allowed to ensure natural behavior (Tatler et al., 2011; Tatler, 2014). Such unconstrained experiments became possible with the introduction of wearable eye-tracking devices that allowed the user to move freely in the environment (Land, 2006). Subsequently, studies on visuomotor coordination have confirmed that also in this setting actions are typically preceded by a visual fixation on the involved objects. This was verified during a block-copying task (Smeets et al., 1996; Pelz et al., 2001), while drinking from various objects (Belardinelli et al., 2016), during an object displacement task (Belardinelli et al., 2016; Lavoie et al., 2018), during pick-and-place of a bar (Johansson et al., 2001), and when grasping (Brouwer et al., 2009). Similar goal-oriented gaze strategies were also reported during ADLs, such as tea-making and sandwich-making (Land and Hayhoe, 2001), walking (Patla and Vickers, 2003), driving (Land and Lee, 1994), and sports (Land and McLeod, 2000; Hayhoe et al., 2012). Although all studies confirm the anticipatory nature of gaze, they do not always agree on the exact timing of the motor execution after the first visual fixation, for instance when the hand reaches the object. These discrepancies can probably be explained by differences in experimental setting (Smeets et al., 1996; Pelz et al., 2001), variability due to a small number of subjects, or difficulty in accurately analyzing a large number of trials.

Only a few studies have investigated the gaze behavior of amputees. In a small case study, Sobuh et al. (2014) observed that the amputated participants did not use gaze to proactively plan subsequent actions in a task. Instead, they tend to switch their gaze more often between the object and the prosthetic hand to visually monitor its proper functioning (Bouwsema et al., 2012; Hebert et al., 2019). This increased visual attention is most likely to compensate for the lack of tactile and proprioceptive feedback from their prostheses. A similar finding was also reported when able-bodied subjects were engaged in similar tasks using a prosthetic simulator (Blank et al., 2010; Sobuh et al., 2014; Parr et al., 2018, 2019). Almost all of these studies investigated this disruption in eye-hand coordination precisely for this reason, namely to measure the subject's proficiency in controlling the prosthesis. More visual attention to the hand area during reaching and manipulation is considered indicative of a lower level of skill and confidence in the control of the prosthesis. Conversely, it should therefore also be expected that gaze behavior will "normalize" with an increasing confidence in the control response of the prosthesis. Indeed, Chadwell et al.

(2016) noted that one participant who used a prosthesis daily showed more natural gaze behavior than another less experienced participant, while Sobuh et al. (2014) observed a shorter fixation on the hand area with increasing practice.

In the present study, we investigate eye-hand coordination during reaching and grasping to determine the window of opportunity in which gaze can provide useful information for intent recognition. We used the data of the recently acquired dataset, in which 15 transradial amputees were asked to try to grasp and manipulate various household objects to the best of their ability with their missing limb. In addition, it contains data from 30 able-bodied control subjects who performed the same grasps and manipulation tasks with their right arm. Throughout the exercise, gaze, and visual data were recorded via eye-tracking glasses, while the muscular activity of the arm was recorded via surface electromyography (sEMG) electrodes. Contrary to prior work, asking amputees to perform "movements without movement" (Raffin et al., 2012b) allows us to investigate to which extent the amputees' eye-hand coordination has changed as a result of the amputation, rather than due to difficulties controlling a prosthesis. Given the similarity of movements executed with the phantom limb compared with those executed with intact limb (Raffin et al., 2012a,b), we also expect the eye-hand coordination of movements involving the missing limb to be highly similar to those involving the intact limb. This "ideal" setting does not imply that the results are not relevant for the prosthetic setting; the disruption of gaze strategies is actually characterized by a markedly longer reaching phase, while still maintaining the majority of the fixations on the target object (Sobuh et al., 2014; Hebert et al., 2019). The window of opportunity in the prosthetic setting is therefore expected to be considerably longer than the one we identify here.

The total size of the dataset exceeds 70 h of video, which is far too large to be analyzed and annotated manually within reasonable time. However, quantifying the distances between gaze point, target object, and the forearm is of fundamental importance for the present study. We therefore employed state-of-the-art deep learning techniques to automatically detect and segment all objects of interest in all videos. This procedure consisted of an efficient method to collect representative training data and the subsequent finetuning of a pretrained object detector to this data. A beneficial side-effect of detecting object locations in the video is that we can reliably determine fixations even in the presence of head movements.

In the following, we describe the dataset and the methods employed in the analysis in section 2. In section 3, we then present the results of our analysis, which are discussed more thoroughly in section 4. Finally, we conclude and summarize the paper in section 5.

2. MATERIALS AND METHODS

To investigate the visuomotor behavior during manipulation actions we relied on a large, recently acquired dataset. In the following, we describe how the data were used in the context

---

Abbreviations: ADL, activity of daily living; sEMG, surface electromyography; IVT, Identification Velocity Threshold; FPN, Feature Pyramid Network; COCO, Common Objects in COntext; AP, average precision; IoU, Intersection over Union.

TABLE 1 | The characteristics of the amputated participants considered in the present study.

| ID | Age | Gender | Handedness | Amputation | | | | |
				Side	Cause	Years	Prosthesis	Limb [%]
101	52	M	Right	Right	Electrocution	2	Cosmetic	60–80
102	39	M	Right	Right	Electrocution	4	Cosmetic	60–80
103	63	M	Ambidextrous	Right	Trauma	3	Myoelectric	60–80
104	49	M	Right	Right	Trauma	18	Myoelectric	80–100
105	73	M	Right	Right	Trauma	6	Body-powered	40–60
106	70	M	Left	Left	Trauma	5	Body-powered	80–100
107	36	M	Right	Left	Trauma	7	Body-powered	20–40
108	35	M	Right	Right	Trauma	9	Myoelectric	0–20
109	65	M	Right	Left	Trauma	1	Cosmetic	80–100
110	38	M	Right	Left	Trauma	14	Myoelectric	20–40
111	38	M	Right	Right	Trauma	10	Myoelectric	40–60
112	33	F	Right	Left	Oncological	13	Cosmetic	60–80
113	28	M	Right	Left	Trauma	7	Myoelectric	40–60
115	36	F	Right	Left	Burn	8	Cosmetic	n/a

The table reports the ID of the subjects in the MeganePro dataset, their age, their gender, and their handedness. Among the clinical parameters we report the amputation side, its cause, the number of years since amputation, the type of prosthesis used, and the relative length of the residual limb with respect to the contralateral limb.
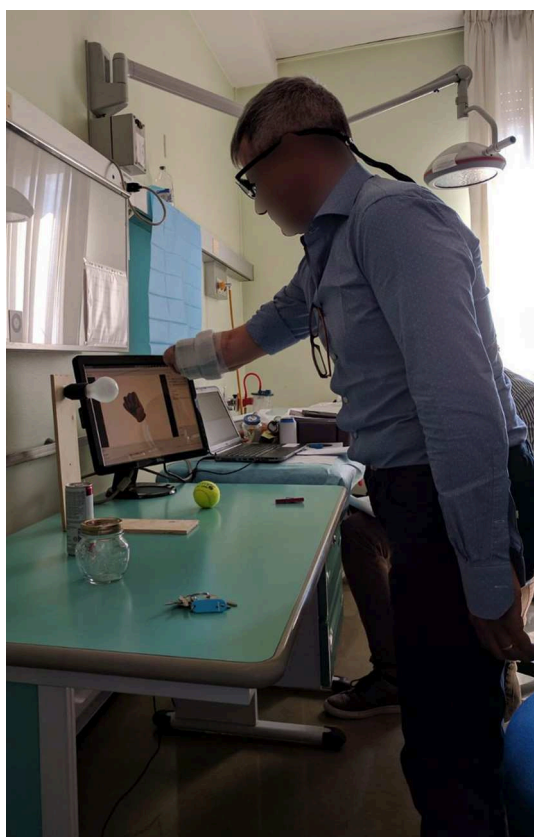
FIGURE 1 | An overview of the experimental setup.

of the present study. Due to the large amount of video data contained in this dataset, we devised a procedure to automatically detect and segment all objects of interest via deep learning.

This procedure is outlined and we formulate how the resulting segmentation masks were used to determine distances.

2.1. MeganePro Dataset

The MeganePro dataset was acquired with the aim of investigating the use of gaze and visual information to improve prosthetic control (Cognolato et al., 2019). It contains data of 15 transradial amputees [13 M, 2 F; age: (47.13 ± 14.16) years] and a frequency matched control group of 30 able-bodied subjects [27 M, 3 F; age: (46.63 ± 15.11) years] who performed grasps and manipulation tasks with a variety of household items. The gaze data for one of the amputated subjects was unreliable due to strabismus; this subject was therefore excluded from our analyses. The characteristics of the remaining amputated subjects is shown in **Table 1**, including information on the amputation and prosthetic use. All of them reported to experience phantom limb sensations, but only 12 had some voluntary control over the phantom limb.

During the experiment, the subjects wore a Tobii Pro Glasses 2 eye-tracker (Tobii AB, Sweden) to record the gaze behavior, first person video, and angular velocities of the head. These glasses sample gaze and gaze-related information at 100 Hz, while the video is recorded with a 1920 px × 1080 px resolution at 25 frames per second. On their forearm, they had 12 Delsys Trigno electrodes (Delsys Inc., USA) arranged in an array of eight equidistant electrodes at the height of the radiohumeral joint and four more electrodes in a second array 45 mm more distally. These electrodes record sEMG at 1926 Hz and contain an integrated three axes accelerometer that is sampled at 148 Hz. A picture showing the setup is shown in **Figure 1**.

The experiment consisted of repeatedly grasping or manipulating household items placed on a table in front of the subject. The pairing of grasps and objects was specifically chosen based (1) on the likelihood of their co-occurrence in

ADLs and (2) to attain as much as possible a many-to-many relationship between grasps and objects. In the first part of the experiment, subjects just had to perform a "static" grasp on the object without any manipulation, hold it for a few seconds and then return to a rest posture when instructed. The amputated subjects were asked to attempt to execute the action as naturally as possible "as if their missing limb were still there," rather than just imagining it, to elicit activation of the remaining muscles in their residual limb. Each of the grasps in **Table 2** and its three associated objects were first introduced via a video, after which the subjects were instructed vocally to grasp each object four times while seated and then another four times while standing. The order in which the objects appeared in each repetition block was randomized to avoid habituation. During the second part of the experiment, the same ten grasps were instead executed as part of a "functional" movement, as can be seen in **Table 3**. In this case, the movements were performed either seated or standing, depending on which position would seem more likely in real life.

Given the scope of the present paper, we only use sEMG from the second and seventh electrode, which were placed approximately on the extensor and flexor digitorum superficialis. Besides having relatively high activations, these electrodes also indicate roughly whether the hand was opening or closing. To aid visualization, both channels were rectified with a moving root mean square with a window-length of 29 ms (i.e., 57 samples) (Merletti, 1999). With respect to accelerometry, we note that the accelerations of all electrodes were highly correlated due to their positioning around the forearm. We therefore use accelerations only from the first electrode and normalize them with respect to the inertial frame of the initial position in each trial (Tundo et al., 2013).

2.2. Gaze Velocity

A common method to classify gaze events in fixations and saccades is based on the evaluation of the angular gaze velocity (Salvucci and Goldberg, 2000). Given two consecutive 3-dimensional gaze vectors $\boldsymbol{g}_{i-1}$ and $\boldsymbol{g}_i$, the angular difference between them can easily be calculated by means of their dot product (Duchowski, 2007)

$$\alpha_i = \arccos\left(\frac{\boldsymbol{g}_i \cdot \boldsymbol{g}_{i-1}}{\|\boldsymbol{g}_i\|\|\boldsymbol{g}_{i-1}\|}\right), \quad \forall i \in \{2, ..., N\} \ . \quad (1)$$

An approximation of the instantaneous gaze velocity at time t_i then follows as

$$v_i = \frac{\alpha_i}{t_i - t_{i-1}}, \quad \forall i \in \{2, ..., N\} \ . \quad (2)$$

Although the Tobii glasses provide a unit gaze vector for both eyes, we instead use the gaze point in world coordinates to estimate the common angle of the eyes. These world coordinates had fewer missing data and were slightly cleaner in practice due to onboard processing. They are however relative to the position of the scene camera rather than the eyes. Since this camera is located on top of the frame of the glasses, this may lead to some inaccuracy at small gaze distances. We therefore map the gaze

TABLE 2 | Overview of the static tasks.

Grasp	Object		
Medium wrap	Bottle	Door handle	Can
Lateral	Mug	Key	Pencilcase
Parallel extension	Plate	Book	Drawer
Tripod grasp	Bottle	Mug	Drawer
Power sphere	Ball	Bulb	Key
Precision disk	Jar	Bulb	Ball
Prismatic pinch	Clothespin	Key	Can
Index finger extension	Remote	Knife	Fork
Adducted thumb	Screwdriver	Remote	Wrench
Prismatic four finger	Knife	Fork	Wrench

For each row, the grasp and the associated objects are indicated. The subjects were asked to grasp each object with the given hand configuration while both seated and standing.

points to a coordinate system that is centered between the left and right pupils

$$\widehat{\boldsymbol{g}}_i = \boldsymbol{g}_i - \bar{\boldsymbol{p}}_i, \quad \forall i \in \{1, ..., N\} \ , \quad (3)$$

Grasp	Vocal instruction	Position	Category
Medium wrap	Drink from the **can**	Standing	Lifting
	Open and close the **door handle**		In place
Lateral	Turn the **key** in the lock	Standing	In place
	Open and close the **pencil case**		In place
Parallel extension	Lift the **plate**	Standing	Lifting
	Lift the **book**		Lifting
Tripod grasp	Open and close the cap of the **bottle**	Standing	In place
	Open and close the **drawer**		In place
Power sphere	Move the **ball** to the right and back	Standing	Displacement
	Move the **keys** forwards and backwards		Displacement
Precision disk	Open and close the lid of **jar**	Seated	In place
	Screw and unscrew the **light bulb**		In place
Prismatic pinch	Squeeze the **clothespin**	Seated	In place
	Move the **keys** forwards and backwards		Displacement
Index finger extension	Press a button on the **remote control**	Seated	In place
	Cut bread with the **knife**		In place
Adducted thumb	Turn the **screwdriver**	Seated	In place
	Move the **wrench** to the right and back		Displacement
Prismatic four finger	Move the **knife** forwards and backwards	Seated	Displacement
	Move the **fork** to the right and back		Displacement

The vocal instruction in English indicates the task that had to be performed for each object-grasp pair, while the position denotes whether the subject performed the task while seated or standing. The last column indicates the movement category as per the description in section 3.3.

where $\bar{p}_i$ is the average of the left and right pupil locations relative to the scene camera. To limit the impact of missing data for the pupils, we linearly interpolated gaps shorter than 0.075 s (Olsen and Matos, 2012).

2.3. Object Detection and Segmentation

To determine whether the subject is fixating the target object at any given time, we need a precise segmentation of this latter object throughout the exercise. Since the videos for each subject totaled around 90 min or 135 000 frames, this would be very time consuming to annotate manually. We therefore employed a deep learning algorithm to automatically segment and classify all instances of our objects of interest (see **Table 2**). Finetuning this algorithm to our data still required at least a few dozen segmentations per object class. Rather than creating these manually, we instead used a second deep learning algorithm to facilitate the creation of this dataset.

2.3.1. Creation of the Training Dataset

SiamMask is a recently proposed method for object tracking and semi-supervised video object segmentation (Wang et al., 2019). By marking just a bounding box around an object in one frame, this deep convolutional algorithm (1) segments the object from the background and (2) tracks it in the following frames in a video sequence. Although it may seem tempting to run this algorithm on an entire video annotating each object only at its

first occurrence, in practice the object tracking does not work reliably on such long time scales. We therefore used this method to amplify our manual annotations; with just a single bounding box annotation per object, we obtain 10 to 20 times as many binary segmentation masks for our training set.

For our approach, we embedded the official implementation of SiamMask[1] with a default ResNet-50 backend in a custom application. This software allows the user to select a frame in a video and to annotate several objects with their bounding box and their class identity. Based on this initialization, SiamMask processes the initial frame and subsequent frames one by one. At each frame, the output is presented to the user for validation, who can either accept or refuse the proposed segmentation. This procedure is shown schematically in **Figure 2**. In practice, we accepted sequences up to about 15 frames. Applying this procedure repeatedly, we processed in total 2,422 frames with 11,726 segmented object instances chosen from 15 subjects. To include as much variability as possible in our dataset, we captured the objects from different perspectives, with different backgrounds, and while partially occluded. Furthermore, besides the eighteen objects in **Table 2**, we also included segmentations for a "person" class, which is primarily used to detect the subject's own limb.

---

[1]https://github.com/foolwood/SiamMask

FIGURE 2 | The procedure to acquire the training set of segmentation masks. We first select an arbitrary frame from a video and annotate each object with its bounding box and object identity. This information is passed to SiamMask, which produces segmentation masks for this initial frame and the subsequent frames in the video sequence. At each frame, the user can choose whether or not to include the frame and its segmentations in the training set or to move to a new initial frame.

2.3.2. Training and Inference of Mask R-CNN

The data we acquired in this manner were used to train Mask R-CNN on our objects of interest. This method detects and segments all instances of the known objects in an image (He et al., 2017). Rather than training a model "from scratch," we bootstrapped from a model that was supplied with the implementation of Mask R-CNN by Massa and Girshick (2018). This model used a relatively standard ResNet-50-FPN backbone (Lin et al., 2017) and was pretrained on the COCO dataset (Lin et al., 2014), a large scale generic dataset for object detection, segmentation, and classification. As is common with finetuning, we replaced the final classification layer of the model with a random initialization and then performed additional training iterations with a reduced learning rate of 0.0025 to tailor the model to our custom dataset. The data of ten subjects were used for training, while the validation set consisted of the data of the remaining five subjects, which were chosen to be as representative as possible for the entire dataset. We chose to use the model that minimized the loss on the validation set (i.e., early stopping), which was obtained after just 4,000 iterations[2]. The performance of this model is compared in **Table 4** with the average precision (AP) metrics of the pretrained model on the original COCO dataset. Note that due to the limited domain of our dataset and the smaller number of classes our performance compares favorably to the larger COCO dataset. After training, we employed the model in inference mode to detect and segment objects in all videos of all subjects, as shown graphically in **Figure 3**.

2.3.3. Distances

The segmentation masks for all videos were stored to disk and then combined with the gaze data to calculate various distances. In the following, we restrict ourselves to segmentations that were recognized with a certainty score of at least 0.8. The distances that are of interest for our analyses are the following.

---

[2]The model is publicly available online (Gregori and Gijsberts, 2019).

TABLE 4 | Comparison of Mask R-CNN's detection accuracy on the COCO dataset and the accuracy of our finetuned model on the MeganePro dataset.

Dataset	AP [%]	AP50 [%]	AP75 [%]	Source
MeganePro	77.5	92.7	87.6	This work
COCO	33.6	55.2	35.3	He et al., 2017

The AP is the average precision over Intersection over Union (IoU) from 0.5 to 0.95 evaluated at steps of 0.05. AP50 and AP75 represent the average precision when the threshold of IoU is 0.5 or 0.75. A detailed description of these metrics can be found on the website of the COCO dataset (http://cocodataset.org).

- The *gaze-target* distance, which is the distance between the gaze point in frame coordinates and the target object for a grasp trial, if visible in the frame. If multiple instances of the same target class were recognized, then we chose the largest in terms of area.
- The *gaze-limb* distance denotes the distance between the gaze point and the hand or residual limb of the participant, if visible. We only consider instances identified as "human" that fall in the lower half of the image frame and again prefer the largest one.
- When applicable, the *limb-target* distance indicates the distance between the subject's hand or residual limb and the target object, as defined in the previous two distances.

Note that with the term "distance" we intend the minimum Euclidean distance in pixels between a point and the contour of a binary mask or between the contours of two binary masks. If these overlap, then the distance is 0. Note that with the scene camera of the Tobii glasses we find that $1\,\text{px} \simeq 0.72\,\text{mm}$ at a typical manipulation distance of 0.8 m.

2.4. Events

The profile of these distances and the modalities described previously were used to determine the timing of visuomotor events with respect to the stimulus, such as the first fixation on the target object or the onset of the arm movement. These events

FIGURE 3 | The procedure to segment the entire dataset. By means of the previously selected training set we fine-tune the Mask R-CNN model. Later we feed the network with new frames and it provides the segmented object instances as output.
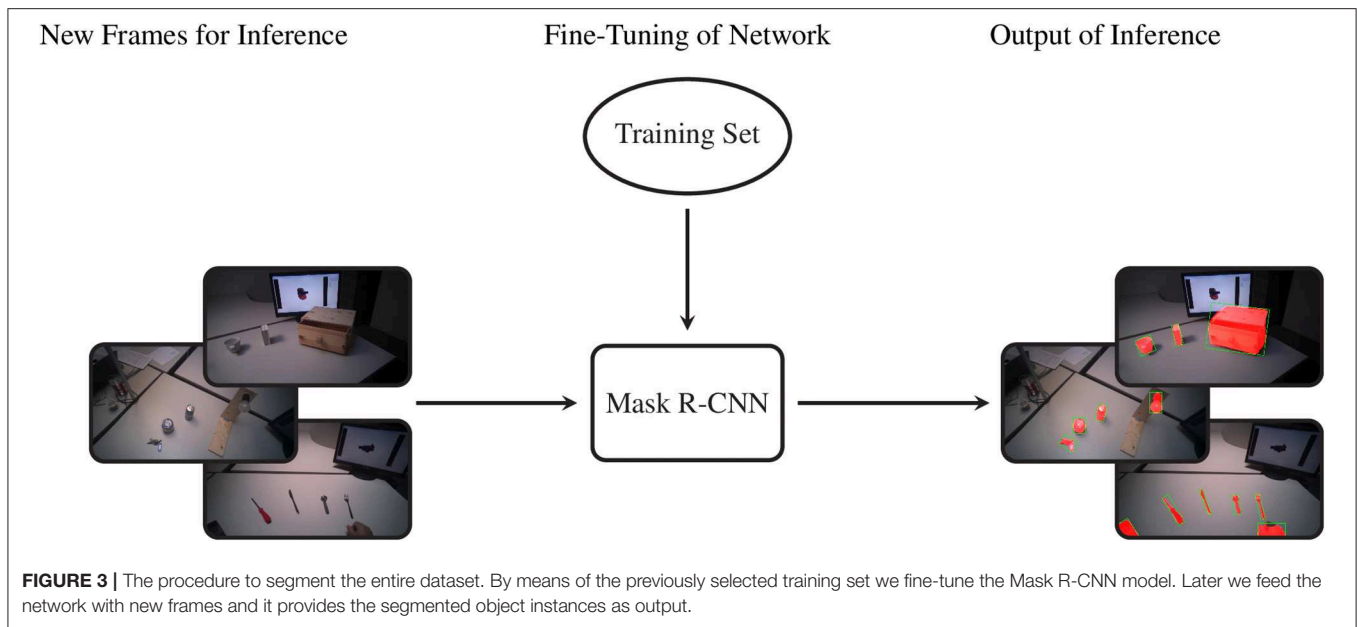
allow us to quantitatively describe the time interval between the activation of the eyes, head, and limb. The analysis window for each trial ranges from 2 s before until 2.5 s after the end of the corresponding vocal instruction with a resolution of 20 ms. We define the following events.

- The first *fixation* is defined as the first of at least two successive samples where the *gaze-target* distance is <20 px. This threshold was chosen to accommodate for some systematic error in the gaze tracking and is roughly twice the average gaze tracking accuracy (Cognolato et al., 2019). The requirement for two successive samples that fall below the threshold is to ignore occasional outliers.

- The *saccade* to the target object is assumed to initiate at the last sample where the gaze velocity was <70 °/s (Komogortsev et al., 2010), starting from 500 ms prior to the target *fixation*. This definition in terms of the last preceding fixation rather than the first saccade makes it robust against missing data from the eye tracker during saccades. Furthermore, we require this saccade to start from a *gaze-target* distance of at least 100 px to avoid occasional trials where the subject was already fixating the target object.

- The start of the *head* movement is defined as the first of two successive samples where the Euclidean norm of the angular velocity vector of the Tobii glasses exceeds 12 °/s. This threshold was chosen manually to be insensitive to systematic errors in the measurements of the gyroscope in the Tobii glasses.

- The movement of the *arm* starts at the first of two successive samples where the Euclidean norm of the three-axis accelerations exceeds 0.07 g. Also in this case the threshold was tuned manually to be insensitive to the baseline level of noise of the accelerometers.

- The activation of the *forearm muscles* starts when either of the myoelectric signals exceeds 4 times its baseline level for two

successive samples. This baseline level is taken as the average activation in the rest period from 2 s to 1 s before the vocal instruction ended.

- Finally, the first *grasp* occurs when there are two successive samples where the *limb-target* distance is <5 px. This threshold was chosen to allow for a small error margin in the detected segmentation masks.

Whenever the conditions for an event were not satisfied it was marked as missing for the corresponding trial. Furthermore, we invalidate all events that were found within the first 100 ms of the analysis window, as it implies that the subject was not in a rest position or was already fixating the target object.

3. RESULTS

In the following, we analyze the eye-hand coordination of the subjects in response to the grasp stimulus during the reach-to-grasp and manipulation phases. In other words, we relate movement of the eyes and head with that of the forearm. Before moving to these analyses, we verified that subjects effectively looked at the target object during a grasp trial. Thanks to the deep learning approach described previously, we determined that in 95.9% of the trials the *gaze-target* distance was <20 px at least once. Manual evaluation of the remaining 4.1% of the trials revealed that these were caused by a low accuracy of the gaze tracking that exceeded our threshold rather than lack of subject engagement.

3.1. Statistical Analysis

The first objective in this paper is to determine the window of opportunity in which gaze can provide useful information about an upcoming grasp. **Table 5** shows that for intact subjects there is a median interval of 561 ms between the *fixation* event and the subsequent *grasp* event. The same interval increases to more

TABLE 5 | Statistical description of the intervals in seconds between various events.

Interval	Intact				Amputated				Significance
	#	Q1	Med.	Q3	#	Q1	Med.	Q3	
fixation → grasp	8,144	0.321	0.561	0.842	1,942	0.581	1.042	1.644	$KS = 0.724, p = 2.602 \times 10^{-5}$
saccade → fixation	5,625	0.080	0.160	0.301	2,522	0.060	0.140	0.281	$KS = 0.190, p = 0.811$
saccade → head	5,419	−0.301	0.020	0.160	2,367	−0.461	−0.020	0.140	$KS = 0.338, p = 0.173$
head → arm	7,929	0.020	0.120	0.301	3,507	0.000	0.140	0.371	$KS = 0.262, p = 0.447$
arm → muscles	7,907	−0.020	0.080	0.401	3,576	0.200	0.581	1.042	$KS = 0.829, p = 4.524 \times 10^{-7}$

The count refers to the number of trials where both events were recognized, out of a total of 9,703 trials for intact and 4,482 trials for amputated subjects, respectively. For the Kolmogorov-Smirnov test the intervals were averaged per subject to guarantee independent samples.

than a second for amputated subjects, although this difference is because they did not physically interact with the objects and the *limb-target* distance therefore did not as often converge to within the 5 px threshold. Not surprisingly, a Kolmogorov-Smirnov test on the average interval per subject indicated that this difference between both subject groups was statistically significant. This is in contrast to the coordination between the initial saccade, the head, and the arm movements, for which we fail to find a significant difference between both groups. The saccade to the target object leads to its fixation in approximately[3] 150 ms. Concurrently with the eyes, also the head starts to move. This head movement is then followed by acceleration of the arm around 130 ms later. In intact subjects, the activation of the forearm muscles comes only 80 ms after the onset of the arm movement in the median case. This interval is more than half a second longer for amputated subjects and this difference is found to be statistically significant.

3.2. Reach-to-Grasp Phase

The coordination during the reaching phase of all "static" and "functional" grasps is visualized in **Figure 4** for both intact and amputated participants. Whereas the previous statistical analysis was intended to provide a quantitative assessment of the relative timings in eye-hand coordination, this figure instead complements those numbers by demonstrating how this coordination evolves over time. It does so by showing the median and quartiles of the distribution over all trials from all subjects in either group from 1.5 s before to 2.5 s after the conclusion of the vocal instruction. For both types of subjects, we observe an increase in gaze velocity from −0.5 s to 1 s. This increase also marks a sharp decrease in the distance between the gaze and the target object, which leads to a fixation soon after. From this moment on, the subjects retain their fixation on the object of interest. Based on the median profiles, we see again that the onset of the head movement starts around the same time as the eye movement and continues for 1.5 s.

The delay of the arm movement with respect to the eyes is slightly larger for amputated subjects, as shown by the median profile of the forearm's acceleration in **Figure 4**. Shortly after the arm starts to move, we also observe an increase in sEMG activity, with initially an emphasis on the extensor and later on

the flexor. For able-bodied subjects, the profile of the *limb-target* distance confirms our earlier finding that the limb arrives at the object 500 ms after its fixation. Although this result is not directly comparable with that for amputated subjects, we observe that the convergence between their residual limb and the target object appears more gradual and is characterized by a much larger variability.

A noteworthy observation is that the activation of the eyes always preceded the end of the vocal stimulus. The reason is that subjects could typically deduce the target object already before the end of the instruction. This does not affect our results, since we are interested in the relative delay between eyes, head, and forearm rather than reaction times to the stimulus. The differences in reaction time to the vocal instructions do increase however the dispersion of the distributions. We also note that the relative contribution among the three axes of the acceleration profile differs between able-bodied and amputated subjects. The reason is that we normalized this profile with respect to the initial position of the forearm, which is typically different for both types of subjects. In the present study, we use accelerometry to determine when the arm starts to move and rely on the limb-target distance to measure its convergence to the target object.

3.3. Manipulation Phase

In **Figure 5**, we focus on the behavior of intact and amputated subjects during the functional tasks to further investigate the similarities in gaze strategy. These figures start from 2 s before the vocal instruction until 7 s after, which is enough to cover the entire manipulation action. We group the MeganePro movements into three categories based on the type of task and the associated visual behavior, as shown in **Table 3**. These categories are *in place* manipulation actions, *lifting* actions, and finally *displacement* actions.

3.3.1. In Place Actions

The *in place* actions concern manipulation tasks that do not require moving the object, like opening an object, cutting bread, or pressing a button of the remote control. The aggregated profiles of all modalities for these actions are shown in **Figure 5A** for able-bodied subjects and in **Figure 5B** for amputees. During this type of action, the gaze remains fixed on the target object throughout the entire duration of the manipulation, as can also be seen in the example in **Figure 6** that overlays gaze and object

---

[3]This is likely a slight overestimation, considering our definition of the *saccade* and missing values in the gaze data from the Tobii glasses.
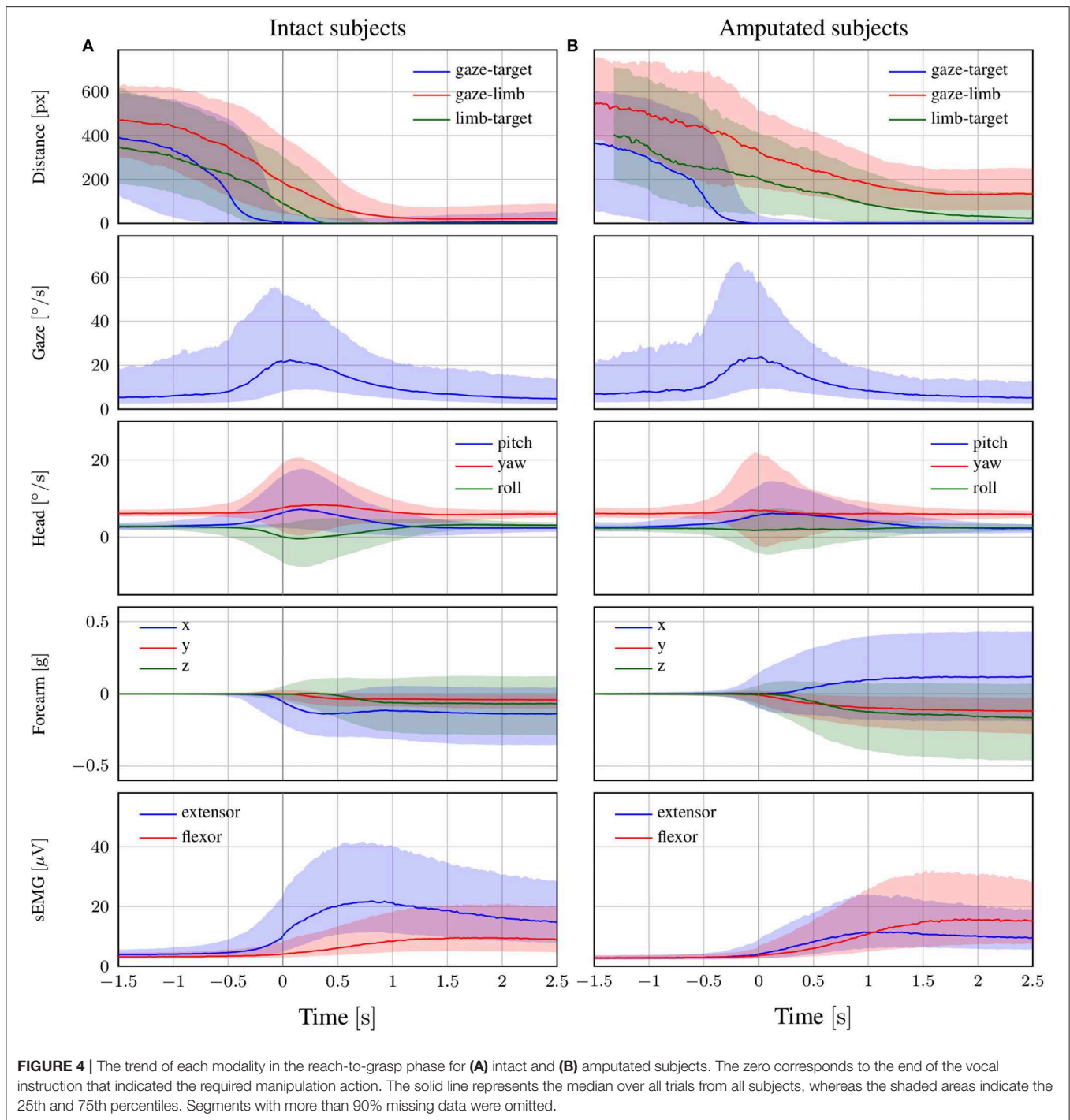
FIGURE 4 | The trend of each modality in the reach-to-grasp phase for **(A)** intact and **(B)** amputated subjects. The zero corresponds to the end of the vocal instruction that indicated the required manipulation action. The solid line represents the median over all trials from all subjects, whereas the shaded areas indicate the 25th and 75th percentiles. Segments with more than 90% missing data were omitted.

segmentations on representative frames of the first person video. As expected, the hand remains on the target for the entire duration in case of able-bodied subjects, whereas for amputes there remains a constant subject-dependent distance between the residual limb and the target. Head movements are limited to the initial reach-to-grasp phase to center the object in the field of view, after which the head remains fixed until the end of the manipulation.

3.3.2. Lifting Actions
The second group is composed of *lifting* actions, in which the subject was required to lift an object up and then place it back in its initial position. As can be seen in **Figures 5C,D**, also in this case the gaze anticipates head and forearm movement. More interestingly, we see a clear movement in the pitch orientation of the head. Since these actions are executed while standing, the subjects first lower their head to locate the target object on
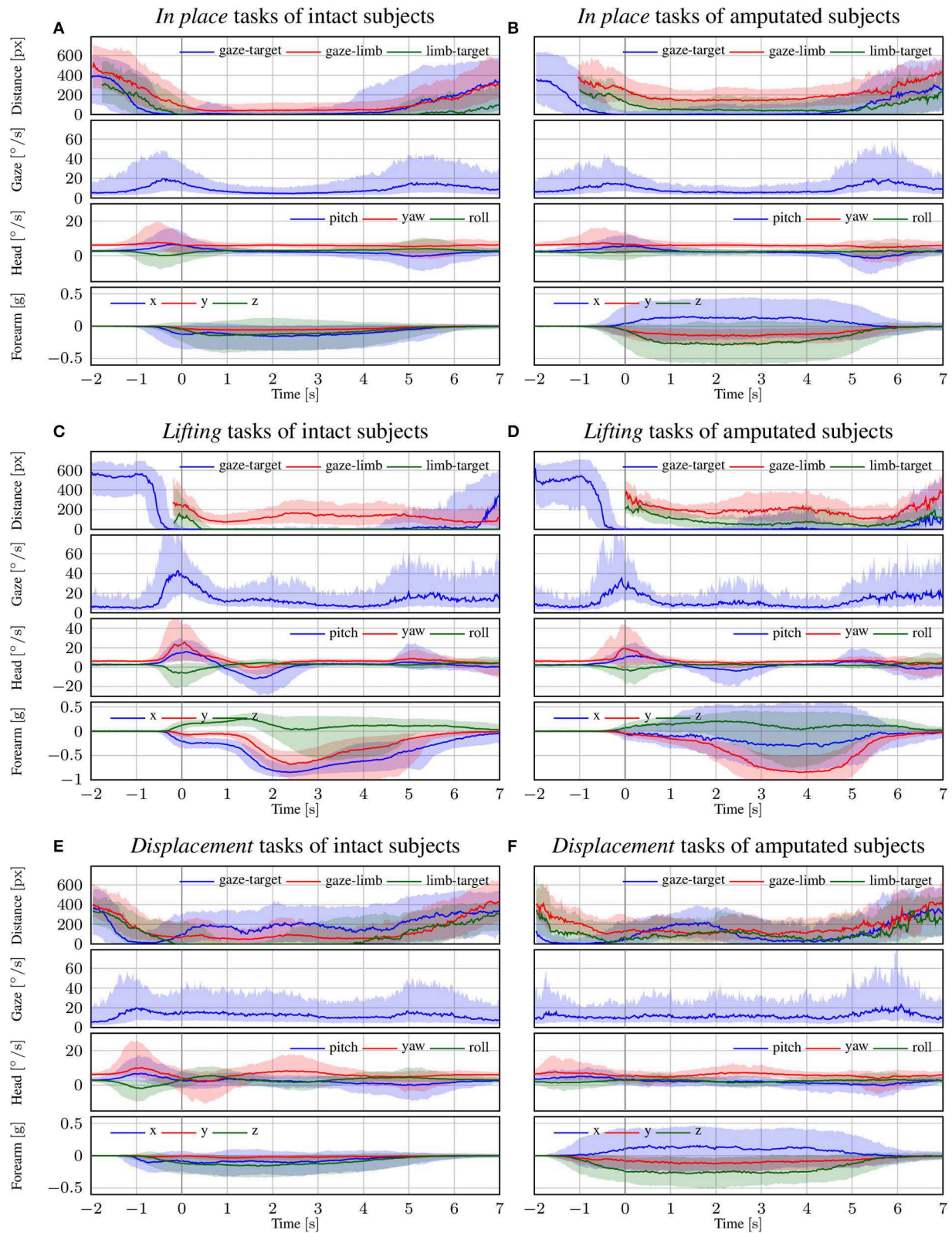
FIGURE 5 | The trend of each modality for intact and amputated subjects for the **(A,B)** *in place*, **(C,D)** *lifting*, and **(E,F)** *displacement* functional tasks. The zero corresponds to the end of the vocal instruction that indicated the required manipulation action. The solid line represents the median over all trials from all subjects, whereas the shaded areas indicate the 25th and 75th percentiles. Segments with more than 90% missing data were omitted.
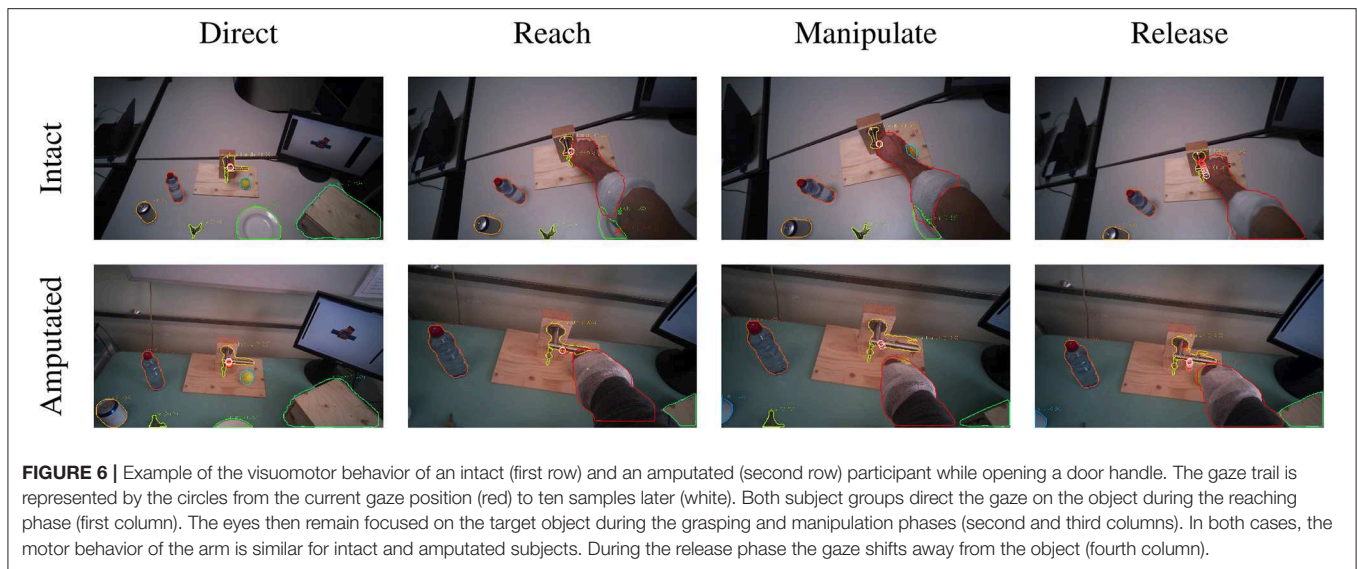
FIGURE 6 | Example of the visuomotor behavior of an intact (first row) and an amputated (second row) participant while opening a door handle. The gaze trail is represented by the circles from the current gaze position (red) to ten samples later (white). Both subject groups direct the gaze on the object during the reaching phase (first column). The eyes then remain focused on the target object during the grasping and manipulation phases (second and third columns). In both cases, the motor behavior of the arm is similar for intact and amputated subjects. During the release phase the gaze shifts away from the object (fourth column).

the table. Then, when they have located and grasped the object, they raise their head again with a peak pitch velocity at 1.7 s for able-bodied subjects and slightly later for amputated subjects. This head movement coincides with a modestly increased gaze velocity and is due to the tracking motion of the lifting action. In some cases, this tracking strategy even caused an amputated subject's *gaze-target* distance to increase, as can also be seen in the example in **Figure 7**. Finally, the subjects lower their head again when tracking the release of the object at the end of the trial.

3.3.3. Displacement Actions

The final category are the so-called *displacement* actions. During these tasks, the subjects had to grasp the objects, move them horizontally to another position, and then place them back in the initial position. We note that the gaze and motor behavior starts earlier with respect to the vocal instruction. For this category of tasks, the name of the object happens to appear at the beginning of the instruction (see **Table 3**), thus allowing subjects to initiate the task early. For intact subjects, we see in **Figure 5E** that 200 ms before the hand reaches the object the *gaze-target* distance starts to increase again. The gaze, in this case, shifts already to the destination position for the displacement action, as demonstrated in the second panel in **Figure 8**. Although less pronounced, the same pattern repeats itself at around 1.5 s when the subject initiates the return movement. The profiles for the amputated subjects in **Figure 5F** show different behavior, with an overall increase in *gaze-target* distance throughout the entire duration of the movement. As intact subjects did, their gaze anticipates the path of the hand rather than the path of the object, which is not physically displaced. This strategy is demonstrated clearly in the bottom row of **Figure 8**.

4. DISCUSSION

The objective of this paper was to determine the window of opportunity for exploiting gaze as contextual information in

decoding the grasp intent of amputees. A related question was to which extent the natural gaze strategies of amputees and able-bodied subjects were similar. After comparing our results with related work, we discuss both topics. Finally, we argue for the use of recent developments in deep learning in the analysis of large-scale visuomotor studies.

4.1. Visuomotor Strategy and Comparison With Related Work

In section 3.2, we presented the results of eye, head, and limb coordination during reaching and grasping. The eyes are the first to react to the vocal stimulus by exhibiting an increasing saccade-related activity, leading to a fixation on the target in about 150 ms. When the eyes start moving, also the head follows almost immediately. Such short delays between movement of the eyes and the head have been reported in the literature, ranging from 10 ms to 100 ms during a block-copying task (Smeets et al., 1996) or in reaction to visual stimuli (Goldring et al., 1996; Di Cesare et al., 2013). This behavior is however strongly dependent on the experimental setting and even small variations therein can change the outcome. For instance, Pelz et al. (2001) found that depending on the exercise's instruction the head may both precede (by about 200 ms) or follow the eyes (by about 50 ms) in the same block-copying task.

After the activation of the eyes and the head we observe the movement onset of the arm 130 ms later. Similar values ranging from 170 ms to 300 ms were also reported by Smeets et al. (1996) and Pelz et al. (2001) in a block-copying task and by Belardinelli et al. (2016) in a pick and place task. Land et al. (1999) instead found a median delay of 0.56 s during a tea-making activity. Rather than movement onset, the time the hand takes to reach the target is more interesting for our scope. For the intact subjects, the hand typically starts to occlude the target object around 500 ms after the first fixation. Although occlusion does not necessarily already imply a completed grasp, especially given the first person perspective, we do expect the grasp to follow

FIGURE 7 | Example of the visuomotor behavior of an intact (first row) and an amputated (second row) participant lifting a plate. The gaze trail is represented by the circles from the current gaze position (red) to ten samples later (white). The eyes focus on the manipulation point to plan the hand's approach (first and second columns). During the lifting phase, the eyes move away from the reaching point and the amputee's gaze even exceeds the mask boundary of the plate (third column). The object is fixated again during the release (fourth column).

FIGURE 8 | Example of the visuomotor behavior of an intact (first row) and an amputated (second row) participant while moving a ball. The gaze trail is represented by the circles from the current gaze position (red) to ten samples later (white). The gaze focuses on the object until the hand's arrival (first column), when the grasping phase begins the eyes shift away toward the destination (second column). When the hand reaches the destination the gaze shifts back to the initial location (third column) to release the target (fourth column).

not much later. These results confirm that visual attention on objects anticipates manipulation. In previous studies concerning displacements (Johansson et al., 2001; Belardinelli et al., 2016; Lavoie et al., 2018) and grasping activities (Brouwer et al., 2009), a variable delay ranging from 0.53 s to 1.3 s was found between the eye and hand. Also in these cases, the exact value of the delay depends on the characteristics of the experiment.

In section 3.3, we concentrated on the visuomotor strategy adopted by amputated and able-bodied subjects to interact with the objects during three groups of functional tasks. We can characterize the strategies associated with these groups in terms of the types of fixations defined by Land et al. (1999) and Land and Hayhoe (2001), namely *locating*, *directing*, *guiding*, and *checking*. A fixation to *locate* is typically done at the beginning of an action, to mentally map the location of objects that is to be used. Instead, a fixation to *direct* is meant to detect an object

that will be used immediately after. Fixations to *guide* are usually multiple and occur when the gaze shifts among two or more objects that are approaching each other. Finally, there are long *checking* fixations to monitor the state of an action waiting for its completion.

The visual strategy of the *in place* actions is relatively straightforward. In these tasks, subjects initiate with a fixation to *direct* the attention to the target object. Subsequently, their fixation remains on the manipulated object to *check* the correct execution of the task. Note that this visual attention seems focused on the target object rather than the subject's hand, as can be seen comparing the *gaze-target* and *gaze-limb* distances in **Figures 5A,B**. Indeed, Land et al. (1999) noted that the hands themselves are rarely fixated.

Also the *lifting* actions start with a *directing* fixation to locate the object of interest. However, whereas the initial fixation is

focused on the intended grasp location (cf. the left column in **Figure 7**), the gaze shifts upwards when the hand has grasped the object. This coincides with the transition from the *directing* fixation to visually *checking* the lifting action. This is in line with observations by Voudouris et al. (2018), who noted that people may fixate higher when grasping and lifting an object to direct their gaze to where the object will be in the future.

Finally, *displacement* actions are the ones most investigated in the literature. Previous studies on pick and place tasks (Belardinelli et al., 2016; Lavoie et al., 2018) and on the block-copying task (Smeets et al., 1996; Pelz et al., 2001) fall in this category. In this case, we observe in **Figure 5E** that the *gaze-target* and *gaze-limb* distances have three minima for intact subjects, namely at the initial pick-up, the destination, and at the release again at the initial position. All three minima indicate fixations that are meant to *direct* the approach of the hand, either for (1) grasping the object, (2) displacing it, or finally (3) releasing it. This behavior can clearly be seen for both intact as well as amputated subjects in the example in **Figure 8**. We also notice that the eyes did not wait for the completion of the pick-up action, moving instead toward the position of the destination around 200 ms in advance. This proactive role of the eyes was highlighted by Land et al. (1999), who measured the gaze moving on to the next object between 0 s to 1 s before the current object manipulation was terminated. Also Pelz et al. (2001) observed the eyes departing from the target object 100 ms to 150 ms before the arrival of the hand.

4.2. Comparison Between Intact and Amputated Subjects

One of the aims of this work was to understand if a transradial amputation has introduced important changes in the visuomotor behavior of amputees. During the reach-to-grasp phase, the overall behavior of intact and amputated subjects is comparable. Even if the coordination timeline between eyes, head, and limb is similar, there are some minor discrepancies between the two groups. The main observed difference concerns the delayed activation of the forearm muscles during the reaching phase for amputated subjects, which was found to be statistically significant. Similarly, during the lifting tasks we noted slower pitch movements of the head. It is likely that some subjects interpreted the instruction to perform the grasp with their missing limb by activating their phantom limb. Such movements executed with the phantom limb are known to be slower than those executed with the intact hand (Raffin et al., 2012b; De Graaf et al., 2016).

Throughout the manipulation phase, we observe a striking similarity in visuomotor strategy between the amputated subjects and the control group. The differences that we noted in the results are not due to an alternative gaze strategy, but rather because the objects were not physically moved during the interaction. For instance, in the *lifting* task visualized in **Figure 5D** we saw an increase in *gaze-target* distance in the range from 2 s to 5 s. This increase was due to an upward shift in the gaze location to track where the object would have been if it had been lifted

for real. Similarly, during the displacement task in **Figure 5F** we do not observe a minimum in *gaze-target* distance at around 1.5 s, as was the case for intact subjects (see **Figure 5E**). Instead, around the same time we observe a peak for the amputated subjects, solely because the target object is still at its original position whereas their gaze has shifted to the intermediate destination. The examples for these gaze strategies in **Figures 7, 8** demonstrate how similar intact and amputated subjects behaved.

It would be interesting to understand how these results relate to the disrupted eye-hand coordination when using a prosthetic device. Previous studies (Bouwsema et al., 2012; Sobuh et al., 2014; Parr et al., 2018) have underlined that prosthetic users are more fixated on guiding the current manipulation, rather than planning the follow-up action. This behavior is most likely caused by the fact that amputated people rely almost exclusively on visual feedback. However, since only a small number of subjects were engaged in the previous studies more research will be needed to fully understand the disruption of the visuomotor strategy. In particular, whether or not this strategy improves when the user develops trust in the prosthesis (Chadwell et al., 2016) merits attention. Another equally interesting question is to which extent the proactive gaze behavior can be restored by integrating tactile or proprioceptive feedback in the prosthesis (Cipriani et al., 2011; Marasco et al., 2018; Markovic et al., 2018, among others).

4.3. Integration of Vision in Prostheses to Improve Intent Recognition

The estimated time interval from *fixation* to *grasp* in section 3.1 shows that the window of opportunity is 500 ms for intact subjects. This interval cannot be accurately determined for amputated subjects, as they executed the movement with their missing limb and therefore lacked physical contact with the target object. Although **Figure 4B** suggests that this window will at least be as long for amputated users, one may argue that this result is not representative for movements performed with a prosthesis. However, previous studies showed without exception that prosthetic users still fixate the target object for the majority of the reaching phase (Bouwsema et al., 2012; Sobuh et al., 2014; Chadwell et al., 2016; Hebert et al., 2019; Parr et al., 2019), albeit alternating it more often with fixations on the hand (i.e., the "switching" strategy). Moreover, this reaching phase may actually take more than twice as long as compared to the same movement performed with the anatomical limb (Sobuh et al., 2014; Hebert et al., 2019). These findings suggest that the target object will still be fixated proactively by a prosthetic user and that the window of opportunity will more likely be longer than shorter.

Exploiting this anticipatory gaze behavior is appealing because it comes naturally and therefore does not require specific attention from the user. The success of this approach relies however on the ability to distinguish informative fixations from those that are not necessarily related to any grasp intent. Gigli et al. (2018) attempted to address this problem by including the onset of the arm movement as an additional condition, which we have shown here to shorten the window of opportunity. Also the method that is used to detect fixations may shorten this window. Thanks to the frame-by-frame segmentations in the

present study, we could accurately and instantaneously recognize object fixations by measuring the distance between the object's segmentation mask and the gaze point. In contrast, common fixation classifiers, such as (IVT) (Salvucci and Goldberg, 2000), define a fixation simply as the lack of eye movement. In reality, gaze shifts more commonly involve not only eye movement, but also head and sometimes even trunk movements (Morasso et al., 1973; Land, 2006). When the head moves, the optokinetic and vestibulo-ocular reflexes cause the eyes to counteract the head movement to maintain a stable gaze point (Lappe and Hoffmann, 2000). It is exactly due to such coordinated gaze movements that the initial object fixation in **Figure 4** actually coincides with a *peak* in gaze velocity. The need to detect fixations as early as possible therefore implies a detection method that uses more information than eye movement alone. Whether this is best done by compensating for head movements (Kinsman et al., 2012; Larsson et al., 2014) or by comparing the visual object at the gaze point as in the present study is an open question.

A final consideration is regarding technical and practical concerns of a prosthetic solution that integrates eye tracking. Myoelectric control of prostheses has a long history and a solution that decodes natural muscle activations via pattern recognition is commercially available (Coapt, LLC, 2015). Tracking a user's gaze continuously and reliably in a variety of conditions will pose a bigger problem, however. The Tobii glasses used for the MeganePro dataset resulted in 10.7% of missing data on average, caused discomfort to the subjects after wearing them for about 2 h, and needed a battery replacement after 1.5 h to 2 h of continuous acquisition. Recent developments have seen considerable improvements however in terms of weight, cost, and aesthetic appeal (Pupil Labs GmbH, 2019).

4.4. Advantages of Deep Learning for the Automatic Analysis of Visual Behavior

Without the deep learning approach described in section 2.3 it would have been extremely labor intensive to analyze 70 h of video and data from 44 subjects. Manufacturers of eye-tracking devices often provide applications for semi-automatic analyses, but these do not allow the level of automation nor precision as the procedure described here. Although the object segmentations produced by Mask R-CNN were occasionally mistaken, the segmentations seen in the examples from **Figures 6–8** are illustrative for the overall performance. It may easily be overlooked that data from research studies, such as the present, often contain much less visual variability than the datasets on which these algorithms are trained and evaluated. With minimal finetuning efforts, it is therefore likely to obtain levels of performance that considerably exceed those reported in the literature, as was seen in **Table 4**.

5. CONCLUSIONS

In this study, we analyzed the coordination of eye, head, and limb movements of amputated and able-bodied participants engaged in manipulation tasks of household objects. Our aim was to understand the anticipatory role of gaze in the visuomotor strategy and to determine whether this could potentially be used

to aid in the grasp intent recognition for upper limb prostheses. We found that a fixation on the target object typically preceded the subsequent grasp by 500 ms in intact subjects and possibly longer for amputees. Moreover, the visuomotor strategies of amputees were similar to those of intact subjects both during the reach-to-grasp phase as well as during functional manipulation tasks. In future work, we aim to use the knowledge gained in this study to integrate vision with the (sEMG) modality to verify whether we can realize an effective improvement in recognizing grasp intentions during the reaching phase.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Commission of the canton of Valais in Switzerland and by the Ethics Commission of the Province of Padova in Italy. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

The consortium contributed the initial conception of the study. VG and AG contributed conception and design of the study. MC, GS, MA, and the consortium collected the dataset. VG devised and implemented the automated segmentation under the supervision of AG. VG collected the segmentation dataset. VG performed the visuomotor analyses under the supervision of AG. VG wrote the first draft of the manuscript. AG rewrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

CONSORTIUM

Aside from the authors, the MeganePro consortium consists of Katia Giacomino and Anne-Gabrielle Mittaz Hager (University of Applied Sciences Western Switzerland "HES-SO Valais," Leukerbad, Switzerland), Diego Faccio, Cesare Tiengo, and Franco Bassetto (Padova University Hospital, Padova, Italy), Barbara Caputo (Istituto Italiano di Tecnologia, Turin, Italy and Politecnico di Torino, Turin, Italy), Peter Brugger (University Hospital of Zurich, Zurich, Switzerland), and Henning Müller (University of Applied Sciences Western Switzerland "HES-SO Valais," Sierre, Switzerland).

REFERENCES

Admoni, H., and Srinivasa, S. (2016). "Predicting user intent through eye gaze for shared autonomy," in *2016 AAAI Fall Symposia* (Arlington, VA).

Belardinelli, A., Stepper, M. Y., and Butz, M. V. (2016). It's in the eyes: planning precise manual actions before execution. *J. Vis.* 16:18. doi: 10.1167/16.1.18

Blank, A., Okamura, A. M., and Kuchenbecker, K. J. (2010). Identifying the role of proprioception in upper-limb prosthesis control: studies on targeted motion. *ACM Trans. Appl. Percept.* 7:15. doi: 10.1145/1773965.1773966

Bouwsema, H., Kyberd, P. J., Hill, W., van der Sluis, C. K., and Bongers, R. M. (2012). Determining skill level in myoelectric prosthesis use with multiple outcome measures. *J. Rehab. Res. Dev.* 49, 1331–1347. doi: 10.1682/JRRD.2011.09.0179

Brouwer, A.-M., Franz, V. H., and Gegenfurtner, K. R. (2009). Differences in fixations between grasping and viewing objects. *J. Vis.* 9, 18.1–18.24. doi: 10.1167/9.1.18

Castellini, C. and Sandini, G. (2006). "Gaze tracking for robotic control in intelligent teleoperation and prosthetics," in *COGAIN 2006* (Turin), 73–77.

Chadwell, A., Kenney, L., Thies, S., Galpin, A., and Head, J. (2016). The reality of myoelectric prostheses: understanding what makes these devices difficult for some users to control. *Front. Neurorobot.* 10:7. doi: 10.3389/fnbot.2016.00007

Cipriani, C., Controzzi, M., and Carrozza, M. C. (2011). The smarthand transradial prosthesis. *J. Neuroeng. Rehab.* 8:29. doi: 10.1186/1743-0003-8-29

Coapt, LLC (2015). *Coapt Engineering.* Available online at: https://www.coaptengineering.com

Cognolato, M., Gijsberts, A., Gregori, V., Saetta, G., Giacomino, K., Mittaz Hager, A.-G., et al. (2019). Gaze, visual, myoelectric, and inertial data of grasps for intelligent prosthetics. medRxiv [Preprint]. Cold Spring Harbor Laboratory Press. doi: 10.1101/19010199

De Graaf, J. B., Jarrassé, N., Nicol, C., Touillet, A., Coyle, T., Maynard, L., et al. (2016). Phantom hand and wrist movements in upper limb amputees are slow but naturally controlled movements. *Neuroscience* 312, 48–57. doi: 10.1016/j.neuroscience.2015.11.007

Duchowski, A. T. (2007). *Eye Tracking Methodology: Theory and Practice.* London, UK: Springer.

Gigli, A., Gregori, V., Cognolato, M., Atzori, M., and Gijsberts, A. (2018). "Visual cues to improve myoelectric control of upper limb prostheses," in *IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob)* (Enschede), 783–788. doi: 10.1109/BIOROB.2018.8487923

Goldring, J. E., Dorris, M. C., Corneil, B. D., Ballantyne, P. A., and Munoz, D. P. (1996). Combined eye-head gaze shifts to visual and auditory targets in humans. *Exp. Brain Res.* 111, 68–78. doi: 10.1007/BF00229557

Gregori, V., and Gijsberts, A. (2019). Mask-RCNN Object Segmentation for MeganePro V1. Harvard Dataverse. doi: 10.7910/DVN/STIXXJ

Hayhoe, M. M., McKinney, T., Chajka, K., and Pelz, J. B. (2012). Predictive eye movements in natural vision. *Exp. Brain Res.* 217, 125–136. doi: 10.1007/s00221-011-2979-2

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2961–2969. doi: 10.1109/ICCV.2017.322

Hebert, J. S., Boser, Q. A., Valevicius, A. M., Tanikawa, H., Lavoie, E. B., Vette, A. H., et al. (2019). Quantitative eye gaze and movement differences in visuomotor adaptations to varying task demands among upper-extremity prosthesis users. *JAMA Netw. Open* 2:e1911197. doi: 10.1001/jamanetworkopen.2019.11197

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932. doi: 10.1523/JNEUROSCI.21-17-06917.2001

Kinsman, T., Evans, K., Sweeney, G., Keane, T., and Pelz, J. (2012). "Ego-motion compensation improves fixation detection in wearable eye tracking," in *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, CA: ACM), 221–224. doi: 10.1145/2168556.2168599

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., and Gowda, S. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Trans. Biomed. Eng.* 57, 2635–2645. doi: 10.1109/TBME.2010.2057429

Koochaki, F., and Najafizadeh, L. (2018). "Predicting intention through eye gaze patterns," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (Cleveland, OH: IEEE), 1–4. doi: 10.1109/BIOCAS.2018.8584665

Land, M., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 1311–1328. doi: 10.1068/p2935

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Prog. Retinal Eye Res.* 25, 296–324. doi: 10.1016/j.preteyeres.2006.01.002

Land, M. F., and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vis. Res.* 41, 3559–3565. doi: 10.1016/S0042-6989(01)00102-X

Land, M. F., and Lee, D. N. (1994). Where we look when we steer. *Nature* 369:742. doi: 10.1038/369742a0

Land, M. F., and McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nat. Neurosci.* 3:1340. doi: 10.1038/81887

Lappe, M., and Hoffmann, K.-P. (2000). Optic flow and eye movements. *Int. Rev. Neurobiol.* 44, 29–50. doi: 10.1016/S0074-7742(08)60736-9

Larsson, L., Schwaller, A., Holmqvist, K., Nyström, M., and Stridh, M. (2014). "Compensation of head movements in mobile eye-tracking data using an inertial measurement unit," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (Seattle, WA: ACM), 1161–1167. doi: 10.1145/2638728.2641693

Lavoie, E. B., Valevicius, A. M., Boser, Q. A., Kovic, O., Vette, A. H., Pilarski, P. M., et al. (2018). Using synchronized eye and motion tracking to determine high-precision eye-movement patterns during object-interaction tasks. *J. Vis.* 18:18. doi: 10.1167/18.6.18

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 2117–2125. doi: 10.1109/CVPR.2017.106

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: common objects in context," in *European Conference on Computer Vision* (Zurich: Springer), 740–755. doi: 10.1007/978-3-319-10602-1_48

Marasco, P. D., Hebert, J. S., Sensinger, J. W., Shell, C. E., Schofield, J. S., Thumser, Z. C., et al. (2018). Illusory movement perception improves motor control for prosthetic hands. *Sci. Transl. Med.* 10:eaao6990. doi: 10.1126/scitranslmed.aao6990

Markovic, M., Dosen, S., Cipriani, C., Popovic, D., and Farina, D. (2014). Stereovision and augmented reality for closed-loop control of grasping in hand prostheses. *J. Neural Eng.* 11:046001. doi: 10.1088/1741-2560/11/4/046001

Markovic, M., Dosen, S., Popovic, D., Graimann, B., and Farina, D. (2015). Sensor fusion and computer vision for context-aware control of a multi degree-of-freedom prosthesis. *J. Neural Eng.* 12:066022. doi: 10.1088/1741-2560/12/6/066022

Markovic, M., Schweisfurth, M. A., Engels, L. F., Bentz, T., Wüstefeld, D., Farina, D., et al. (2018). The clinical relevance of advanced artificial feedback in the control of a multi-functional myoelectric prosthesis. *J. Neuroeng. Rehab.* 15:28. doi: 10.1186/s12984-018-0371-1

Massa, F., and Girshick, R. (2018). *Maskrcnn-Benchmark: Fast, Modular Reference Implementation of Instance Segmentation and Object Detection algorithms in PyTorch.* Available online at: https://github.com/facebookresearch/maskrcnn-benchmark

Merletti, R. (1999). Standards for reporting EMG data. *J. Electromyogr. Kinesiol.* 9, 3–4.

Morasso, P., Bizzi, E., and Dichgans, J. (1973). Adjustment of saccade characteristics during head movements. *Exp. Brain Res.* 16, 492–500.

Olsen, A., and Matos, R. (2012). "Identifying parameter values for an i-vt fixation filter suitable for handling data sampled with various sampling frequencies," in *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12* (New York, NY: ACM), 317–320.

Parr, J. V. V., Vine, S. J., Harrison, N. R., and Wood, G. (2018). Examining the spatiotemporal disruption to gaze when using a myoelectric prosthetic hand. *J. Motor Behav.* 50, 416–425. doi: 10.1080/00222895.2017.1363703

Parr, J. V. V., Vine, S. J., Wilson, M. R., Harrison, N. R., and Wood, G. (2019). Visual attention, EEG alpha power and t7-fz connectivity are implicated in prosthetic

hand control and can be optimized through gaze training. *J. Neuroeng. Rehab.* 16:52. doi: 10.1186/s12984-019-0524-x

Patla, A. E., and Vickers, J. N. (2003). How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Exp. Brain Res.* 148, 133–138. doi: 10.1007/s00221-002-1246-y

Pelz, J., Hayhoe, M., and Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Exp. Brain Res.* 139, 266–277. doi: 10.1007/s002210100745

Pupil Labs GmbH (2019). *Pupil Labs Invisible.* Available online at: https://pupil-labs.com/products/invisible/

Raffin, E., Giraux, P., and Reilly, K. T. (2012a). The moving phantom: motor execution or motor imagery? *Cortex* 48, 746–757. doi: 10.1016/j.cortex.2011.02.003

Raffin, E., Mattout, J., Reilly, K. T., and Giraux, P. (2012b). Disentangling motor execution from motor imagery with the phantom limb. *Brain* 135, 582–595. doi: 10.1093/brain/awr337

Salvucci, D. D., and Goldberg, J. H. (2000). "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, FL: ACM), 71–78.

Saran, A., Majumdar, S., Shor, E. S., Thomaz, A., and Niekum, S. (2018). "Human gaze following for human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 8615–8621.

Scotto Di Cesare, C. S., Anastasopoulos, D., Bringoux, L., Lee, P.-Y., Naushahi, M., and Bronstein, A. M. (2013). Influence of postural constraints on eye and head latency during voluntary rotations. *Vis. Res.* 78, 1–5. doi: 10.1016/j.visres.2012.11.011

Smeets, J. B., Hayhoe, M. M., and Ballard, D. H. (1996). Goal-directed arm movements change eye-head coordination. *Exp. Brain Res.* 109, 434–440.

Sobuh, M. M., Kenney, L. P., Galpin, A. J., Thies, S. B., McLaughlin, J., Kulkarni, J., et al. (2014). Visuomotor behaviours when using a myoelectric prosthesis. *J. Neuroeng. Rehab.* 11:72. doi: 10.1186/1743-0003-11-72

Tatler, B. W. (2014). "Eye movements from laboratory to life," in *Current Trends in Eye Tracking Research* (Cham: Springer), 17–35.

Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *J. Vis.* 11:5. doi: 10.1167/11.5.5

Tundo, M. D., Lemaire, E., and Baddour, N. (2013). "Correcting smartphone orientation for accelerometer-based analysis," in *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (Gatineau, QC: IEEE), 58–62.

Voudouris, D., Smeets, J. B. J., Fiehler, K., and Brenner, E. (2018). Gaze when reaching to grasp a glass. *J. Vis.* 18:16. doi: 10.1167/18.8.16

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., and Torr, P. H. (2019). "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA).

Excessive Lateral Trunk Lean in Patients With Cerebral Palsy: Is It Based on a Kinematic Compensatory Mechanism?

Roman Rethwilm [1,2]*, Harald Böhm [1], Chakravarthy U. Dussa [1] and Peter Federolf [2]

[1] Orthopedic Children's Hospital Aschau, Aschau im Chiemgau, Germany, [2] Department of Sport Science, University of Innsbruck, Innsbruck, Austria

Introduction: Excessive lateral trunk lean is a commonly observed gait deviation in children with cerebral palsy (CP), with implications for energy expenditure and the development of back pain. While the trunk lean toward the stance leg is widely interpreted as a compensatory strategy to unload the hip, in CP the relation to hip abductor muscle strength is only weak. Therefore, other mechanisms may play a role in the prevalence of excessive trunk lean in CP, or it could be a primary motor function deficit.

Research Question: Is the excessive lateral trunk lean in patients with CP part of an underlying biomechanical mechanism?

Materials and Methods: Patients with bilateral CP ($N = 255$; age 13.6 ± 6.6 years) were retrospectively included and divided into a group with ($n = 174$) and without ($n = 81$) excessive lateral trunk lean. Ten lower-extremity joint angle waveforms were analyzed using a principal component analysis (PCA) to identify patterns of correlated deviations from average angle waveforms. Binary logistic regressions were performed to determine the discriminative capacity of the identified patterns.

Results: The PCA identified correlated kinematic patterns, with lower-order patterns showing more common gait pathologies, such as torsional malalignments and crouch gait pattern. Within five patterns, significant ($p < 0.0025$) group differences were identified. Interestingly, the trunk lean was not always distinctive in these patterns and despite the significant differences their effect sizes were small. The logistic regression was unable to reliably classify patients based on their trunk lean patterns.

Discussion: The current study identified multiple trunk lean-related patterns, however, excessive trunk lean was not attributable to a distinctive CP related gait pathology or to a specific compensatory strategy. Overall, the results do not support the hypothesis that excessive trunk lean is part of a biomechanical mechanism. Therefore, it seems more likely that excessive lateral trunk lean is based on other disease specific dysfunctions, influenced by the severity of the disease.

Keywords: 3D gait analysis, gait pattern, principal component analysis, trunk control, Duchenne gait, motor function, compensatory strategy, cerebral palsy

INTRODUCTION

An excessive lateral trunk lean is a commonly observed gait deviation in patients with cerebral palsy (CP) (Attias et al., 2015), with a prevalence as high as 72% in children with CP (Klum et al., 2015). This excessive trunk lean has been shown to result in an increased cost of locomotion (Salami et al., 2017) and may also contribute to the development of back pain, which is one of the most common pain sites in adults with CP (Opheim et al., 2009).

From a clinical perspective, an increased lateral trunk lean is widely viewed as a compensatory mechanism for hip abductor muscle weakness (Schmid et al., 2013), also known as Duchenne gait. This mechanism describes the inclination of the trunk toward the ipsilateral stance leg, which shifts the center of mass (COM) laterally and reduces the hip abduction moments significantly (Salami et al., 2017).

It is often challenging to identify the cause of an observed gait deviation and to determine whether the nature of a specific abnormal pattern is a primary feature of the disease or rather a compensatory strategy adopted to cope with an underlying gait problem (Schmid et al., 2013). Identifying a mechanism as primary or compensatory has far-reaching implications. Not knowing the underlying source of an abnormal movement pattern could result in untreated primary pathology or, even worse, in unnecessary treatment of a compensatory mechanism that would resolve when the primary pathology is addressed directly.

The lateral trunk lean in CP gait is a case where the source is not as clear and apparent as in some other gait phenomena. For example, compared to the name-giving Duchenne muscle dystrophy, studies in CP found only weak—yet significant—correlations with hip abductor muscle strength (Krautwurst et al., 2013; Klum et al., 2015). This indicates that muscle weakness contributes to the occurrence of an excessive lateral trunk lean, but also suggests that other factors may play a role or could possibly be more prevalent. Accordingly, some researchers suspected further contributing causes, such as bony deformities (Salami et al., 2017) or an underlying primary motor control deficit (Heyrman et al., 2014).

To gain further insights into potential underlying mechanisms for an excessive trunk lean, the current study explored if excessive trunk lean is part of a kinematic movement pattern that can be frequently observed in the gait of CP patients. To determine kinematic movement patterns, we applied a principal component analysis (PCA), a statistical method that identifies correlated patterns (PC-eigenvectors) in multi-dimensional data (Daffertshofer et al., 2004; Eskofier et al., 2013; Robertson et al., 2014). Furthermore, a score for each pattern and patient is computed, indicating the extent to which each individual patient exhibits a particular pattern. For such a pattern to represent a functional mechanism, we postulated three criteria: (1) excessive trunk lean needs to be a part of the pattern. (2) If CP patients are classified into a group showing excessive trunk lean (eTL) and a group non-excessive trunk lean (nTL), then we expected to find significant differences and medium or high effect sizes for the scores that patients in these groups receive. (3) We considered that there could be more than one mechanism that produces

excessive trunk lean as part of its kinematic pattern, however, we expected that based on the scores subjects receive for these patterns, it should be possible to successfully classify the subjects into the eTL and nTL groups.

We postulated that if gait patterns can be found that satisfy these three criteria, then causative biomechanical relations can be established, that are indicative for the underlying origins of an excessive trunk lean. The absence of such patterns, in turn, would suggest other disease related deficits of the trunk control that are not correlated with other deviations in the kinematic movement pattern of CP gait.

In summary, the aim of the current study was to explore correlated patterns in the kinematic variables that characterize CP gait. We hypothesized that patterns can be found that satisfy the three postulated criteria, which would suggest that excessive trunk lean is part of a kinematic pattern caused by an underlying biomechanical mechanism. Understanding the mechanisms related to an excessive trunk lean and whether these are primary or compensatory could aid in the clinical decision-making and improve the management of CP.

MATERIALS AND METHODS

Patients with bilateral CP who were at least 5 years of age were retrospectively included from the database of the gait laboratory, which were measured between 2009 and 2018. Written consent was provided for research purposes by the patients and the study was approved by the ethics commission of the local medical association. Inclusion criteria was a Gross Motor Function Classification System (GMFCS) level I or II; accordingly, patients were able to walk freely without assistive devices or help. Excluded were obese patients according to the age-dependent body mass index thresholds suggested by the WHO (de Onis and Lobstein, 2010) and patients with documented spine deformities.

Patient data included instrumented 3D gait analysis data, where the kinematics had been captured with an 8-camera Vicon MX system at a sampling rate of 200 Hz. Joint angles were computed based on the Vicon Plug-In gait model including the trunk. All patients walked barefoot at self-selected speed along a 13 m walkway and at least 3 consistent step cycles needed to be present.

For the analysis mean angle waveforms of the upper body and of the lower limb of the more pronounced side (greater lateral trunk lean) were processed. In detail, 10 angles were included: trunk lean and the 9 angles of the Gait Profile Score (GPS) of the lower extremity (Baker et al., 2009) (seen in **Figures 1**, **2**), that represent clinically relevant joint angles in patients with CP. Furthermore, the GPS was facilitated as a descriptive measure to assess severity differences.

For further analysis the patients were divided into patients with and without an excessive lateral trunk lean (eTL and nTL, respectively), where "excessive" was defined as the trunk lean range of motion (ROM) angle exceeding 3 SD from a typically developed norm collective (TD; $n = 24$). The ROM was chosen over the maximum lateral trunk lean to account for spine deformities resulting in a constant lateral side bending.

FIGURE 1 | Illustration of the 10 analyzed angles. Angles referring to the global reference frame are indicated by solid and dashed lines. Other angles are referenced to the more proximal segment. obl, obliquity; rot, rotation; abd/add, abduction/adduction; flx, flexion; prog, progression.
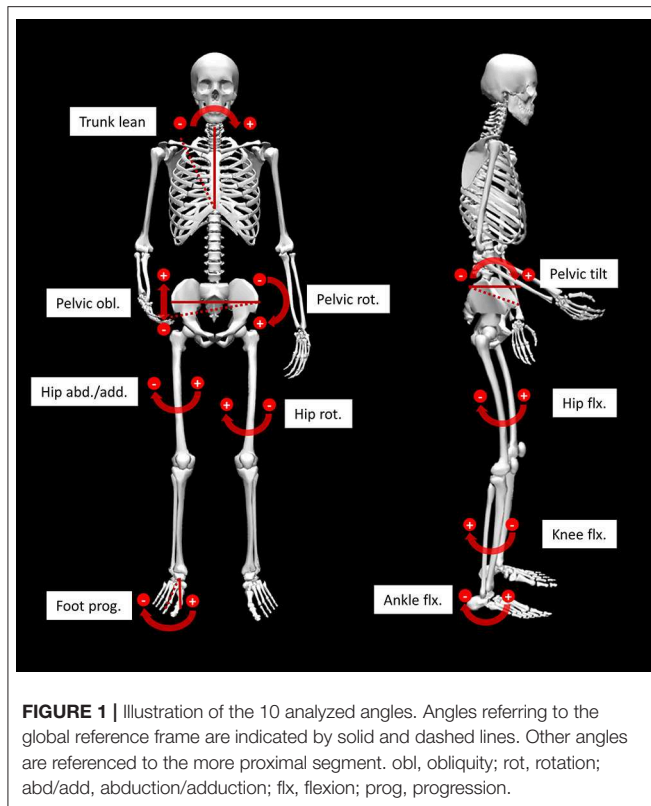
The 3SD cutoff was chosen based on the small TD trunk lean standard deviation and a lower cutoff would be hardly visible. The resulting threshold to classify for the eTL group was a lateral trunk lean ROM exceeding 8.4° (TD ROM + 3SD).

To identify patterns of correlated deviations from average kinematic angle waveforms, the principal component analysis (PCA) was facilitated. The PCA has already been successfully used in different contexts in CP, for example, to identify CP-related gait pathologies (Carriero et al., 2009a,b), to study multi segmental gait deviations (Zago et al., 2017) or to evaluate post-operative changes after multilevel surgery (Steppacher et al., 2018).

The input data for the PCA were the time normalized angle waveforms, each angle consisting of 101 discrete points, concatenated to a vector of 1,010 columns. Each gait cycle of every patient contributed a new row for the PCA input matrix. This approach has two advantages: firstly, it allows the systematic identification of kinematic gait patterns [principal components (PCs)] within and between the 10 angle waveforms, and secondly, a score is generated for each patient, expressing the coincidence of the patients movement pattern with each PC (Daffertshofer et al., 2004; Federolf et al., 2013). The individual PC-scores can then be used to further investigate the group differences between patients exhibiting an excessive trunk lean and patients who are not.

As part of the further investigation of the PC-scores, t-tests were used to identify group differences within the kinematic gait patterns and the standardized mean differences effect size

Hedges's g (d) with 95% confidence interval (CI) were calculated. Due to multiple testing, the alpha error was conservatively adapted with the Bonferroni correction (alpha level $p < 0.0025$). Further, the PC-scores with significant group differences were facilitated to investigate classification rates using binary logistic regressions. This last step was used to verify that the identified kinematic patterns are distinctive for a lateral trunk lean mechanism since this would result in high classification rates. We considered the third criterion of the compensatory hypothesis for excessive trunk lean to be satisfied, when the classification rates exceeded the proportional chance criterion (PCC) for logistic regressions. For the analysis MATLAB (MathWorks Inc., Natick, USA) was used for the PCA and group comparisons and SPSS Statistics (IBM Corp., USA) for the binary logistic regression. For the skeletal joint angle visualization (**Figure 1**) OpenSim was facilitated (Delp et al., 2007; Seth et al., 2018).

RESULTS

For the current study $n = 255$ patients met the inclusion criteria (mean age 13.6 ± 6.6 years, 155 males 100 females). Of these patients $n = 174$ (68%) exhibited an excessive lateral trunk lean within their gait pattern while $n = 81$ (32%) stayed below the cutoff of 8.4° trunk lean ROM (TD ROM + 3SD). In terms of anthropometrics the two groups were very similar (**Table 1**). As for the functional status, the excessive trunk lean group (eTL) had a larger proportion of patients rated as GMFCS II (82.2%), compared to 64.2% in the normal trunk lean group (nTL). This was also reflected by the GPS, where the eTL showed higher deviations from the norm. The differences in GPS were highly significant ($p < 0.001$) when comparing the eTL and nTL groups.

Joint angle waveforms averaged over the two groups are displayed in **Figure 2**. The mean maximal trunk lean ROM was 16.68 ± 6.4 in the eTL, nTL 6.63 ± 2.8 and 4.18 ± 1.4 in TD.

Within the between-patient variations in the kinematic variables the PCA identified correlated and mutually orthogonal patterns (**Figure 3**). The lower-order patterns, specifically PC-vectors 1 and 2 represented CP-gait deviations that did not change over the gait cycle, such as internal rotation malalignment of the lower extremity (PC1) and crouch gait (PC2). Higher-order PCs increasingly represented phase-dependent systematic gait deviations, including the trunk lean (PC8, PC12, PC17). More specifically, PC1, accounting for 31% of variability, showed a kinematic pattern dominated by the rotational malalignments of the lower extremity, expressed particularly as correlated deviations in hip rotation and foot progression. The contribution of these angle waveforms is visualized by the bar plots in **Figure 3**, showing that 63% in contributed by hip rotation and foot progression. The positive correlation between hip rotation and foot progression is visualized by the Eigenvector graph being negative in both cases. This means that the foot progressing increases with increased hip rotation or vice versa. Whether the correlation is positive or negative can also be seen in the right column mean angle visualization indicated by the colored lines. Is the colored line in both cases above or below the mean angle, the correlation is positive. Is the colored line opposite

FIGURE 2 | Angle waveforms averaged for the excessive eTL (excessive trunk lean) and nTL (trunk lean < 3SD of norm) groups of CP patients. In gray: typical developed ± 1SD. obl, obliquity; rot, rotation; abd/add, abduction/adduction; flx, flexion; prog, progression.
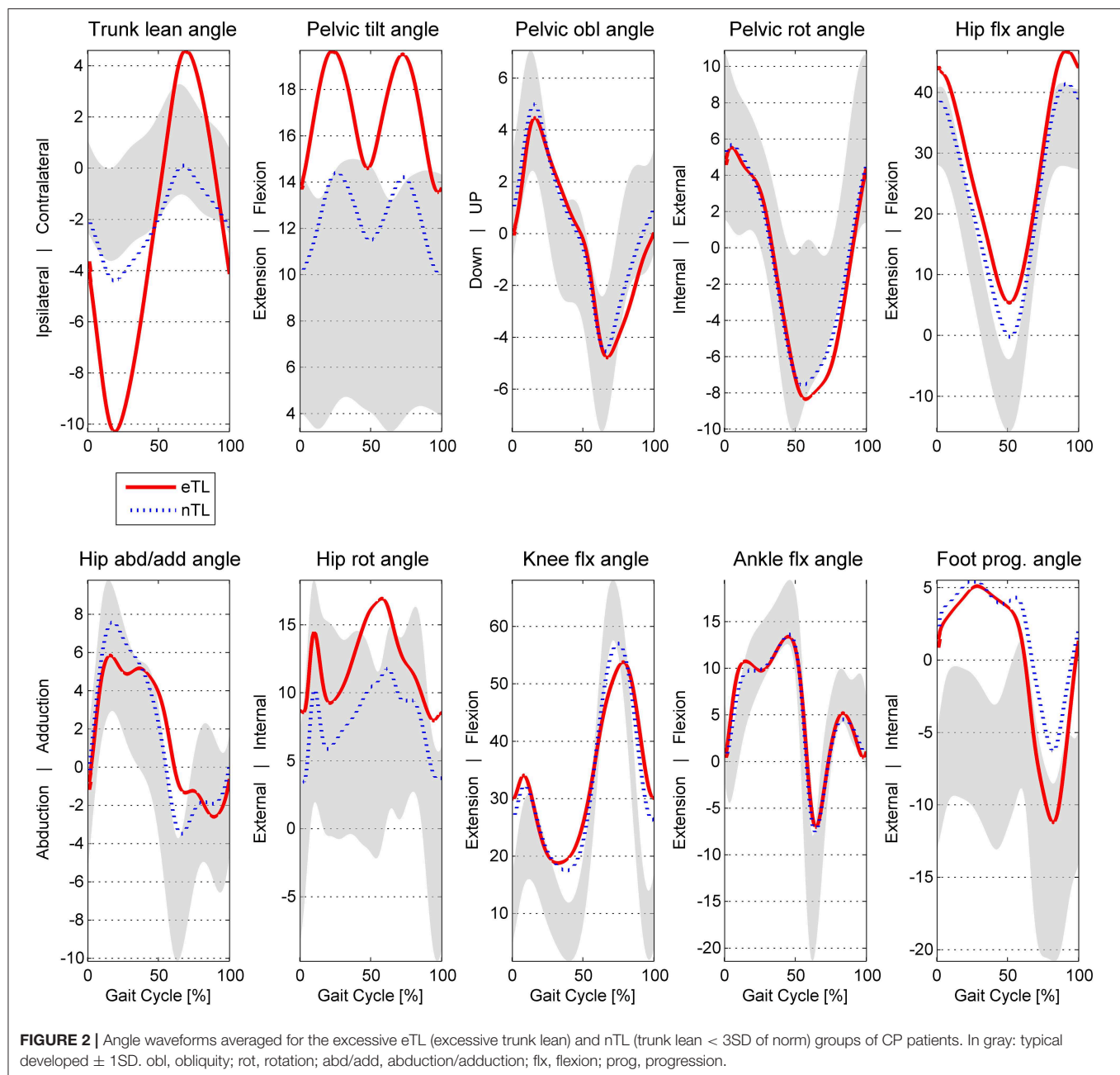
TABLE 1 | Anthropometrics and severity.

Group characteristics	eTL (*n* = 174)	nTL (*n* = 81)	*p*-value
Age (years)	13.6 (6.7)	13.6 (6.4)	0.80
Height (cm)	148 (19)	148 (16)	0.62
BMI (kg/m^2)	18.4 (3.3)	18.5 (3.4)	0.99
GPS (°)	9.4 (2.1)	7.9 (2.8)	**<0.001**
GMFCS I (%)	17.8	35.8	**<0.001***
GMFCS II (%)	82.2	64.2	

$^*\chi^2$-test.
eTL, excessive trunk lean; nTL, non-excessive trunk lean.
Significant group differences in bold.

sides, the correlation is negative. PC2 depicted a combination of hip and knee flexion, coupled with hip rotation and foot progression pronounced during swing phase. PC3 expressed between-patient variations in pelvic tilt coupled with variations in hip and knee flexion.

Within the first 20 PCs, 5 kinematic patterns with significant differences ($p < 0.0025$) between the scores of the eTL and nTL groups were identified (**Figure 4**). The trunk lean was distinctive in 4 of these kinematic patterns (PC: 3, 8, 12, 17), whereas PC2, while being significantly different between the eTL and nTL groups, did not entail the trunk lean within its kinematic pattern (**Figure 3**). Overall, the first 20 PCs explained 96% of the
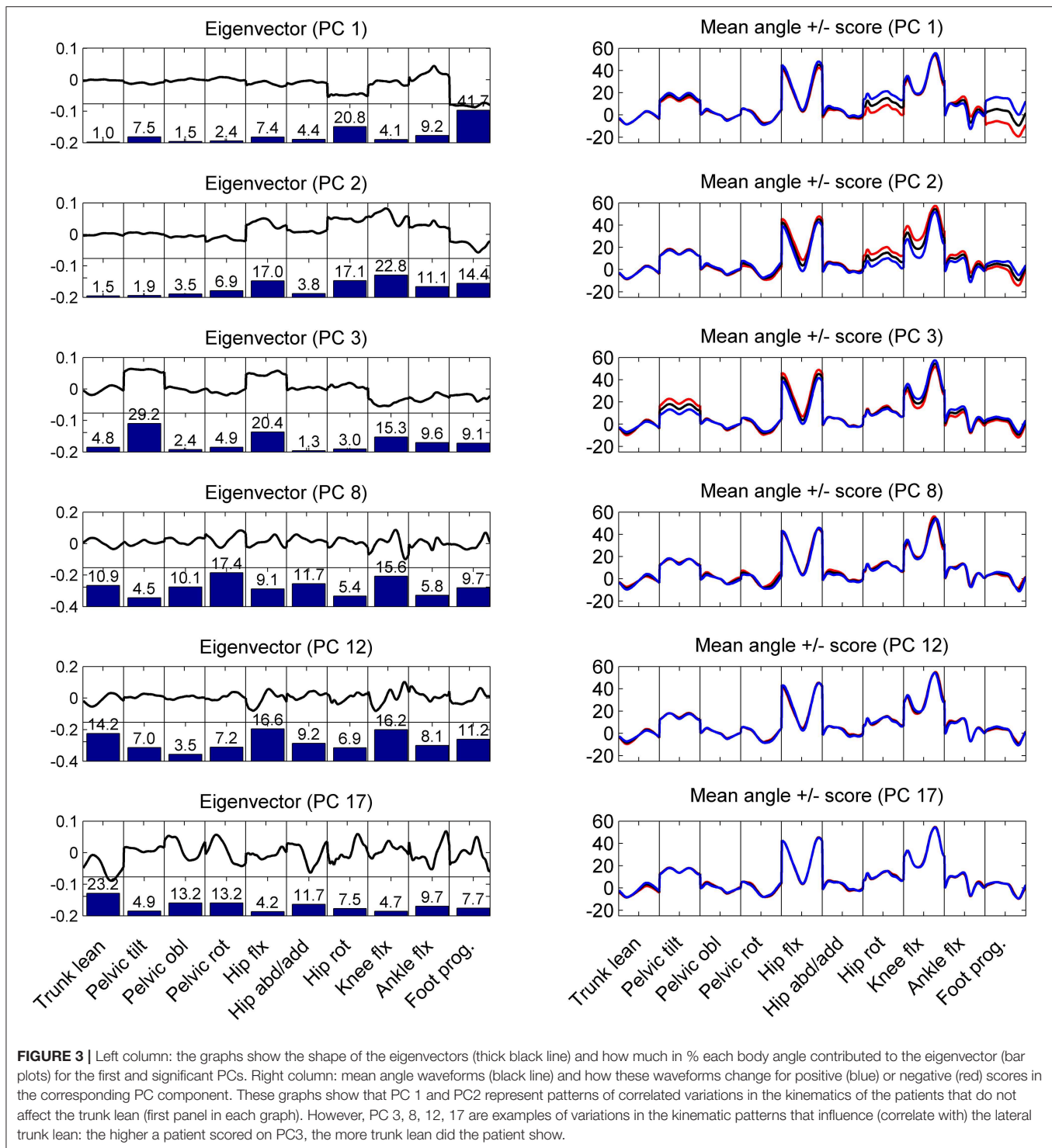
FIGURE 3 | Left column: the graphs show the shape of the eigenvectors (thick black line) and how much in % each body angle contributed to the eigenvector (bar plots) for the first and significant PCs. Right column: mean angle waveforms (black line) and how these waveforms change for positive (blue) or negative (red) scores in the corresponding PC component. These graphs show that PC 1 and PC2 represent patterns of correlated variations in the kinematics of the patients that do not affect the trunk lean (first panel in each graph). However, PC 3, 8, 12, 17 are examples of variations in the kinematic patterns that influence (correlate with) the lateral trunk lean: the higher a patient scored on PC3, the more trunk lean did the patient show.

variability between patients. The PC components that differed between groups together explained 34% of the variability. The 4 kinematic patterns that contained trunk lean as part of their pattern, as visible in PC 3, 8, 12, 17 in **Figure 3**, accounted for 14% of the between-patient variability. However, although highly significant group differences were found within the scores of the

different kinematic patterns (PCs) a substantial overlap of the score distributions of the eTL and nTL groups were observed (**Figure 4**), corresponding to small effect sizes for the group differences (**Table 2**).

Individually, none of the PCs had sufficient predictive power in a logistic regression model (**Table 3**)—particularly
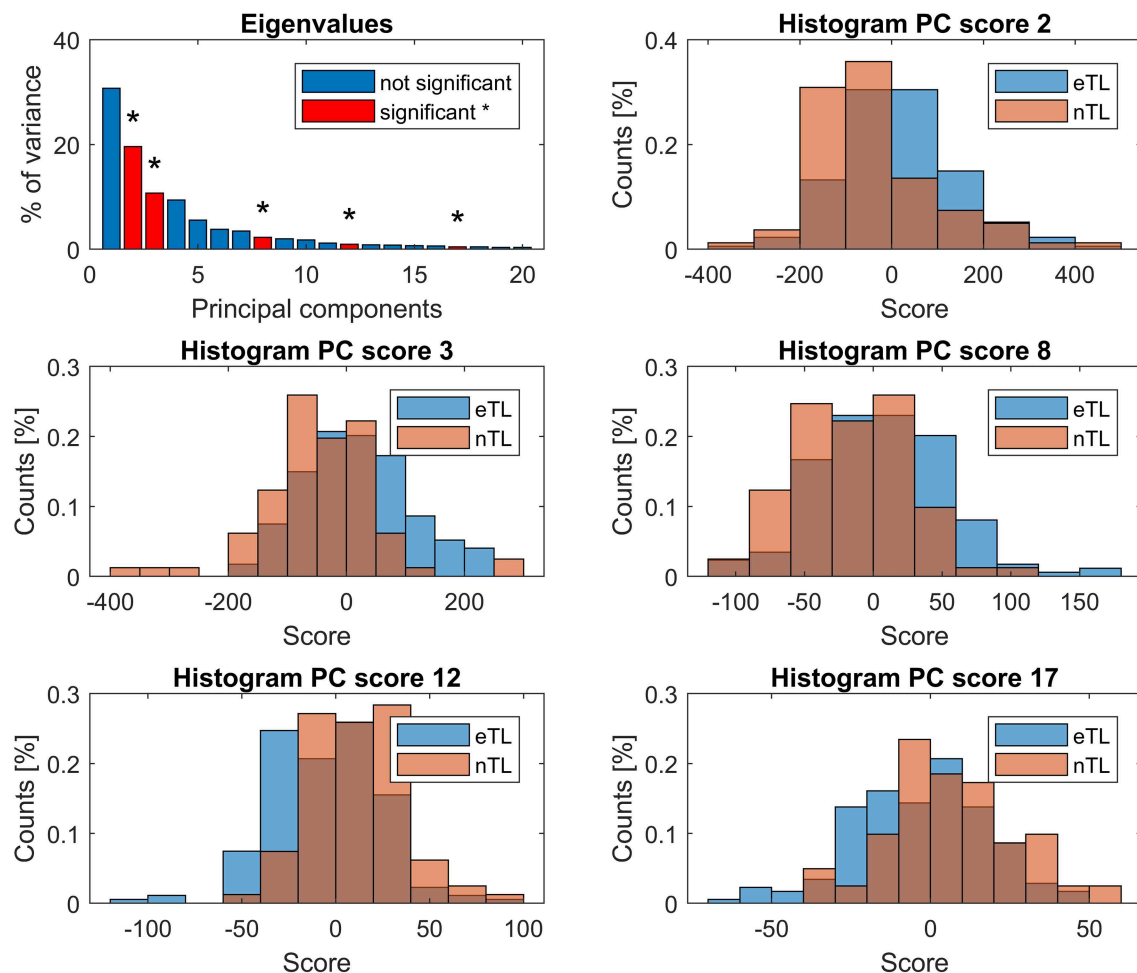
FIGURE 4 | First panel: eigenvalues, showing the explained variance by each PC. PCs whose scores differed significantly ($p < 0.0025$) between the eTL and nTL groups are displayed in red and marked with *. Remaining panels: histograms showing the group score distribution and their overlap.

TABLE 2 | Overview of the patterns (PCs) with significant score group differences.

PCs	P-value	Effect size (ES) (d)	ES 95% CI	
			Lower	Upper
2	<0.001	0.20	0.08	0.33
3	<0.001	0.29	0.18	0.40
8	<0.001	0.24	0.13	0.35
12	<0.001	0.28	0.17	0.38
17	<0.001	0.22	0.10	0.33

nTL patients were falsely classified at a very high rate. The binary logistic regression model combining PC 2, 3, 8, 12, and 17 revealed a significant association between the lateral trunk lean and the 5 significant PC scores [$X^2_{(5)} = 96.196$, $p < 0.001$] and explained 44% of variance (Nagelkerke R^2). The overall predictability of the model was 81.2% (**Table 3**), however, whereas eTL patients were correctly predicted by the

logistic regression at a rate of 89.7%, nTL patients were predicted correctly at rates of 63.0%. For the given group sizes—n(eTL) = 174; n(nTL) = 81—the proportional chance criterion (PCC), necessitates a nTL classification rate of at least a rate of 72% for an acceptable accuracy of an 25% improvement over a by-chance classification.

DISCUSSION

The current study analyzed the between-patient variability in CP gait patterns with a focus on lateral trunk lean. We hypothesized that excessive trunk lean might be part of a functional kinematic mechanism, and postulated that then we should find patterns in which trunk lean is correlated with other deviations in kinematic variables (criterion 1), in which patient groups with excessive (eTL) or normal trunk lean (nTL) would score significantly and substantially (effect size) different (criterion 2), and which would together allow to reliably classify individual patients as belonging to the eTL or nTL group (criterion 3). The PCA

TABLE 3 | Logistic regression models for the individual significant scores and for the combined model, with odds ratios (OR) and classification results for the excessive (eTL) and non-excessive (nTL) trunk lean groups.

PCs	P-value	OR (95% CI)	Classification rate (%)	
			eTL	nTL
2	0.002	0.995 (0.993–0.998)	97.7	3.7
3	<0.001	1.010 (1.006–1.014)	93.7	14.8
8	<0.001	0.982 (0.973–0.990)	96.0	14.8
12	<0.001	1.032 (1.018–1.046)	95.4	12.3
17	0.001	0.968 (0.953–0.984)	96.6	11.1
Combined model		Overall	89.7	63.0
2 + 3 + 8 + 12 + 17		Combined	81.2	

analysis conducted in the current study successfully identified several patterns (PC 3, 8, 12, 17) in which trunk lean was correlated to changes in other kinematic variables (criterion 1), and we found that subject groups scored significantly different on these four PCs (criterion 2a). However, we found that the score distributions of the two patient groups on all four PCs largely overlapped, corresponding with small effect sizes for the group differences (criterion 2b not satisfied). Furthermore, we found that the logistic regression model—despite yielding a significant association between the PC scores and excessive trunk lean—was not reliably able to classify patients into the correct group. Particularly patients with a normal trunk lean were to a high percentage falsely classified as eTL patients (criterion 3 not satisfied). Overall, these findings suggest that an excessive lateral trunk lean in CP is not based on a kinematic compensatory mechanism but, in a large fraction of the patients, more likely the result of other motor functional deficits (Panibatla et al., 2017).

Interestingly, no significant differences between eTL and nTL were found within the first kinematic pattern (PC1) entailing rotational malalignments of the lower extremity. Therefore, it can be concluded that lateral trunk lean mechanics are unaffected by these rotational malalignments. This finding stands in contrast with the assumption that hip internal rotation is a compensatory mechanism for hip muscle lever arm dysfunctions of the often anteverted hip in CP (Arnold et al., 1997). Since hip muscle weakness accounts, at least to some degree, for the prevalence of a trunk lean pattern (Krautwurst et al., 2013), a connection between hip internal rotation and trunk lean would have been plausible but could not be established in the current study. Another study found a significant negative correlation between hip rotation and foot rotation (Gaston et al., 2011) and propose the internal hip rotation as result of a distal foot external rotation. While the PC1 pattern contrarily shows a positive correlation of hip rotation and foot progression, the PC2 pattern shows the, by Gaston et al. proposed, functional relation of internal hip rotation and foot external rotation, combined with crouch gait characteristics, due to the lever arm dysfunction of the plantar-flexion knee-extension couple (Sangeux et al., 2015). The fact that the PCA revealed two different rotational patterns, the presence of two individual mechanisms is likely.

In the 2nd pattern (PC2) changes in trunk lean angle was not part of the PC-vector, however, the eTL and nTL groups scored significantly different when projected onto this pattern. The observation that excessive trunk lean was not part of this pattern, implies that more severely affected patients were more likely to also show excessive trunk lean (Attias et al., 2015; Swinnen et al., 2016), without excessive trunk lean itself being correlated to the PC2 gait variables. Other studies also found an increased trunk lean with increasing impairment (Attias et al., 2015), which is in line with our findings of more severe gait deviations within the eTL group expressed by the highly significant GPS differences. Additionally, the proportion of patients rated GMFCS II was about 20% higher in the trunk lean group, which further corroborates the severity explanation and was also observed in other studies (Swinnen et al., 2016).

The other 4 kinematic patterns (PC 3, 8, 12, 17) that differed significantly between eTL and nTL groups did not appear to show clear functional mechanisms. This could be partly due to the fact that the trunk lean strategy appears to be present in a variety of gait pathologies. The heterogeneous patient groups with diverse combinations of different gait pathologies is likely to result in multiple patterns, of which some will also contain trunk lean as part of their pattern. However, this does not imply a causative nature of the trunk lean, describing functional patterns. What can be said is that patients showing a particular combination of angle deviations (specific for the pattern), usually also exhibited an excessive lateral trunk lean. Since the identified patterns do not entail clear functional mechanisms, such as relations between frontal plane trunk, pelvis and hip kinematics, it seems plausible to conclude that neither a universal trunk lean mechanism exists, nor that a specific trunk lean strategy exists that is attributable to certain CP related gait pathologies. In a clinical sense, these findings support a multifactorial cause of a lateral trunk lean, implying that there is no single solution for addressing or correcting excessive trunk lean.

Limitations

To provide further insight into how the results of the current study can be interpreted, some reservations should be mentioned and kept in mind.

One of these limitations is the circumstance that the gait speed was self-selected and different velocities result in altered angle patterns (Schwartz et al., 2008), including altered trunk kinematics (Thummerer et al., 2012). Despite the gait speed being not significantly different between the groups, it may still have some effect on the PCA results. PCA has been shown to be able to detect running speed differences (Maurer et al., 2012).

Arm movement was not measured. Arm movements in CP, however, may influence the trunk kinematics and might entail further information for the prevalence of an excessive trunk. Children with CP often show flexed elbow positions and increased shoulder abduction, which might be caused by spasticity but is also thought as compensatory strategy for balance and guarding purposes (Galli et al., 2014). Despite these general influences of arm movements on posture, the measured trunk lean should be largely unaffected, since the Plug-In gait model

uses only the thorax markers, without the shoulder markers to calculate the trunk lean.

Furthermore, excessive trunk lean was defined as exceeding 3SD from norm. This definition is not based on a clinical classification of a certain degree of trunk lean being pathological. Hence, defining a meaningful cutoff will require further research and insight into the underlying mechanisms of excessive lateral trunk lean in CP.

CONCLUSION

The PCA was able to identify kinematic patterns that were significantly related to the lateral trunk lean based on the group differences. However, despite these findings, a clear kinematic mechanism leading to excessive trunk lean was not found. The current study does not provide conclusive evidence against a kinematic compensatory mechanism. However, the absence of such patterns makes it more likely that excessive lateral trunk lean in CP could be the result of disease related motor functional deficits. More research is necessary to clarify this issue. Our study does provide evidence that rotational malalignments present independently of the trunk lean and that the prevalence of an excessive lateral trunk lean in CP depends on the severity of the disease.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. Written consent included only the instituional use of the data.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Arnold, A. S., Komattu, A. V., and Delp, S. L. (1997). Internal rotation gait: a compensatory mechanism to restore abduction capacity decreased by bone deformity. *Dev. Med. Child Neurol.* 39, 40–44.

Attias, M., Bonnefoy-Mazure, A., Lempereur, M., Lascombes, P., de Coulon, G., and Armand, S. (2015). Trunk movements during gait in cerebral palsy. *Clin. Biomech.* 30, 28–32. doi: 10.1016/j.clinbiomech.2014.11.009

Baker, R., McGinley, J. L., Schwartz, M. H., Beynon, S., Rozumalski, A., Graham, H. K., et al. (2009). The gait profile score and movement analysis profile. *Gait Posture* 30, 265–269. doi: 10.1016/j.gaitpost.2009.05.020

Carriero, A., Zavatsky, A., Stebbins, J., Theologis, T., and Shefelbine, S. J. (2009a). Correlation between lower limb bone morphology and gait characteristics in children with spastic diplegic cerebral palsy. *J. Pediatr. Orthop.* 29, 73–79. doi: 10.1097/BPO.0b013e31819224d

Carriero, A., Zavatsky, A., Stebbins, J., Theologis, T., and Shefelbine, S. J. (2009b). Determination of gait patterns in children with spastic diplegic cerebral palsy using principal components. *Gait Posture* 29, 71–75. doi: 10.1016/j.gaitpost.2008.06.011

Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., and Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clin. Biomech.* 19, 415–428. doi: 10.1016/j.clinbiomech.2004.01.005

de Onis, M., and Lobstein, T. (2010). Defining obesity risk status in the general childhood population: which cut-offs should we use? *Int. J. Pediatr. Obes.* 5, 458–460. doi: 10.3109/17477161003615583

Delp, S. L., Anderson, F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., et al. (2007). OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE Trans. Biomed. Eng.* 54, 1940–1950. doi: 10.1109/TBME.2007.901024

Eskofier, B. M., Federolf, P., Kugler, P. F., and Nigg, B. M. (2013). Marker-based classification of young-elderly gait pattern differences via direct PCA feature extraction and SVMs. *Comput. Methods Biomech. Biomed. Engin.* 16, 435–442. doi: 10.1080/10255842.2011.624515

Federolf, P. A., Boyer, K. A., and Andriacchi, T. P. (2013). Application of principal component analysis in clinical gait research: identification of systematic differences between healthy and medial knee-osteoarthritic gait. *J. Biomech.* 46, 2173–2178. doi: 10.1016/j.jbiomech.2013.06.032

Galli, M., Cimolin, V., Albertini, G., Piccinini, L., Turconi, A. C., Romkes, J., et al. (2014). Kinematic analysis of upper limb during walking in diplegic children with cerebral palsy. *Eur. J. Paediatr. Neurol.* 18, 134–139. doi: 10.1016/j.ejpn.2013.09.007

Gaston, M. S., Rutz, E., Dreher, T., and Brunner, R. (2011). Transverse plane rotation of the foot and transverse hip and pelvic kinematics in diplegic cerebral palsy. *Gait Posture* 34, 218–221. doi: 10.1016/j.gaitpost.2011.05.001

Heyrman, L., Feys, H., Molenaers, G., Jaspers, E., Monari, D., Nieuwenhuys, A., et al. (2014). Altered trunk movements during gait in children with spastic diplegia: compensatory or underlying trunk control deficit? *Res. Dev. Disabil.* 35, 2044–2052. doi: 10.1016/j.ridd.2014.04.031

Klum, E., Streicher, H., Böhm, H., Wagner, P., and Döderlein, L. (2015). Ursachen und Prädiktoren eines Duchenne-Hinkens bei Jugendlichen und jungen Erwachsenen mit Zerebralparese. *Z. Orthop. Unfall.* 153, 85–92. doi: 10.1055/s-0034-1396186

Krautwurst, B. K., Wolf, S. I., Heitzmann, D. W. W., Gantz, S., Braatz, F., and Dreher, T. (2013). The influence of hip abductor weakness on frontal plane motion of the trunk and pelvis in patients with cerebral palsy. *Res. Dev. Disabil.* 34, 1198–1203. doi: 10.1016/j.ridd.2012.12.018

Maurer, C., Federolf, P., Tscharner, V., von Stirling, L., and Nigg, B. M. (2012). Discrimination of gender-, speed-, and shoe-dependent movement patterns in runners using full-body kinematics. *Gait Posture* 36, 40–45. doi: 10.1016/j.gaitpost.2011.12.023

Opheim, A., Jahnsen, R., Olsson, E., and Stanghelle, J. K. (2009). Walking function, pain, and fatigue in adults with cerebral palsy: a 7-year follow-up study. *Dev. Med. Child Neurol.* 51, 381–388. doi: 10.1111/j.1469-8749.2008.03250.x

Panibatla, S., Kumar, V., and Narayan, A. (2017). Relationship between trunk control and balance in children with spastic cerebral palsy: a cross-sectional study. *J. Clin. Diag. Res.* 11, YC05–YC08. doi: 10.7860/JCDR/2017/28388.10649

Robertson, D. G. E., Caldwell, G. E., Hamill, J., Kamen, G., and Whittlesey, S. N. (2014). *Research Methods in Biomechanics*. Champaign, IL: Human Kinetics.

Salami, F., Niklasch, M., Krautwurst, B. K., Dreher, T., and Wolf, S. I. (2017). What is the price for the Duchenne gait pattern in patients with cerebral palsy? *Gait Posture* 58, 453–456. doi: 10.1016/j.gaitpost.2017.09.006

Sangeux, M., Rodda, J., and Graham, H. K. (2015). Sagittal gait patterns in cerebral palsy: the plantarflexor-knee extension couple index. *Gait Posture* 41, 586–591. doi: 10.1016/j.gaitpost.2014.12.019

Schmid, S., Schweizer, K., Romkes, J., Lorenzetti, S., and Brunner, R. (2013). Secondary gait deviations in patients with and without neurological involvement: a systematic review. *Gait Posture* 37, 480–493. doi: 10.1016/j.gaitpost.2012.09.006

Schwartz, M. H., Rozumalski, A., and Trost, J. P. (2008). The effect of walking speed on the gait of typically developing children. *J. Biomech.* 41, 1639–1650. doi: 10.1016/j.jbiomech.2008.03.015

Seth, A., Hicks, J. L., Uchida, T. K., Habib, A., Dembia, C. L., Dunne, J. J., et al. (2018). OpenSim: simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS Comput. Biol.* 14:e1006223. doi: 10.1371/journal.pcbi.1006223

Steppacher, R., North, D., Künzle, C., Lengnick, H., Klima, H., Mündermann, A., et al. (2018). Retrospective evaluation of changes in gait patterns in children and adolescents with cerebral palsy after multilevel surgery. *J. Child Neurol.* 33, 453–462. doi: 10.1177/0883073818766681

Swinnen, E., Goten, L. V., de Koster, B., and Degelaen, M. (2016). Thorax and pelvis kinematics during walking, a comparison between children with and

without cerebral palsy: a systematic review. *NeuroRehabilitation* 38, 129–146. doi: 10.3233/NRE-161303

Thummerer, Y., von Kries, R., Marton, M.-A., and Beyerlein, A. (2012). Is age or speed the predominant factor in the development of trunk movement in normally developing children? *Gait Posture* 35, 23–28. doi: 10.1016/j.gaitpost.2011.07.018

Zago, M., Sforza, C., Bona, A., Cimolin, V., Costici, P. F., Condoluci, C., et al. (2017). How multi segmental patterns deviate in spastic diplegia from typical developed. *Clin. Biomech.* 48, 103–109. doi: 10.1016/j.clinbiomech.2017.07.016

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Check for
updates

Sensor Data Required for Automatic Recognition of Athletic Tasks Using Deep Neural Networks

Allison L. Clouthier, Gwyneth B. Ross and Ryan B. Graham*

School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada

Movement screens are used to assess the overall movement quality of an athlete. However, these rely on visual observation of a series of movements and subjective scoring. Data-driven methods to provide objective scoring of these movements are being developed. These currently use optical motion capture and require manual pre-processing of data to identify the start and end points of each movement. Therefore, we aimed to use deep learning techniques to automatically identify movements typically found in movement screens and assess the feasibility of performing the classification based on wearable sensor data. Optical motion capture data were collected on 417 athletes performing 13 athletic movements. We trained an existing deep neural network architecture that combines convolutional and recurrent layers on a subset of 278 athletes. A validation subset of 69 athletes was used to tune the hyperparameters and the final network was tested on the remaining 70 athletes. Simulated inertial measurement data were generated based on the optical motion capture data and the network was trained on this data for different combinations of body segments. Classification accuracy was similar for networks trained using the optical and full-body simulated inertial measurement unit data at 90.1 and 90.2%, respectively. A good classification accuracy of 85.9% was obtained using as few as three simulated sensors placed on the torso and shanks. However, using three simulated sensors on the torso and upper arms or fewer than three sensors resulted in poor accuracy. These results for simulated sensor data indicate the feasibility of classifying athletic movements using a small number of wearable sensors. This could facilitate objective data-driven methods that automatically score overall movement quality using wearable sensors to be easily implemented in the field.

Keywords: human activity recognition, wearable sensors, machine learning, neural network, movement screens

INTRODUCTION

Movement screens are used to assess the overall movement quality of an athlete. Typically, the athlete will perform a series of movements while a trained rater visually observes and scores the movements. The goals of movement screens are to predict injury risk and identify performance deficits that can be targeted in training. While interrater and intrarater reliabilities for movement screens such as the Functional Movement Screen (FMS™) are good (Minick et al., 2010; Teyhen et al., 2012), interrater reliability for subtest components can be poor and dependent on rater experience (Smith et al., 2013; Gulgin and Hoogenboom, 2014; Bonazza et al., 2017). Furthermore, concerns have been raised that grading criteria can be somewhat ambiguous

(Frost et al., 2015; Bonazza et al., 2017) and scores may not be sensitive enough to detect movement abnormalities (Clifton et al., 2013). Recent work has aimed to develop objective scoring methods for movement screens (Ross et al., 2018). Data-driven approaches have the potential to improve the repeatability of scoring and increase the ability to detect subtle differences in movement patterns. However, current methods require manual processing of motion capture data before scoring can be performed, including cropping trials to isolate each movement. Additionally, the reliance on optical motion capture could be a barrier to implementation of these methods in the field.

Wearable sensors are an attractive alternative to optical motion capture for motion analysis applications. They are cost-effective and portable, allowing for the collection of motion data outside of a laboratory and over large capture volumes. Furthermore, wearable sensors have the potential to be less cumbersome than optical markers depending on the number and placement of sensors. Previous work investigated optimal placement and number of sensors to classify activities of daily living (Pannurat et al., 2017), everyday activities (Kern et al., 2003; Olguin and Pentland, 2006; Atallah et al., 2011; Cleland et al., 2013), and fall detection (Gjoreski et al., 2011). However, which sensors are necessary to best classify movement screening tasks remains unclear.

Human activity recognition is an area of research that seeks to automatically identify human activities by applying machine learning techniques to motion data. Methods have been developed to classify movements including hand gestures (Kim and Toomajian, 2016), activities of daily living (Hammerla et al., 2016), and movements typical in various sports (Nguyen et al., 2015; Kautz et al., 2017). Previously, activity recognition methods employed techniques that required hand-selected features as input (Bulling et al., 2014). However, convolutional neural networks (CNNs), a type of deep neural network (DNN), are now commonly used to automatically generate features through deep learning (Zeng et al., 2014; Yang et al., 2015; Lee et al., 2017). CNNs have shown promising results in activity recognition; however, they are unable to capture time dependencies. Recurrent neural networks are a type of neural network that include a memory component that allows them to model temporal dependencies. The combination of CNNs to extract features with long-short-term memory (LSTM) recurrent networks to capture temporal dependences has provided improved classification performance over CNNs alone (Ordóñez and Roggen, 2016).

The use of deep neural networks in movement screens would allow for a continuous data collection during a movement screen. Individual movements could then be automatically identified and segmented as a preparation for further analysis or scoring. This would decrease the manual effort required for the analysis process and increase the utility of these objective measurement techniques. The ability to perform the movement classification and scoring based on data from a minimal set of wearable sensors would further increase the applicability of data-driven movement screens. Therefore, our first aim was to use a deep neural network to identify when movements typical of movement screens occur within motion data. Our second aim was to compare networks trained using optical motion capture data with those trained using data available from wearable sensors.

METHODS

Data Collection and Processing

Optical motion capture data were collected from 417 athletes performing a series of movement tests by Motus Global (Rockville Center, NY). The athletes competed in a variety of

FIGURE 1 | Architecture of the deep neural network used to classify athletic movements. The network combines convolutional and recurrent layers (Ordóñez and Roggen, 2016). Tensor sizes and function inputs based on the OPT data and final architecture parameters are shown. PyTorch functions and inputs are shown for each layer. SWS, sliding window size; C, number of CNN channels; N, number of columns in the input data; k, CNN kernel size; L, LSTM cells; Nclasses, number of movements classified.

sports, including baseball, basketball, soccer, golf, tennis, track and field, squash, cricket, lacrosse, football, and volleyball. They ranged in skill level from recreational athletes to those playing in major professional sports leagues (e.g., NBA, MLB, PGA, etc.). Participants provided informed consent for future use of their data for research before completing the protocol. The secondary use of the data was approved by the University of Ottawa Research Ethics Board. Forty-five retroreflective markers were placed on the athlete for motion tracking (Ross et al., 2018) and data were recorded at 120 Hz using an eight-camera Raptor-E (Motion Analysis, Santa Rosa, CA) motion capture system. Each athlete performed a series of movement tests consisting of 21 unique movements. The 13 movements most likely to challenge mobility and stability were selected for analysis in this study, including hop down right/left (HDR, HDL), bird-dog right/left (BDR, BDL), drop jump (DJ), T-balance right/left (TBR, TBL), step-down right/left (SDR, SDL), L-hop right/left (LHR, LHL), and lunge right/left (LR, LL) (Ross et al., 2018). Individual trials were collected for each movement.

Start and end time points were manually identified for each trial (Ross et al., 2018) for use as a ground truth of when each activity was performed. The optical motion data used in the analysis (OPT) included global x, y, z coordinates for 45 markers. To simulate data that can be obtained using inertial measurement units (sIMU), marker trajectories were processed in Visual3D (C-Motion, Inc., Germantown, MD) and global angular orientation Euler angles and the Euclidean norm of the center of mass linear acceleration and angular velocity for each body segment were calculated. The Euclidean norm of the velocity and acceleration

was used to reduce the reliance on accurate sensor alignment. Accelerations and velocities were low-pass filtered at 15 Hz with a zero-lag second order Butterworth filter.

Deep Neural Network

Athletes were randomly separated into training (67%, $n = 278$), validation (33%, $n = 69$), and test (33%, $n = 70$) subsets. A single matrix was created for each subset by concatenating data from all movement trials performed by all athletes in the subset. Each variable was normalized by subtracting the mean and dividing by the standard deviation of all data frames across athletes and movements in the training set for that variable. A sliding window approach was used to divide the subset data into data segments containing an equal number of data frames. The stride for the sliding window was 1/4 the window size. Each data segment was

TABLE 1 | Learning and architecture parameter values tested for hyperparameter tuning.

SGD optimizer parameter tuning		
Learning Rate	0.0001, 0.001, 0.01, 0.1, 1	
Momentum	0.5, 0.7, 0.9, 0.95, 0.98	
DNN parameter tuning		
Window Size (frames)	24	48
CNN Kernel Size (frames)	5, 6	6, 8
CNN Channels	32, 64, 96	
LSTM Cells	64, 128, 192	

TABLE 2 | Combination of body segments used to train and test the DNN for the simulated IMU data.

Data input	Body segments
sIMU1	Torso
sIMU2	Torso, pelvis
sIMU3L	Torso, shanks (lower body)
sIMU3U	Torso, upper arms (upper body)
sIMU4	Torso, pelvis, thighs
sIMU4D	Forearms, shanks (distal segments)
sIMU4P	Upper arms, thighs (proximal segments)
sIMU5	Torso, forearms, shanks
sIMU13	Head, torso, pelvis, upper arms, forearms, thighs, shanks, feet
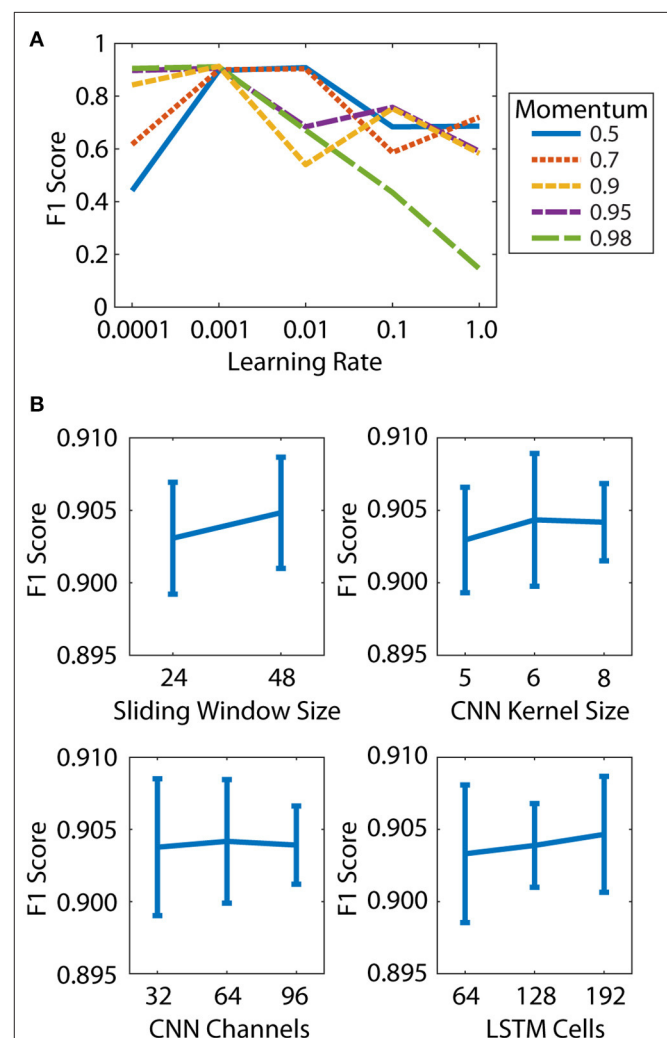
FIGURE 2 | Hyperparameter tuning was performed in two steps: learning parameters **(A)** and architecture parameters **(B)**. **(A)** Effect of learning rate and momentum on micro-averaged F1 score. **(B)** Effect of sliding window size, CNN kernel size, CNN channels, and LSTM cells on micro-averaged F1 score. Mean and standard deviation of all DNNs at each parameter level are shown.

assigned a label according to the movement that was performed for the majority of the data segment. A "Null" label was included to describe times when none of the movements were being performed for a total of 14 classes.

A deep neural network based on the work of Ordóñez and Roggen (2016) was implemented in PyTorch (Paszke et al., 2017). The architecture combines convolutional layers to extract features with recurrent layers to model the temporal dynamics. The network includes four convolutional layers, two long-short-term memory (LSTM) recurrent layers, a linear fully connected layer, and a softmax classifier (**Figure 1**). The input to the network is the windowed time series data. The length of the input data was the sliding window size and the number of columns depended on the data used: 3*45 for OPT (x, y, and z component of each trajectory) and 5 * number of body segments for the sIMU data (3 Euler angles + 1 angular velocity norm + 1 linear acceleration norm).

For network training, a mini-batch size of 100 was used. A stochastic gradient descent (SGD) optimizer with momentum was used for training with a cross-entropy loss criterion. The DNN was trained to classify the movement performed during a given windowed data segment.

Hyperparameter Tuning

Hyperparameter tuning was performed using a grid search with the validation set of the optical motion data (OPT). The learning parameters were tuned first as these have a larger impact on classifier performance (Hammerla et al., 2016). The learning parameters were the learning rate and momentum of the SGD optimizer. Five values of each were explored (**Table 1**)

resulting in 25 DNNs trained on the OPT training set. The micro-averaged F1 score was calculated for the validation set to assess the performance of each DNN. The F1 score is a measure of classification accuracy that is the harmonic mean of precision and recall ($F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$). The micro-averaged F1 score calculates the mean across the classes by considering all individual predictions, which is suitable for classes of different sizes. The micro-averaged F1 score is equivalent to the micro-averaged precision, micro-averaged recall, and classification accuracy.

After selecting the learning rate and momentum that produced the best F1 score, the architecture parameters were tuned. Two to three values were tested for each of the following parameters: sliding window size, CNN kernel filter size, CNN channels, and LSTM cells (**Table 1**). Note that CNN kernels of size 5 and 6 were used with window size 24 and CNN kernels of size 6 and 8 were used with window size 48. Models were assessed based on the micro-averaged F1 score.

Comparison of Simulated IMU Sensor Data

Once the final learning and architecture parameters were determined, the final model was used to identify movements in the test set. In this case, the DNN was used individually on each athlete. All trials of athlete's data were combined and then segmented using sliding windows and the DNN was used to classify each window. Then for each frame of data, the class probabilities from each window containing that frame were averaged, and a final classification was made for that frame of data.

FIGURE 3 | (A) Micro-averaged F1 score for DNNs trained using simulated IMU (sIMU) data from various combinations of body segments. **(B)** F1 score for each movement for a selection of DNNs trained on sIMU data. Scores were calculated on the test set based on classification of individual data frames. Movements are HDR/L, hop down right/left; BDR/L, bird-dog right/left; SDR/L, step-down right/left; LHR/L, L-hop right/left; DJ, drop jump; LR/L, lunge right/left; TBR/L, T-balance right/left.

DNNs using the final learning and architecture parameters were also trained for the simulated IMU data on the training subset. Different combinations of simulated sensor locations were examined (**Table 2**). DNNs were evaluated on the test subset following the procedure outlined above. For each DNN, the confusion matrix, accuracy, precision, recall, and F1 score were calculated. Micro and macro averages and metrics for each class were produced.

RESULTS

Hyperparameter Tuning

The learning rate and momentum of the SGD optimizer both had a large effect on the micro-averaged F1 score for the OPT validation set (**Figure 2A**). The best F1 score was obtained for a learning rate of 0.001 and momentum of 0.9, and these values were used for all subsequent models. The DNN parameters had a relatively small effect on the F1 scores, with values ranging from 0.895 to 0.911 (**Figure 2B**). The best results were obtained for a sliding window size of 48 (0.04 s), CNN kernel size of 6 frames, 32 CNN channels, and 64 LSTM cells. These parameters were selected for use in the final DNN.

Comparison of Simulated IMU Sensor Data

Deep neural networks trained using optical data (OPT) and all 13 body segments of sIMU data (sIMU13) had similar micro-averaged F1 scores (0.901 and 0.902, respectively). In general, including more body segments improved performance (**Figure 3**), although only small improvements were obtained by including more than four body segments. Bird-dog (BDR/L) movements were predicted well (F1 score > 0.76) for all networks, while drop jumps (DJ) tended to be more poorly identified in general.

The effect of including upper or lower limb data can be observed in the confusion matrices for the sIMU3U and sIMU3L models (**Figure 4**). With the torso and upper arms included (sIMU3U), the DNN frequently confuses left and right versions of tasks. Tasks involving jumping were also confused. The network using the torso and shanks (sIMU3L) is better able to distinguish between left and right, but occasionally confuses the T-balance (TBR/L) and lunge tasks (LR/L). L-hops (LHR/L) are sometimes classified as hop downs (HDR/L) in both three-segment networks (sIMU3L, sIMU3U).

The true and predicted movements over time for the OPT, sIMU1, sIMU3L, and sIMU13 models are shown in **Figure 5** for a representative athlete. OPT, sIMU13, and sIMU3L were better able to predict the entire duration of movements. Networks with fewer body segments tended to switch between predictions. The misclassification between movements and Null largely occurs at the beginning and end of a movement.

The complete set of precision, recall, F1 scores, and confusion matrices are included in the **Supplementary Material.**

DISCUSSION

The deep neural network (DNN) combining convolutional and recurrent layers was able to successfully identify athletic
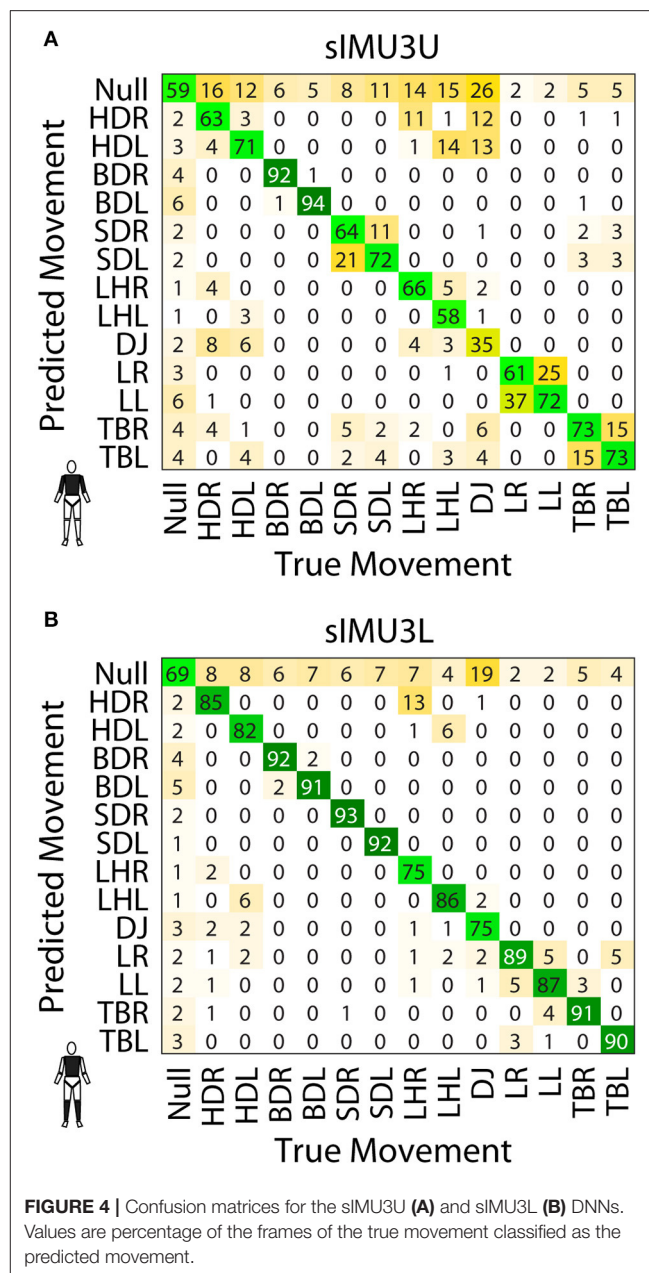
FIGURE 4 | Confusion matrices for the sIMU3U **(A)** and sIMU3L **(B)** DNNs. Values are percentage of the frames of the true movement classified as the predicted movement.

movements for both optical motion capture trajectories and simulated inertial measurement unit (sIMU) data. DNNs trained using optical motion capture data (OPT) and full body simulated IMU (sIMU13) data had similar performance with F1 scores of approximately 0.90. Classification accuracy was poor (<70%) if fewer than three body segments were included or the lower limbs were not included in the sIMU data.

There was minimal difference between micro-averaged F1 scores for the DNNs trained using five or 13 body segments. This indicates that it is not necessary to include measurements from the head or more than one segment from each upper or lower limb. This is encouraging as the use of fewer sensors would simplify the set-up before a movement screen and would be less
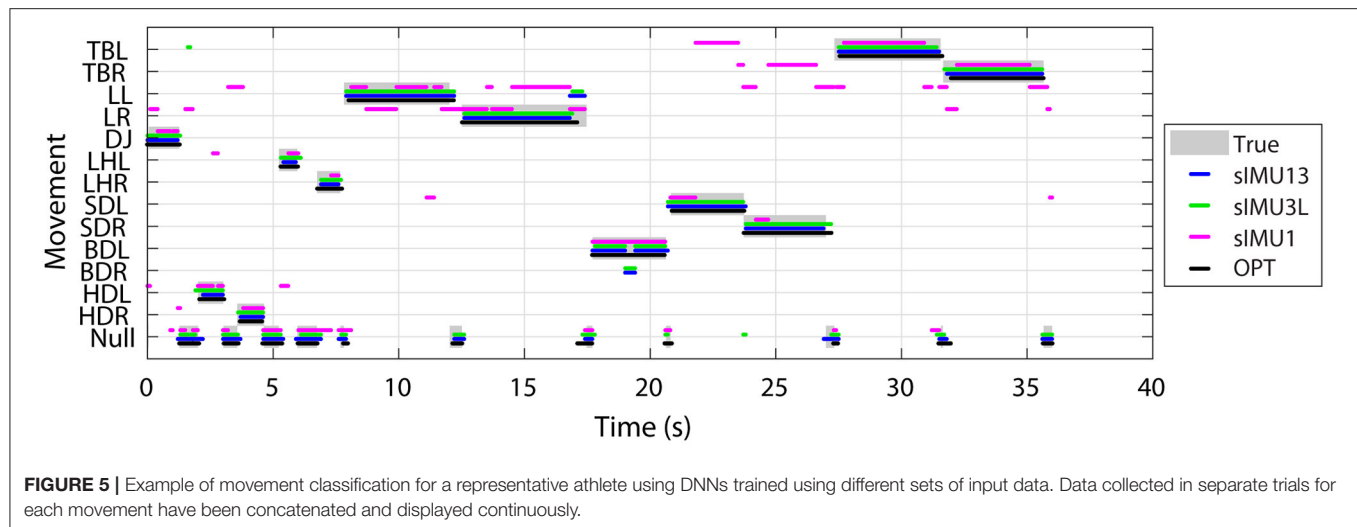
FIGURE 5 | Example of movement classification for a representative athlete using DNNs trained using different sets of input data. Data collected in separate trials for each movement have been concatenated and displayed continuously.

of a hindrance to the athlete's motion. The F1 score for sIMU3L, which used the torso and shanks, was only ∼0.04 less than for the full body DNN. Therefore, depending on the desired accuracy, classification rates may be sufficient using only three sensors.

Some movements, such as the bird-dogs, were more easily identified by the DNN, even for networks trained on sIMU data from one or two segments. This is likely because trunk motion for these movements is substantially different from that of the other movements, with the trunk horizontal and relatively stationary throughout the motion. Including arm segments prevented confusion between T-balances and lunges. The drop jump was classified particularly poorly when few segments were used, often being classified as the null condition. This may in part be due to the way the start and end of the L-hop motion was defined. The L-hop involved the athlete jumping horizontally forward, landing on one foot, then jumping laterally and landing on the opposite foot. This movement was defined to begin when the athlete had reached their maximum height during the initial jump. Without sufficient data, the DNN was unable to differentiate between the end of the drop jump movement, which involved a vertical jump, and the initial jump of the L-hop which was included in the null condition.

The F1 score we achieved in classifying athletic movements is similar to previously reported human activity classification results. The architecture of the DNN used in this study was based on the work of Ordóñez and Roggen (2016), who achieved an F1 score of 0.895 on a dataset including various modes of locomotion. Other work has reported classification accuracies ranging from 83 to 100% for everyday activities (Pärkkä et al., 2006; Yeoh et al., 2008; Attal et al., 2015; Yang et al., 2015) and 79–93% for movements involved in various sports (Schuldhaus et al., 2015; Groh et al., 2016; Anand et al., 2017; Cust et al., 2019).

Previous work on classification of everyday activities, such as walking, jogging, sitting, stair climbing, etc., has identified one sensor placed at the waist as producing the best classification accuracy (Cleland et al., 2013; Pannurat et al., 2017). In the current study, we found that a single simulated torso sensor

resulted in a poor classification accuracy of 48%. This discrepancy can likely be attributed to the differences in activities included, as optimal sensor placement depends on the activity (Atallah et al., 2011; Attal et al., 2015). The activities classified in the previous studies involve activities that are repetitive and take place over a relatively long period of time. The athletic movements included in our study, however, are short single movements. Furthermore, the need to differentiate right and left versions of the movements makes classification with a single torso-mounted sensor more challenging.

The sIMU DNNs relied on simulated IMU data generated based on optical motion tracking markers. Therefore, these results likely represent a best-case scenario for classification of these athletic movements using wearable sensors. Sensor drift is a common issue with IMUs and therefore it is possible that misclassification rates would be larger using real sensor data, particularly for long data collections as drift increases over time. Care would also need to be taken to standardize sensor placement on each body segment. While we have used the Euclidean norm of the angular velocity and linear acceleration, error would be introduced into the angular orientation of the body segments by misaligned sensors. Additionally, it may be possible to mitigate sensor misalignment issues using a static or dynamic calibration at the beginning of the data collection. Despite the reliance on simulated sensor data, the results presented here highlight the potential for movement classification using wearable sensors and provide guidance for sensor placement in future work.

In this study, separate data trials were recorded for each motion and these were combined for the classification. As a result, the amount of null data frames included was relatively small. It may be necessary to included more null condition training data, including transitions between movements, for the DNN to be used successfully on continuously collected data.

Accurate classification of movements is critical for this DNN approach to be used with no manual intervention in combination with data-driven assessments of movement quality, as the quality could only be assessed on properly identified movements. Some

errors may be possible to correct with additional processing, such as when the classification jumps to another movement for a few frames in the middle of an otherwise accurately classified movement. We observed that a large source of error was over- or under-estimating the start or end points of a movement with misclassification between the movement and the null condition. It is possible that movement quality could still be quantified with these slight errors in start and end points, but future work will be required to verify this. Alternately, a small amount of manual intervention could be used to verify task identification before proceeding to quantification of movement quality.

The favorable classification rates obtained in this work using simulated sensor data demonstrates the feasibility of classifying athletic tasks typical of movement screens using wearable sensors. Using simulated IMU data, we observed the best classification accuracy by including data from all body segments; however, we obtained good results using as few as three simulated sensors. This indicates that classification of these athletic movements using real IMU data would require at least three sensors and should include the torso and legs. Implementation of a movement classification DNN with wearable sensor data would facilitate automatic data-driven assessment of movement quality, eliminating subjective scoring, and increasing the ability to detect subtle differences.

DATA AVAILABILITY STATEMENT

Code, sample data, and trained DNNs weights are available at doi: 10.5281/zenodo.3546204.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Ottawa Research Ethics Board. Written informed consent to participate in this study was provided by the participant or their legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AC, GR, and RG conceived of the study, interpreted the results, and critically revised the manuscript. GR collected and processed the data. AC implemented the neural network, analyzed the results, and prepared the manuscript.

FUNDING

ACKNOWLEDGMENTS

The data used in this study were collected by Motus Global.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/article/10.3389/fbioe.2019.00473/full#supplementary-material

REFERENCES

Anand, A., Sharma, M., Srivastava, R., Kaligounder, L., and Prakash, D. (2017). "Wearable motion sensor based analysis of swing sports," in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017* (Cancun), 261–267. doi: 10.1109/ICMLA.2017.0-149

Atallah, L., Lo, B., King, R., and Yang, G. Z. (2011). "Sensor positioning for activity recognition using wearable accelerometers," in *IEEE Transactions on Biomedical Circuits and Systems (IEEE)*, 320–329. doi: 10.1109/TBCAS.2011.2160540

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors* 15, 31314–31338. doi: 10.3390/s151229858

Bonazza, N. A., Smuin, D., Onks, C. A., Silvis, M. L., and Dhawan, A. (2017). Reliability, validity, and injury predictive value of the functional movement screen. *Am. J. Sports Med.* 45, 725–732. doi: 10.1177/0363546516641937

Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* 46, 1–33. doi: 10.1145/2499621

Cleland, I., Kikhia, B., Nugent, C., Boytsov, A., Hallberg, J., Synnes, K., et al. (2013). Optimal placement of accelerometers for the detection of everyday activities. *Sensors* 13, 9183–9200. doi: 10.3390/s130709183

Clifton, D. R., Harrison, B. C., Hertel, J., and Hart, J. M. (2013). Relationship between functional assessments and exercise-related changes during static balance. *J. Strength Cond. Res.* 27, 966–972. doi: 10.1519/JSC.0b013e318260b723

Cust, E. E., Sweeting, A. J., Ball, K., and Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *J. Sports Sci.* 37, 568–600. doi: 10.1080/02640414.2018.1521769

Frost, D. M., Beach, T. A. C., Callaghan, J. P., and McGill, S. M. (2015). FMS scores change with performers' knowledge of the grading criteria - Are general whole-body movement screens capturing "dysfunction"? *J. Strength Cond. Res.* 29, 3037–3043. doi: 10.1097/JSC.0000000000000211

Gjoreski, H., Luštrek, M., and Gams, M. (2011). "Accelerometer placement for posture recognition and fall detection," in *Proceedings - 2011 7th International Conference on Intelligent Environments, IE 2011* (Nottingham), 47–54. doi: 10.1109/IE.2011.11

Groh, B. H., Fleckenstein, M., and Eskofier, B. M. (2016). "Wearable trick classification in freestyle snowboarding," in *BSN 2016 - 13th Annual Body Sensor Networks Conference* (San Francisco, CA: IEEE), 89–93. doi: 10.1109/BSN.2016.7516238

Gulgin, H., and Hoogenboom, B. (2014). The functional movement screening (FMS)™: an inter-rater reliability study between raters of varied experience. *Int. J. Sports Phys. Ther.* 9, 14–20.

Hammerla, N. Y., Halloran, S., and Plötz, T. (2016). "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *IJCAI International Joint Conference on Artificial Intelligence* (New York, NY), 1533–1540.

Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., and Eskofier, B. M. (2017). Activity recognition in beach volleyball using a deep convolutional neural network: leveraging the potential of deep learning in sports. *Data Min. Knowl. Discov.* 31, 1678–1705. doi: 10.1007/s10618-017-0495-0

Kern, N., Schiele, B., and Schmidt, A. (2003). "Multi-sensor activity context detection for wearable computing," in *Lecture Notes in Computer Science*, eds E. Aarts, R. W. Collier, E. van Leonen, and B. de Ruyter (Berlin; Heidelberg: Springer), 220–232.

Kim, Y., and Toomajian, B. (2016). Hand gesture recognition using micro-Doppler signatures with convolutional neural network. *IEEE Access* 4, 7125–7130. doi: 10.1109/ACCESS.2016.2617282

Lee, S.-M., Yoon, S. M., and Cho, H. (2017). "Human activity recognition from accelerometer data using Convolutional Neural Network," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* (Jeju), 131–134. doi: 10.1109/BIGCOMP.2017.7881728

Minick, K. I., Kiesel, K. B., Burton, L., Taylor, A., Plisky, P., and Butler, R. J. (2010). Interrater reliability of the Functional Movement Screen. *J. Strength Cond. Res.* 24, 479–486. doi: 10.1519/JSC.0b013e3181c09c04

Nguyen, L. N. N., Rodríguez-Martín, D., Català, A., Pérez-López, C., Samà, A., and Cavallaro, A. (2015). "Basketball activity recognition using wearable inertial measurement units," in *Proceedings of the XVI International Conference on Human Computer Interaction* (Vilanova i la Geltrú), 1–6. doi: 10.1145/2829875.2829930

Olguin, D. O., and Pentland, A. (2006). "Human activity recognition: accuracy across common locations for wearable sensors," in *IEEE 10th International Symposium on Wearable Computers* (Montreaux), 5–7.

Ordóñez, F. J., and Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16:115. doi: 10.3390/s16010115

Pannurat, N., Thiemjarus, S., Nantajeewarawat, E., and Anantavrasilp, I. (2017). Analysis of optimal sensor positions for activity classification and application on a different data collection scenario. *Sensors* 17:774. doi: 10.3390/s17040774

Pärkkä, J., Ermes, M., Korpipää, P., Mäntyjärvi, J., Peltola, J., and Korhonen, I. (2006). "Activity classification using realistic data from wearable sensors," in *IEEE Transactions on Information Technology in Biomedicine*, 119–128. doi: 10.1109/TITB.2005.856863

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop* (Long Beach, CA).

Ross, G. B., Dowling, B., Troje, N. F., Fischer, S. L., and Graham, R. B. (2018). Objectively differentiating movement patterns between elite and novice athletes. *Med. Sci. Sports Exerc.* 50, 1457–1464. doi: 10.1249/MSS.0000000000001571

Schuldhaus, D., Zwick, C., Körger, H., Dorschky, E., Kirk, R., and Eskofier, B. M. (2015). "Inertial sensor-based approach for shot/pass classification during

a soccer match," in *21st ACM KDD Workshop on Large-Scale Sports Analytics* (Sydney, NSW), 1–4.

Smith, C. A., Chimera, N. J., Wright, N. J., and Warren, M. (2013). Intrarater reliability of the Functional Movement Screen. *J. Strength Cond. Res.* 27, 982–987. doi: 10.1519/JSC.0b013e31825c32a8

Teyhen, D. S., Shaffer, S. W., Lorenson, C. L., Halfpap, J. P., Donofry, D. F., Walker, M. J., et al. (2012). The functional movement screen: a reliability study. *J. Orthop. Sports Phys. Ther.* 42, 530–540. doi: 10.2519/jospt.2012.3838

Yang, J. B., Nguyen, M. N., San, P. P., Li, X. L., and Krishnaswamy, S. (2015). "Deep convolutional neural networks on multichannel time series for human activity recognition," in *International Joint Conference on Artificial Intelligence* (Buenos Aires), 3995–4001.

Yeoh, W. S., Pek, I., Yong, Y. H., Chen, X., and Waluyo, A. B. (2008). Ambulatory monitoring of human posture and walking speed using wearable accelerometer sensors. *Conf. IEEE Eng. Med. Biol. Soc.* 2008, 5184–5187. doi: 10.1109/IEMBS.2008.4650382

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., et al. (2014). "Convolutional neural networks for human activity recognition using mobile sensors," in *International Conference on Mobile Computing, Applications and Services, MobiCASE 2014* (Austin, TX), 197–205. doi: 10.4108/icst.mobicase.2014.257786

A Machine Learning and Wearable Sensor Based Approach to Estimate External Knee Flexion and Adduction Moments During Various Locomotion Tasks

Bernd J. Stetter[1]*, Frieder C. Krafft[1], Steffen Ringhof[1,2], Thorsten Stein[1†] and Stefan Sell[1,3†]

[1] Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany, [2] Department of Sport and Sport Science, University of Freiburg, Freiburg, Germany, [3] Joint Center Black Forest, Hospital Neuenbuerg, Neuenbuerg, Germany

Joint moment measurements represent an objective biomechanical parameter of knee joint load in knee osteoarthritis (KOA). Wearable sensors in combination with machine learning techniques may provide solutions to develop assistive devices in KOA patients to improve disease treatment and to minimize risk of non-functional overreaching (e.g., pain). The purpose of this study was to develop an artificial neural network (ANN) that estimates external knee flexion moments (KFM) and external knee adduction moments (KAM) during various locomotion tasks, based on data obtained by two wearable sensors. Thirteen participants were instrumented with two inertial measurement units (IMUs) located on the right thigh and shank. Participants performed six different locomotion tasks consisting of linear motions and motions with a change of direction, while IMU signals as well as full body kinematics and ground reaction forces were synchronously recorded. KFM and KAM were determined using a full body biomechanical model. An ANN was trained to estimate the KFM and KAM time series using the IMU signals as input. Evaluation of the ANN was done using a leave-one-subject-out cross-validation. Concordance of the ANN-estimated KFM and reference data was categorized for five tasks (walking straight, 90° walking turn, moderate running, 90° running turn and 45° cutting maneuver) as strong ($r \geq 0.69$, rRMSE ≤ 23.1) and as moderate for fast running ($r = 0.65 \pm 0.43$, rRMSE $= 25.5 \pm 7.0\%$). For all locomotion tasks, KAM yielded a lower concordance in comparison to the KFM, ranging from weak ($r \leq 0.21$, rRMSE $\geq 33.8\%$) in cutting and fast running to strong ($r = 0.71 \pm 0.26$, rRMSE $= 22.3 \pm 8.3\%$) for walking straight. Smallest mean difference of classical discrete load metrics was seen for KFM impulse, $10.6 \pm 47.0\%$. The results demonstrate the feasibility of using only two IMUs to estimate KFM and KAM to a limited extent. This methodological step facilitates further work that should aim to improve the estimation accuracy to provide valuable biofeedback systems for KOA patients. Greater accuracy of effective implementation could be achieved by a participant- or task-specific ANN modeling.

Keywords: knee osteoarthritis, biomechanics, knee joint loading, biofeedback, artificial neural networks, accelerometers and gyroscopes, reduced sensor set

INTRODUCTION

Medio-tibiofemoral knee osteoarthritis (KOA) is a major cause of disability in elderly people (Hurley et al., 1997) and accounts for high socio-economic burden in industrial countries (Neogi et al., 2009; Reeves and Bowling, 2011; Ferreira et al., 2015). Symptoms known as knee pain, functional impairment and a loss of mobility, can lead to physical and psychological disability and reduced quality of life (Bennell et al., 2011; Richards et al., 2017).

Mechanical factors, particularly the knee joint load have shown to profoundly influence the severity and progression of KOA (Sharma et al., 1998; Andriacchi and Muendermann, 2006; Foroughi et al., 2009; Bennell et al., 2011; Reeves and Bowling, 2011). A widely used surrogate measure of the compressive load of the medial compartment is the external knee adduction moment (KAM) (Sharma et al., 1998; Bennell et al., 2011; Reeves and Bowling, 2011; Ferreira et al., 2015). Moreover, the knee flexion moment (KFM) has been highlighted as a critical measure to assess the loading of the medial compartment (Walter et al., 2010; Ferreira et al., 2015; Cheung et al., 2018) as well as to quantify the progression of patellofemoral cartilage damage (Teng et al., 2015; Crossley et al., 2016).

Beside other non-pharmacological conservative treatments (e.g., bracing or footwear interventions) (Sarzi-Puttini et al., 2005; Reeves and Bowling, 2011), gait modification approaches by gait retraining therapies (e.g., modifying the foot progression angle) have shown to be effective to reduce the KAM during walking and to improve the symptoms of patients (Barrios et al., 2010; Cheung et al., 2018; Karatsidis et al., 2018). Richards et al. (2017) stated in their systematic review that a strong potential exists for the development of biofeedback systems for reducing KAM and pain and for improving knee joint function in KOA patients. The development of assistive devices (e.g., a smart knee sleeve to monitor the knee load in combination with a smartphone-based user feedback system) could help to provide effective disease-enhancing interventions to slow down the loss of cartilage volume (Shull et al., 2014). Additionally, as exercise is a key component of the KOA management (Bennell et al., 2011; Ferreira et al., 2015; Richards et al., 2017), assistive devices could be beneficial in supporting therapeutical exercise. However, most of the current studies with respect to the assessment of knee joint loading were conducted in a laboratory setting using motion capture and force plate measurements (Barrios et al., 2010; Richards et al., 2017; Cheung et al., 2018). The major shortcoming of such laboratory-based methods is that they cannot be completely included into a patients' habitual environment (Muro-de-la-Herran et al., 2014; Shull et al., 2014).

As a consequence, alternative measurement technologies have been provided progressive advances over the past years (Muro-de-la-Herran et al., 2014; Wong et al., 2015). One of the first studies toward a wearable measurement tool was done by van den Noort et al. (2011). The authors tested the effect of an instrumented force shoe in combination with an optoelectronic marker system on target variables (e.g., KAM) in 20 KOA patients. Therein, the authors stated the necessity of additional measurement equipment (e.g., inertial sensors) to obtain joint positions and orientations as a complement to ground reaction force (GRF) measurements in order to calculate the KAM. Karatsidis et al. (2016) compared GRF estimation accuracies of a full-body inertial motion capture and optical motion capture system due to the importance of the GRF measures as input in biomechanical analysis to estimate joint kinetics. Their results showed comparable results between the two systems. Therefore, the authors concluded that the inertial sensor-based system has a high potential in monitoring critical biomechanical parameters in habitual conditions. Yang and Mao (2015) postulated a method for evaluating the intersegmental forces and moments acting on the lower limbs during walking solely based on posture data obtained from seven inertial sensors placed on the lower limbs and trunk in combination with a 3D analytical model. In 2018 Karatsidis et al. proposed and evaluated a wearable visual feedback system for gait retraining using inertial sensing with seven inertial measurement units (IMUs) and augmented reality technologies. The foot progression angle was used for visual feedback and was tracked by the wearable system with a root mean square error of $2.4°$, compared to an optical motion capture system. Knee joint kinetics were not analyzed in this study. A further approach of a mobile assessment of knee joint biomechanics in natural environment was recently provided by Konrath et al. (2019). The authors estimated the KAM and the tibio-femoral joint contact force during activities of daily living by means of combining musculoskeletal modeling with inertial motion capture (17 IMUs). The results showed comparable estimation accuracies for the IMU-based approach compared to the same musculoskeletal model using optical motion capture and force plate measurements.

The majority of applied methods require modeling of the musculoskeletal system to a certain degree, with mandatory embedded subject-specific anthropometric data (e.g., mass, dimensions, and center of mass of the body segments). However, such modeling processes inevitably introduce inaccuracies (van den Noort et al., 2013; Faber et al., 2016; Ancillao et al., 2018). In contrast, machine learning-based approaches do not need an *a priori* knowledge of the model as they build up their model by using training data (Sivakumar et al., 2016; Ancillao et al., 2018; Halilaj et al., 2018). Accurate predictions for new data can be made by learning the relationship between a set of independent variables (e.g., IMU signals) and one or more dependent variables (e.g., KAM) (Lin et al., 2016; Halilaj et al., 2018). Several studies have shown that machine learning techniques, such as artificial neural networks (ANN), are powerful tools to deduce biomechanical variables based on measured accelerations or angular velocities of body segments (Leporace et al., 2015; Guo et al., 2017; Ancillao et al., 2018; Wouda et al., 2018; Stetter et al., 2019). The study by Wouda et al. (2018) used an ANN approach to estimate vertical GRFs and sagittal knee kinematics during running, based on three inertial sensors placed at the lower legs and the pelvis. The estimated force-time profiles and flexion/extension profiles showed high agreement with the optical and GRF reference measure. In a recent study we presented an ANN approach to estimate knee joint forces in sport movements (Stetter et al., 2019). Good agreement between ANN-estimated outcomes and inverse dynamics-calculated vertical and anterior-posterior knee joint forces were shown, which highlights

the feasibility of an ANN approach to estimate internal loadings on the knee joint structures.

Although the above described studies have estimated joint kinematics and kinetics during locomotion, no study has directly estimated biomechanical surrogate measures for knee joint load in KOA using an ambulatory minimal body-worn sensor setup so far. Therefore, the purpose of this study was to develop an ANN that estimates KFM and KAM during various locomotion tasks based on data obtained by two wearable sensors integrated in a knee sleeve. The findings of this study could help to (1) overcome current restrictions in the mobile assessment of knee joint loading in KOA patients and (2) open new possibilities in diagnosing the patients' habitual life, which could help to improve disease treatment strategies and minimizing the risk of non-functional overreaching (e.g., pain).

MATERIALS AND METHODS

Participants

The current study used data from the sample presented in Stetter et al. (2019) and forms a secondary dataset analysis. The sample consisted of 13 healthy males (age, 26.1 $\pm$ 2.9 years; height, 178.7 $\pm$ 5.5 cm; body mass, 78.4 $\pm$ 5.9 kg). All participants exhibited bowlegs (minimum inter-knee distance of 0.05 m), which mimics the common varus malalignment of medial KOA patients (Bennell et al., 2011). All participants gave written informed consent in accordance with the Declaration of Helsinki. The study was approved by the ethics committee of the Karlsruhe Institute of Technology.

Experimental Protocol

Measurements were performed at the BioMotion Center, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany. Two identical custom-built 6DOF IMUs (1,500 Hz, ± 8 g accelerometer, $\pm 2,000°$/s gyroscope) were attached to each participant's right leg while they performed six different locomotion tasks at self-selected speed: walking straight, 90° walking turn, moderate running, fast running, 90° running turn and 45° cutting maneuver. Participants were instructed to perform the 90° turns in clockwise direction. A detailed description of the right orientated cutting maneuver (named as v-cut) is described by Neptune et al. (1999). Participants were instructed to perform at least three successful trials of each task. A trial was considered successful when the right foot landed cleanly within the boundaries of a force plate. The IMUs were positioned in two patch pockets at the upper and lower frontal end of a customized knee sleeve (**Figure 1**). This positioning was chosen in order to capture IMU signals closely related to knee kinematics and dynamics, as the recent study by Matijevich et al. (2019) has highlighted that a targeted approach is necessary to obtain structure-specific loading.

Full body kinematics and GRFs (1,000 Hz, AMTI Inc., Watertown, MA) were collected synchronously using a marker-based motion capture system (11 MX-13 cameras, 200 Hz, Vicon, Oxford, UK) in order to perform biomechanical modeling. A total of 42 spherical reflective markers were placed on the participants' body segments in accordance to the ALASKA Dynamicus protocol (ALASKA, INSYS GmbH,

FIGURE 1 | A participant wearing the knee sleeve on the right leg. The two inertial measurement units were placed in the patch pockets at the upper and lower frontal end of the knee sleeve.

Germany) (Härtel and Hermsdorf, 2006; Willwacher et al., 2017). Prior to the attachment of the data collection equipment, standardized anthropometric measurements were exhibited. The measurements consisted of a total of 22 length, width and circumference measures of the body segments. Prior to performing trials, a static calibration trial was recorded for each participant in a natural upright posture.

Biomechanical Model

The 3D marker coordinates and GRF data were reconstructed and filtered with a 15 Hz low-pass filter (zero-phase Butterworth 4th order) (Kristianslund et al., 2012). Inverse dynamics modeling was performed using the full-body Dynamicus 9 model (Härtel and Hermsdorf, 2006; Willwacher et al., 2017). Each participant was individually scaled to the generic linked-segment model using the measured anthropometrics and the static calibration trial (Whittlesey and Robertson, 2014). In a next step, the marker trajectories and GRFs acquired from the dynamic trials were used to determine the knee flexion moment (KFM) and the knee adduction moment (KAM). A 20 N threshold of the vertical GRF was used to extract the stance phase for each locomotion movement (Milner and Paquette, 2015). KFM and KAM time series were time-normalized to 100 time

steps representing 0–100% of the stance phase. Joint moment amplitudes were normalized to body weight and expressed as external moments.

Machine Learning Model

ANN modeling was set up with the Neural Network Toolbox in MATLAB R2019a (The MathWorks, USA). The IMU signals were low-pass filtered (zero-phase Butterworth 4th order filter; cut-off frequency of 15 Hz) and each trial was cropped to contain data for the same phase as the biomechanical data. An IMU signal matrix (rows: 13 participants × three trials × six tasks × 100 time steps; columns: two locations × six spatial dimensions) and a biomechanical data matrix (rows: 13 participants × three trials × six tasks × 100 time steps; columns: two variables) were created by vertically concatenating the IMU signals and KFM and KAM time series of all trials, respectively. An ANN was trained to model the association between the IMU signals and the KFM and KAM time series. The IMU signal matrix served as input and the biomechanical data matrix served as output (target). As a consequence, the ANN had 12 and two variables (i.e., nodes) in its input and output layer, respectively. The ANN architecture was inspired by previous work (Favre et al., 2012; Wouda et al., 2018) and had two hidden layers with 100 and 20 neurons, which were connected to the input and output nodes. The hidden layers and the output layer consisted of hyperbolic tangent sigmoid transfer functions and a linear transfer function, respectively. Initialization of the ANN was done using the Nguyen-Widrow initialization function. The ANN was trained for 1,000 iterations with Levenberg-Marquardt back-propagated error correction (Watson and Moré, 1978) and training was stopped if the gradient did not decrease for six consecutive epochs or if the gradient was smaller than 1×10^{-6}. Evaluation of the ANN was done using a leave-one-subject-out cross-validation (Halilaj et al., 2018). The cross-validation involved training the ANN with all trials from 12 participants (i.e., the training set) and then testing with the trials from the remaining participant (i.e., the test set). As cross-dependencies between the input and output in a combined estimation model for KFM and KAM may affect the estimation accuracy (Wouda et al., 2018), independent models for KFM and KAM were also build. Independent models were trained and evaluated in the same manner as the combined model, beside the fact that only one variable was chosen in its output layer.

Statistical Analysis

According to previous studies, for each movement, the agreement between the ANN-estimated outcomes (KFM* and KAM*) and the inverse dynamics-calculated data (KFM and KAM) was derived from Pearson's correlation coefficients, which were categorized as weak ($r \leq 0.35$), moderate ($0.35 < r \leq 0.67$), strong ($0.67 < r \leq 0.90$) and excellent ($r > 0.90$) (Taylor, 1990; Fluit et al., 2014; Karatsidis et al., 2016). Additionally, the Root Mean Squared Error (RMSE) and relative Root Mean Squared Error (rRMSE) were determined to assess the accuracy of the ANN estimations (Ren et al., 2008). The rRMSE facilitates the comparison between the different locomotion tasks with different moment amplitudes. The averages and standard deviations were calculated for r, RMSE and rRMSE from the 13 cross-validation

subsets. Average r values across participants were computed using Fisher's z transformation (Corey et al., 1998). Mean r values were expressed in the original range from −1 to 1 by reversing the transformation. Furthermore, peak KFM* and KFM* impulse as well as peak KAM and KAM impulse were evaluated as classical discrete load metrics (Bennell et al., 2011; Teng et al., 2015). Impulse represents the area under the corresponding moment-time curve. Percent differences (%Diff) between ANN-estimated and inverse dynamics-calculated peak and impulse metrics were used to provide a pragmatic interpretation.

RESULTS

Estimated Continuous Outcomes

The ANN-estimated KFM* and KAM* time series of the whole stance phase are illustrated in **Figures 2**, **3**, respectively, with the measured references used for comparison. An overview of the estimated accuracy for all movements is presented in **Table 1**.

For the different locomotion tasks, the ANN-estimated time series revealed moderate to strong correlations for the KFM* and weak to strong correlations for the KAM*. The highest correlation for KFM* and KAM* was observed for moderate running ($r = 0.85 \pm 0.43$; mean ± standard deviation) and for walking straight (0.71 ± 0.26), respectively. For all locomotion tasks, the RMSE for KFM* was between 0.26 ± 0.09 and 1.13 ± 0.46 Nm/kg, whereas for KAM*, that was between 0.18 ± 0.06 and 0.92 ± 0.54 Nm/kg. The rRMSE for the different locomotion tasks ranged between $17.2 \pm 3.1\%$ (walking 90° turn) and $25.5 \pm 7.0\%$ (fast running) for KFM* and between $22.3 \pm 8.3\%$ (walking straight) and $37.2 \pm 7.8\%$ (cutting maneuver) for KAM*.

Discrete Load Metrics

The inverse dynamics-calculated and ANN-estimated discrete load metrics (peak moments and moment integrals) are shown in **Table 2**. **Table 3** presents the %Diff results for each of the performed locomotion tasks. The 90° walking turn showed the smallest %Diff ($6.7 \pm 31.3\%$) for the ANN-estimated KFM impulse in comparison to the reference values. In contrast, %Diff of KAM impulse were higher with a minimum value of $42.7 \pm 108.9\%$ for moderate running. The smallest %Diff for the estimation of peak KFM and KAM was $24.7 \pm 33.0\%$ (moderate running) and $39.1 \pm 101.0\%$ (walking straight), respectively. Across all locomotion tasks, mean differences of peak moments and moment integrals were lower for the KFM* in comparison to the KAM* (40.4 ± 56.5 vs. $130.3 \pm 157.3\%$ and 10.6 ± 47.0 vs. $161.4 \pm 252.8\%$, respectively).

Model Comparison

The changes in estimation accuracy due to independent model building for KFM and KAM for each of the analyzed locomotion tasks is presented in **Table 4**. Independent model building resulted in a lower r value for both KFM and KAM in the majority (five out of six) of the analyzed locomotion tasks in comparison to the combined estimation model. Across all locomotion tasks, mean RMSE and mean rRMSE increased for KFM* (RMSE = 0.15, rRMSE = 1.18) and KAM* (mean RMSE = 0.13, rRMSE = 0.26) due to independent model building.
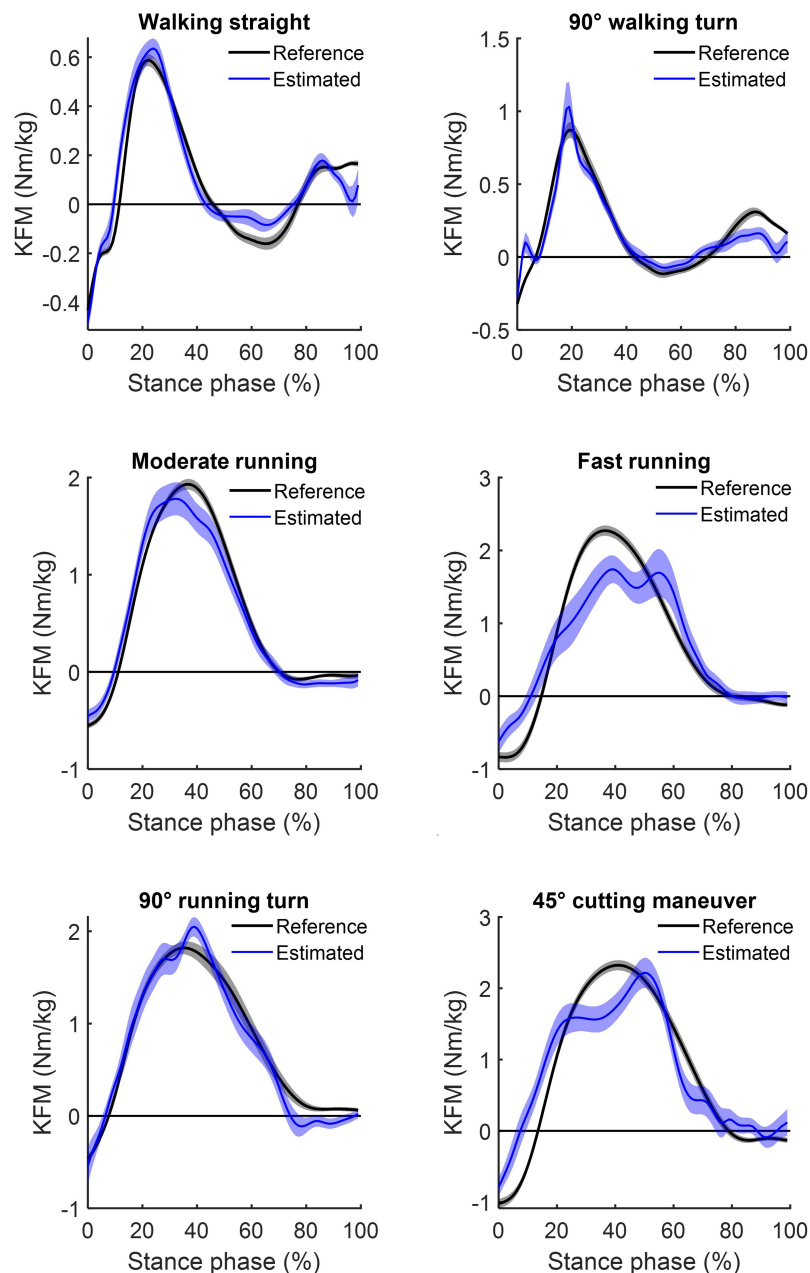
FIGURE 2 | Mean (and standard error) of the estimated knee flexion moments (blue) for the six analyzed locomotion tasks compared to their respective inverse dynamics-calculated values (black). Positive values indicate external flexion moments and negative values indicate external extension moments.

DISCUSSION

This study was aimed to develop and train an ANN model to estimate KFM and KAM during various locomotion tasks based on data obtained by two wearable sensors. The mobile assessment of knee joint loading enlarges the scope of diagnostic methods and disease management of KOA, which could help to improve disease treatment strategies and minimizing the risk of non-functional overreaching (e.g., pain).

The results of the study show a higher estimation accuracy of the KFM compared to the KAM over most locomotion task. However, estimation accuracy highly varied between tasks for both the KFM and the KAM, especially with increasing intensity and movement velocity. Apart from walking straight, for all locomotion tasks, a distinct reduced level of agreement was found between the ANN-estimated outcomes and reference data for the KAM (mean $r = 0.39 \pm 0.32$, mean rRMSE = 29.9% $\pm$ 8.1%) in comparison to the KFM (mean $r = 0.74 \pm 0.36$, mean rRMSE = 20.8% $\pm$ 5.7%). Discrete load metrics highlighted lower %Diff
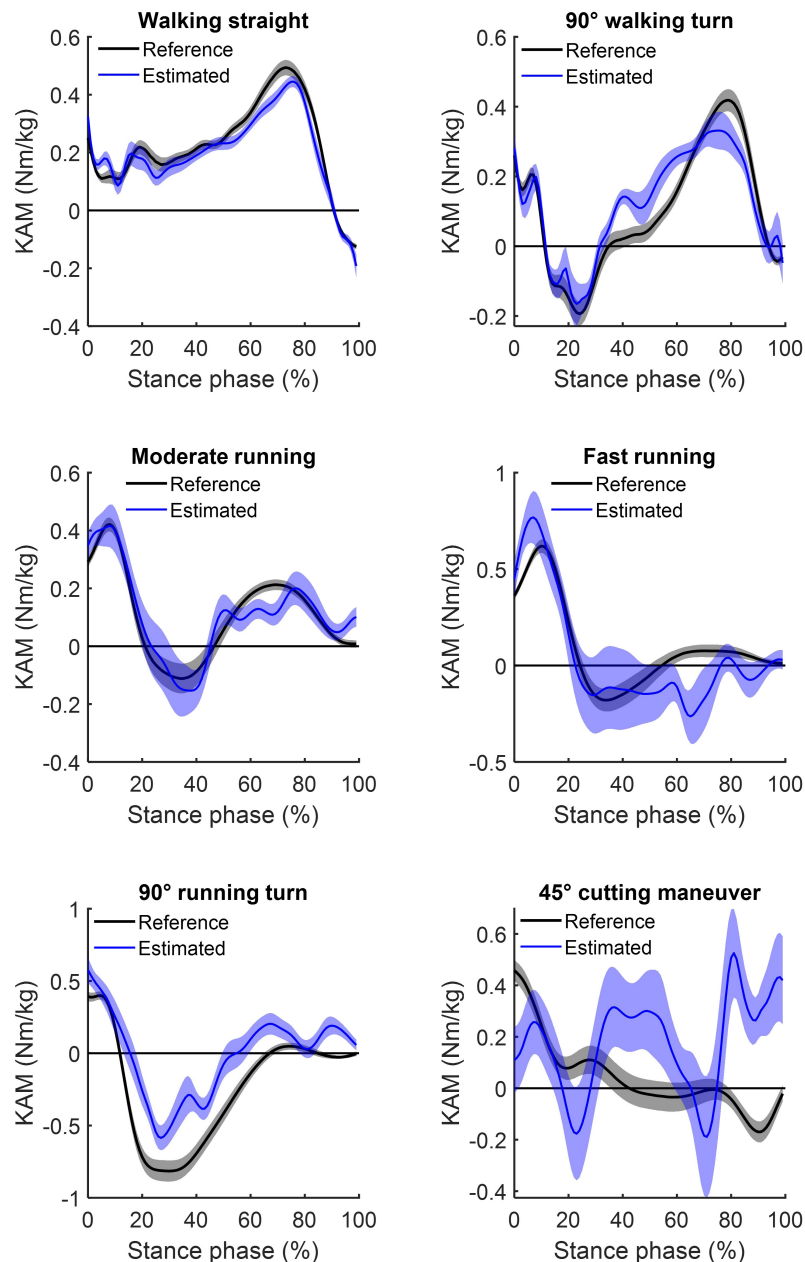
FIGURE 3 | Mean (and standard error) of the estimated knee adduction moments (blue) for the six analyzed locomotion tasks compared to their respective inverse dynamics-calculated values (black). Positive values indicate external adduction moments and negative values indicate external abduction moments.

of KFM impulses in comparison to KFM peaks in all locomotion tasks, whereas %*Diff* of KAM impulses were lower compared to KAM peaks only in three out of the six locomotion tasks.

Estimation Accuracy Across Different Locomotion Tasks

In general, when comparing the estimation accuracy across the different locomotion tasks, predictive power was always better and %*Diff* was always less for KFM than for KAM. On average,

strong correlations ($r = 0.74$) and rRMSE of 20.8% for KFM and moderate correlations ($r = 0.39$) with rRMSE of 29.9% for KAM were found. Nonetheless, distinct differences between KFM and KAM estimation values were evident across the locomotion tasks.

For KFM, highest correlations with the inverse dynamics calculations were found for moderate running ($r = 0.85$), which is reinforced by lowest %Diff for both the peak and impulse of the KFM. The lowest correlations and largest rRMSE were found for fast running ($r = 0.65$; rRMSE $= 25.5$%). Nevertheless, %*Diff* for KFM peaks and impulses during fast running were

TABLE 1 | Accuracy (r, Pearson's correlation coefficient; RMSE, root-mean-squared error; rRMSE, relative root-mean-squared error) of the estimated continuous outcomes [knee flexion moment (KFM*), and knee adduction moment (KAM*)].

Locomotion task	KFM*			KAM*		
	r	RMSE (Nm/kg)	rRMSE (%)	r	RMSE (Nm/kg)	rRMSE (%)
Walking straight	0.72 ± 0.32	0.26 ± 0.09	18.4 ± 5.3	0.71 ± 0.26	0.18 ± 0.06	22.3 ± 8.3
90° walking turn	0.69 ± 0.31	0.32 ± 0.10	17.2 ± 3.1	0.56 ± 0.33	0.29 ± 0.10	23.9 ± 6.4
Moderate running	0.85 ± 0.43	0.58 ± 0.20	19.7 ± 7.9	0.40 ± 0.35	0.37 ± 0.14	34.4 ± 13.5
Fast running	0.65 ± 0.43	1.13 ± 0.46	25.5 ± 7.0	0.21 ± 0.47	0.80 ± 0.46	33.8 ± 8.5
90° running turn	0.79 ± 0.28	0.77 ± 0.20	20.8 ± 4.5	0.51 ± 0.22	0.62 ± 0.19	27.9 ± 3.9
45° cutting maneuver	0.73 ± 0.41	1.05 ± 0.41	23.1 ± 6.5	−0.05 ± 0.30	0.92 ± 0.54	37.2 ± 7.8
Mean	0.74 ± 0.36	0.67 ± 0.24	20.8 ± 5.7	0.39 ± 0.32	0.53 ± 0.25	29.9 ± 8.1

Data are presented as mean ± standard deviations. Mean r and r standard deviation were computed using Fisher's z transformation.

TABLE 2 | Inverse dynamics-calculated (KFM and KAM) and ANN-estimated (KFM* and KAM*) discrete load metrics (peak and impulse).

Locomotion task	KFM		KAM		KFM*		KAM*	
	Peak	Impulse	Peak	Impulse	Peak	Impulse	Peak	Impulse
	(Nm/kg)	(Nm/kg)	(Nm/kg)	(Nm/kg)	(Nm/kg)	(Nm/kg)	(Nm/kg)	(Nm/kg)
Walking straight	0.67 ± 0.13	45.72 ± 14.52	0.54 ± 0.15	69.16 ± 26.03	0.91 ± 0.30	52.31 ± 24.83	0.65 ± 0.18	64.23 ± 13.76
90° walking turn	1.02 ± 0.38	71.79 ± 36.05	0.57 ± 0.18	44.65 ± 21.96	1.55 ± 1.19	70.12 ± 31.23	0.90 ± 0.44	52.06 ± 17.00
Moderate running	2.03 ± 0.34	193.05 ± 58.08	0.52 ± 0.16	43.48 ± 21.81	2.57 ± 0.92	197.00 ± 90.16	0.84 ± 0.39	56.35 ± 50.26
Fast running	2.49 ± 0.35	246.20 ± 71.51	0.77 ± 0.20	51.35 ± 27.01	3.44 ± 1.92	259.80 ± 118.59	1.72 ± 0.99	91.98 ± 62.78
90° running turn	2.20 ± 0.40	240.28 ± 83.01	0.60 ± 0.17	20.80 ± 6.56	3.12 ± 0.88	253.13 ± 91.06	1.45 ± 0.73	61.94 ± 31.19
45° cutting maneuver	2.52 ± 0.50	284.58 ± 85.73	0.61 ± 0.23	43.97 ± 35.24	3.50 ± 1.29	310.16 ± 144.96	2.11 ± 1.38	120.90 ± 110.35
Mean	1.82 ± 0.79	180.27 ± 98.86	0.60 ± 0.09	45.57 ± 15.56	2.52 ± 1.07	190.42 ± 106.46	1.28 ± 0.57	74.58 ± 26.66

Data are presented as mean ± standard deviations; KFM, knee flexion moment; KAM, knee adduction moment.

TABLE 3 | Percent differences (%Diff) of discrete load metrics (peak and impulse).

Locomotion task	KFM		KAM	
	Peak	Impulse	Peak	Impulse
	%Diff	%Diff	%Diff	%Diff
Walking straight	44.3 ± 70.8	27.4 ± 83.9	39.1 ± 101.0	62.0 ± 253.1
90° walking turn	47.1 ± 60.6	6.7 ± 31.3	82.4 ± 110.5	69.3 ± 127.5
Moderate running	24.7 ± 33.0	0.65 ± 37.2	68.7 ± 94.5	42.7 ± 108.9
Fast running	37.2 ± 68.7	6.8 ± 40.7	123.5 ± 124.1	94.2 ± 145.3
90° running turn	44.9 ± 45.2	12.1 ± 46.5	159.8 ± 157.1	230.0 ± 179.9
45° cutting maneuver	44.1 ± 60.7	10.0 ± 42.6	308.2 ± 356.5	470.0 ± 702.0
Mean	40.4 ± 56.5	10.6 ± 47.0	130.3 ± 157.3	161.4 ± 252.8

Data are presented as mean ± standard deviations; KFM, knee flexion moment; KAM, knee adduction moment.

TABLE 4 | Increase (+) or decrease (−) in estimation accuracy (r, Pearson's correlation coefficient; RMSE, root-mean-squared error; rRMSE, relative root-mean-squared error) due to independent model building in comparison to the combined model.

Locomotion task	KFM*			KAM*		
	r	RMSE (Nm/kg)	rRMSE (%)	r	RMSE (Nm/kg)	rRMSE (%)
Walking straight	0.03	0.00	0.50	−0.20	0.05	2.64
90° walking turn	−0.02	0.03	1.56	−0.08	0.07	0.09
Moderate running	−0.02	0.18	1.31	−0.10	0.09	−1.58
Fast running	−0.03	0.15	0.90	−0.04	0.20	1.87
90° running turn	−0.08	0.11	0.85	−0.14	0.16	−0.87
45° cutting maneuver	−0.07	0.44	1.94	0.26	0.22	−0.57
Mean	−0.03	0.15	1.18	−0.05	0.13	0.26

KFM, knee flexion moment; KAM*, knee adduction moment.*

lower than for most of the other locomotion tasks, except for moderate running. Interestingly, the largest %Diff was found for walking straight, while %Diff of moment integrals were in general lower compared to %Diff of peak moments. These findings indicate that our ANN-configuration is more appropriate for estimating knee joint loading over the stance phase than for estimating the peak moment of the stance phase. In particular,

during walking straight, the low knee flexion moment peaks and impulses might account for the strong correlations but large %Diff. Albeit, for KFM generally high agreement was found for ANN-estimated outcomes, with a reduced performance for

the high intensity movements running and cutting maneuvers. In contrast, in these movements lower %Diff occurred to the lower-intensity movements.

For the estimation of KAM, overall weak to strong correlations were found for the analyzed movements. Estimation accuracy was highest in walking straight ($r = 0.71$, rRMSE = 22.3%). Mediocre correlations were found in moderate running as well as 90° walking/running turns ($0.40 \leq r \leq 0.56$), and low or negative correlations in fast running and 45° cutting maneuvers ($-0.05 \leq r \leq 0.21$). With regard to rRMSE, alterations of locomotion speed (walking to running) and direction (turning and cutting) led to slight reductions in accuracy of the ANN estimations. Concomitant, large increases in %Diff along with high variability were detected in fast running, 90° running turns and 45° cutting maneuvers (KAM impulse: 94.2, 230.0, and 470.0%, respectively). A potential reason for the less estimation accuracy and larger differences for movements with increased velocity and changes of direction might be the higher variation in the execution of these movements, while locomotor tasks such as walking or moderate running are performed automatically with repeatable characteristics (Schmidt and Lee, 2011). Similarly, variability in estimation accuracy was also shown by Fluit et al. (2014), evaluating a prediction model for GRFs and moments during various activities of daily living by means of 3D full-body motion.

However, a generalization of the estimation accuracies cannot be deduced, as a reduced estimation accuracy in continuous outcomes does not necessarily result in an inaccurate estimation of discrete variables, as it was seen in the KFM during fast running. Similarly, good agreement in continuous outcomes does not implicate accurate estimation of discrete load metrics, as seen in 90° running turn. Furthermore, it must be noted that most KFM and KAM show high standard deviations, which indicates a wide dispersion across participants. Nonetheless, %Diff of KFM were entirely lower in the impulses compared to the peak values. In contrast, %Diff of KAM impulse were lower compared to the peak values only in three out of the six locomotion tasks (90° walking turn, moderate and fast running). Summarized, KAM estimations were less accurate both for continuous and for discrete outcomes compared to KFM and should therefore be treated with caution. The more pronounced characteristic changes in the KAM time series between locomotion tasks in comparison to the KFM time series are a potential reason for the reduced estimation accuracy in KAM (see **Figures 2**, **3**).

Furthermore, with respect to the comparison of a combined estimation model for KFM and KAM and independent models for KFM and KAM, the results show that an independent model building leads to slightly decreased estimation accuracy of the KFM and a more pronounced decrease of the KAM, concomitant with increased RMSE and rRMSE in the investigated locomotion tasks. Hence, if only one variable was chosen as an output, decreased performance for the model was observed, indicating that cross-dependencies between input and output in the combined estimation model clearly affected the estimation accuracy. Overall, the combined estimation model for KFM and KAM presented a fair estimation accuracy, especially, in the low-intensity movements.

Comparison of Different Wearable Measurement Systems

A novel machine learning based method was developed and applied in this study to estimate KFM and KAM based on data obtained by two wearable sensors integrated in a knee sleeve. Various approaches have experienced progressive advances to assess the mechanical loading of KOA patients in their habitual environment over the past years. The majority of the approaches were based on analytical biomechanical models, which typically determine joint moments by means of the inverse dynamics calculations. As a consequence, GRF measurements and kinematic data are necessary to perform such analysis (Whittlesey and Robertson, 2014).

Van den Noort et al. presented in 2011 an instrumented force shoe as an alternative to force plate measurements. Subsequently, an ambulatory measurement system, consisting of the instrumented force shoe and an inertial measurement system combined with a linked-segment model, was used to compare KAM measures with a laboratory based system in KOA patients (van den Noort et al., 2013). Limited accuracy was shown and the authors concluded that a more advanced calibrated linked-segment model should be investigated (van den Noort et al., 2013). As an alternative to a direct measurement of GRF, Karatsidis et al. (2016) estimated GRF by means of a full-body inertial motion capture system during walking. Their results showed for the comparison with an optical motion capture system higher r values (range 0.82–0.99 and 0.76–0.99 for the inertial and optical motion capture systems, respectively) and lower rRMSE values (range from 5 to 15% for both systems) compared to the KFM and KAM estimations present in this study. More recent studies from Dorschky et al. (2019) and Konrath et al. (2019) used inertial motion capturing and musculoskeletal modeling to estimate biomechanical variables, such as joint kinematics and kinetics without GRF data. Dorschky et al. (2019) presented high correlations for sagittal plain kinematics ($r > 0.93$) and kinetics ($r > 0.90$) in gait and running. In accordance, Konrath et al. (2019) estimated the KAM and the tibio-femoral joint contact force during daily living activities (e.g., stair walking) with moderate to strong correlation coefficients. However, such approaches using inertial sensor data and musculoskeletal models require more IMUs (seven IMUs in Dorschky et al., 2019 and 17 IMUs in Konrath et al., 2019) compared to this study's approach.

Parallel to the analytical model development, an increasing number of machine learning approaches have been explored to simplify data acquisition and modeling strategies to estimate target variables, such as the KAM (Liu et al., 2009; Favre et al., 2012; Wouda et al., 2018). ANN modeling does not require modeling of the musculoskeletal system, as the relationship between the input IMU signals and the target variables is build up during the training process of the model (Halilaj et al., 2018; Wouda et al., 2018). However, ground truth reference data, such as the inverse dynamics-calculated KFMs and KAMs, are required during the supervised learning process of the model. Providing a large amount of known output data is essential

to establish a robust model (Sivakumar et al., 2016; Halilaj et al., 2018). Wouda et al. (2018) used similar ANN modeling to the one used in this study for estimating vertical GRF and sagittal knee kinematics during running. The estimated vertical GRF profiles of their non-personalized ANN developed by eight participants showed a higher correlation ($r > 0.90$) to the actual force time series. The slightly reduced estimation accuracy in the current study ($r < 0.85$) may depend on the variety of locomotion tasks included in the model building. A more locomotion task-specific modeling may lead to an increased estimation accuracy for individual tasks, but has the disadvantage that each task must be modeled by itself (Wouda et al., 2018). In consequence, the combination with an activity recognition approach could help to select individual estimation models in practical applications.

Limitations

Certain limitations of this study need to be considered when interpreting the results. One consideration worth noting is that the estimation accuracy depends on the neural network architecture. The ANN was built on previous work (Favre et al., 2012; Wouda et al., 2018), which highlighted that such configuration is capable of mapping non-linearity between input and output; however, other model specifications may result in an improved estimation accuracy. The ANN was trained with data from multiple participants as well as various locomotion tasks, which should rather lead to a less participant- and task-specific but a more generic model. As a consequence, this approach rather yields a decreased performance due to a lack of individualization, but has the advantage that not every new user needs to perform a training phase (Favre et al., 2012; Wouda et al., 2018). Further research is necessary to assess if a single participant learning approach increases the estimation accuracy. Another limitation is that we included a homogeneous group of participants consisting of only males without any musculoskeletal disorders and the translation of the results to the target group of KOA patients remains speculative. Nonetheless, future clinical studies may benefit from the use of the method developed in this study, especially in low-intensity movements (Richards et al., 2017). Beyond, the sample size was rather small, including 13 participants. Similar investigations included comparable numbers of participants (e.g., sample of eight participants in Wouda et al., 2018 or sample of 17 participants in Leporace et al., 2015). The small sample size potentially limits the outcome, as the robustness of the relationship between the input and output variables of the ANN depends on the amount of training data (Sivakumar et al., 2016; Ancillao et al., 2018; Halilaj et al., 2018). Finally, it cannot be fully ensured that the fixation technique excluded any oscillations or misalignment of the IMUs, even though the exact fit of the sleeve and the sensors was repetitively checked. However, the wearable sensors were integrated in a knee sleeve on purpose to mimic natural effects and to capture IMU signals closely related to the joint under investigation.

CONCLUSION

This study demonstrated the potential of estimating KFM and KAM for various locomotion tasks using a minimal body-worn sensor setup consisting of two IMUs integrated in a knee sleeve. The agreement between the ANN-estimated outcomes and inverse dynamics-calculated data was strong for the majority of analyzed locomotion tasks in the KFM and moderate in the KAM. Overall, higher estimation accuracies were seen for the KFM in comparison to the KAM across all locomotion tasks. The accuracy limitations especially of KAM estimation makes prediction of knee joint loading challenging. In order to reach an acceptable level of accuracy related to critical changes due to KOA, typically characterized by relatively small kinetic differences, a participant- or task-specific modeling could be helpful. This has important implications for the development of wearable devices as well as for scientific research on KOA. The highest estimation accuracy for both KFM and KAM of walking straight matches the main characteristic of KOA therapy and treatment by low-intensity movements (e.g., walking). Looking ahead, wearable technology could serve as a rehabilitation aid for patients with KOA leading to an improved load management, which could result in a slower progression.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ethics committee of the Karlsruhe Institute of Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BS, FK, SR, TS, and SS were involved in the design of the study. BS, FK, and SR carried out all data collection and analysis. BS, FK, SR, and TS were involved in the interpretation and discussion of the results. BS took the lead in writing the manuscript. All authors provided critical feedback and contributed to the final manuscript.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Ancillao, A., Tedesco, S., Barton, J., and O'Flynn, B. (2018). Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: a systematic review. *Sensors* 18:E2564. doi: 10.3390/s18082564

Andriacchi, T. P., and Muendermann, A. (2006). The role of ambulatory mechanics in the initiation and progression of knee osteoarthritis. *Curr. Opin. Rheumatol.* 18, 514–518. doi: 10.1097/01.bor.0000240365.16842.4e

Barrios, J. A., Crossley, K. M., and Davis, I. S. (2010). Gait retraining to reduce the knee adduction moment through real-time visual feedback of dynamic knee alignment. *J. Biomech.* 43, 2208–2213. doi: 10.1016/j.jbiomech.2010.03.040

Bennell, K. L., Bowles, K.-A., Wang, Y., Cicuttini, F., Davies-Tuck, M., and Hinman, R. S. (2011). Higher dynamic medial knee load predicts greater cartilage loss over 12 months in medial knee osteoarthritis. *Ann. Rheum. Dis.* 70, 1770–1774. doi: 10.1136/ard.2010.147082

Cheung, R. T. H., Ho, K. K. W., Au, I. P. H., An, W. W., Zhang, J. H. W., Chan, Z. Y. S., et al. (2018). Immediate and short-term effects of gait retraining on the knee joint moments and symptoms in patients with early tibiofemoral joint osteoarthritis: a randomized controlled trial. *Osteoarthr. Cartil.* 26, 1479–1486. doi: 10.1016/j.joca.2018.07.011

Corey, D. M., Dunlap, W. P., and Burke, M. J. (1998). Averaging correlations: expected values and bias in combined Pearson rs and Fisher's z transformations. *J. Gen. Psychol.* 125, 245–261. doi: 10.1080/00221309809595548

Crossley, K. M., Stefanik, J. J., Selfe, J., Collins, N. J., Davis, I. S., Powers, C. M., et al. (2016). 2016 Patellofemoral pain consensus statement from the 4th International Patellofemoral Pain Research Retreat, Manchester. Part 1: Terminology, definitions, clinical examination, natural history, patellofemoral osteoarthritis and patient-reported outcome measures. *Br. J. Sports Med.* 50, 839–843. doi: 10.1136/bjsports-2016-096384

Dorschky, E., Nitschke, M., Seifer, A.-K., van den Bogert, A. J., and Eskofier, B. M. (2019). Estimation of gait kinematics and kinetics from inertial sensor data using optimal control of musculoskeletal models. *J. Biomech.* 95:109278. doi: 10.1016/j.jbiomech.2019.07.022

Faber, G. S., Chang, C. C., Kingma, I., Dennerlein, J. T., and van Dieën, J. H. (2016). Estimating 3D L5/S1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system. *J. Biomech.* 49, 904–912. doi: 10.1016/j.jbiomech.2015.11.042

Favre, J., Hayoz, M., Erhart-Hledik, J. C., and Andriacchi, T. P. (2012). A neural network model to predict knee adduction moment during walking based on ground reaction force and anthropometric measurements. *J. Biomech.* 45, 692–698. doi: 10.1016/j.jbiomech.2011.11.057

Ferreira, G. E., Robinson, C. C., Wiebusch, M., Viero, C. C., da Rosa, L. H., and Silva, M. F. (2015). The effect of exercise therapy on knee adduction moment in individuals with knee osteoarthritis: a systematic review. *Clin. Biomech.* 30, 521–527. doi: 10.1016/j.clinbiomech.2015.03.028

Fluit, R., Andersen, M. S., Kolk, S., Verdonschot, N., and Koopman, H. F. J. M. (2014). Prediction of ground reaction forces and moments during various activities of daily living. *J. Biomech.* 47, 2321–2329. doi: 10.1016/j.jbiomech.2014.04.030

Foroughi, N., Smith, R., and Vanwanseele, B. (2009). The association of external knee adduction moment with biomechanical variables in osteoarthritis: a systematic review. *Knee* 16, 303–309. doi: 10.1016/j.knee.2008.12.007

Guo, Y., Storm, F., Zhao, Y., Billings, S., Pavic, A., Mazza, C., et al. (2017). A new proxy measurement algorithm with application to the estimation of vertical ground reaction forces using wearable sensors. *Sensors* 17:E2181. doi: 10.3390/s17102181

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. (2018). Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J. Biomech.* 81, 1–11. doi: 10.1016/j.jbiomech.2018.09.009

Härtel, T., and Hermsdorf, H. (2006). Biomechanical modelling and simulation of human body by means of DYNAMICUS. *J. Biomech.* 39:S549. doi: 10.1016/S0021-9290(06)85262-0

Hurley, M., Scott, D., Rees, J., and Newham, D. (1997). Sensorimotor changes and functional performance in patients with knee osteoarthritis. *Ann. Rheum. Dis.* 56, 641–648. doi: 10.1136/ard.56.11.641

Karatsidis, A., Bellusci, G., Schepers, H. M., Zee, M., de Andersen, M. S., and Veltink, P. H. (2016). Estimation of ground reaction forces and moments during gait using only inertial motion capture. *Sensors* 17:75. doi: 10.3390/s17010075

Karatsidis, A., Richards, R. E., Konrath, J. M., van den Noort, J. C., Schepers, H. M., Bellusci, G., et al. (2018). Validation of wearable visual feedback for retraining foot progression angle using inertial sensors and an augmented reality headset. *J. NeuroEng. Rehabil.* 15, 1–12. doi: 10.1186/s12984-018-0419-2

Konrath, J. M., Karatsidis, A., Schepers, H. M., Bellusci, G., Zee, M., de, and Andersen, M. S. (2019). Estimation of the knee adduction moment and joint contact force during daily living activities using inertial motion capture. *Sensors* 19:1681. doi: 10.3390/s19071681

Kristianslund, E., Krosshaug, T., and van den Bogert, A. J. (2012). Effect of low pass filtering on joint moments from inverse dynamics: implications for injury prevention. *J. Biomech.* 45, 666–671. doi: 10.1016/j.jbiomech.2011.12.011

Leporace, G., Batista, L. A., Metsavaht, L., and Nadal, J. (2015). "Residual analysis of ground reaction forces simulation during gait using neural networks with different configurations," in *Conference proceedings: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2015 (Milano), 2812–2815.

Lin, D., Vasilakos, A. V., Tang, Y., and Yao, Y. (2016). Neural networks for computer-aided diagnosis in medicine: a review. *Neurocomputing* 216, 700–708. doi: 10.1016/j.neucom.2016.08.039

Liu, Y., Shih, S.-M., Tian, S.-L., Zhong, Y.-J., and Li, L. (2009). Lower extremity joint torque predicted by using artificial neural network during vertical jump. *J. Biomech.* 42, 906–911. doi: 10.1016/j.jbiomech.2009.01.033

Matijevich, E. S., Branscombe, L. M., Scott, L. R., and Zelik, K. E. (2019). Ground reaction force metrics are not strongly correlated with tibial bone load when running across speeds and slopes: Implications for science, sport and wearable tech. *PLoS ONE* 14:e0210000. doi: 10.1371/journal.pone.0210000

Milner, C. E., and Paquette, M. R. (2015). A kinematic method to detect foot contact during running for all foot strike patterns. *J. Biomech.* 48, 3502–3505. doi: 10.1016/j.jbiomech.2015.07.036

Muro-de-la-Herran, A., Garcia-Zapirain, B., and Mendez-Zorrilla, A. (2014). Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors* 14, 3362–3394. doi: 10.3390/s140203362

Neogi, T., Felson, D., Niu, J., Nevitt, M., Lewis, C. E., Aliabadi, P., et al. (2009). Association between radiographic features of knee osteoarthritis and pain: results from two cohort studies. *BMJ* 339:b2844. doi: 10.1136/bmj.b2844

Neptune, R. R., Wright, I. C., and van den Bogert, A. J. (1999). Muscle coordination and function during cutting movements. *Med. Sci. Sports Exerc.* 31, 294–302. doi: 10.1097/00005768-199902000-00014

Reeves, N. D., and Bowling, F. L. (2011). Conservative biomechanical strategies for knee osteoarthritis. *Nat. Rev. Rheumatol.* 7, 113–122. doi: 10.1038/nrrheum.2010.212

Ren, L., Jones, R. K., and Howard, D. (2008). Whole body inverse dynamics over a complete gait cycle based only on measured kinematics. *J. Biomech.* 41, 2750–2759. doi: 10.1016/j.jbiomech.2008.06.001

Richards, R., van den Noort, J. C., Dekker, J., and Harlaar, J. (2017). Gait retraining with real-time biofeedback to reduce knee adduction moment: systematic review of effects and methods used. *Arch. Phys. Med. Rehabil.* 98, 137–150. doi: 10.1016/j.apmr.2016.07.006

Sarzi-Puttini, P., Cimmino, M. A., Scarpa, R., Caporali, R., Parazzini, F., Zaninelli, A., et al. (2005). Osteoarthritis: an overview of the disease and its treatment strategies. *Semin. Arthritis Rheumat.* 35, 1–10. doi: 10.1016/j.semarthrit.2005.01.013

Schmidt, R. A., and Lee, T. D. (2011). *Motor Control and Learning: A Behavioral Emphasis.* Champaign, IL: Human Kinetics.

Sharma, L., Hurwitz, D. E., Thonar, E. J., Sum, J. A., Lenz, M. E., Dunlop, D. D., et al. (1998). Knee adduction moment, serum hyaluronan level, and disease severity in medial tibiofemoral osteoarthritis. *Arthritis Rheumat.* 41, 1233–1240. doi: 10.1002/1529-0131(199807)41:7<1233::AID-ART14>3.0.CO;2-L

Shull, P. B., Jirattigalachote, W., Hunt, M. A., Cutkosky, M. R., and Delp, S. L. (2014). Quantified self and human movement: a review on the clinical impact of wearable sensing and feedback for gait analysis and intervention. *Gait Posture* 40, 11–19. doi: 10.1016/j.gaitpost.2014.03.189

Sivakumar, S., Gopalai, A. A., Gouwanda, D., and Hann, L. K. (2016). "ANN for gait estimations: a review on current trends and future applications," in *IEEE-EMBS*

Conferences on Biomedical Engineering and Sciences: 4-8 December 2016, Kuala Lumpur (Piscataway, NJ: IEEE), 311–316.

Stetter, B. J., Ringhof, S., Krafft, F. C., Sell, S., and Stein, T. (2019). Estimation of knee joint forces in sport movements using wearable sensors and machine learning. *Sensors* 19:3690. doi: 10.3390/s19173690

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *J. Diagn. Med. Sonogr.* 6, 35–39. doi: 10.1177/87564793900 0600106

Teng, H.-L., MacLeod, T. D., Link, T. M., Majumdar, S., and Souza, R. B. (2015). Higher knee flexion moment during the second half of the stance phase of gait is associated with the progression of osteoarthritis of the patellofemoral joint on magnetic resonance imaging. *J. Orthop. Sports Phys. Ther.* 45, 656–664. doi: 10.2519/jospt.2015.5859

van den Noort, J., van der Esch, M., Steultjens, M. P., Dekker, J., Schepers, M., Veltink, P. H., et al. (2011). Influence of the instrumented force shoe on gait pattern in patients with osteoarthritis of the knee. *Med. Biol. Eng. Comp.* 49, 1381–1392. doi: 10.1007/s11517-011-0818-z

van den Noort, J. J. C., van der Esch, M., Steultjens, M. P. M., Dekker, J., Schepers, M. H. M., Veltink, P. H., et al. (2013). Ambulatory measurement of the knee adduction moment in patients with osteoarthritis of the knee. *J. Biomech.* 46, 43–49. doi: 10.1016/j.jbiomech.2012.09.030

Walter, J. P., D'Lima, D. D., Colwell, C. W., and Fregly, B. J. (2010). Decreased knee adduction moment does not guarantee decreased medial contact force during gait. *J. Orthop. Res.* 28, 1348–1354. doi: 10.1002/jor. 21142

Watson, G. A., and Moré, J. J. (eds.) (1978). *The Levenberg-Marquardt algorithm: Implementation and Theory: Numerical Analysis*. Berlin; Heidelberg: Springer.

Whittlesey, S. N., and Robertson, D. G. E. (2014). "Two-dimentional inverse dynamics," in *Research Methods in Biomechanics*, eds D. G. E. Robertson, G. E.

Caldwell, J. Hamill, G. Kamen, and S. N. Whittlesey (Champaign, IL: Human Kinetics), 103–123.

Willwacher, S., Funken, J., Heinrich, K., Müller, R., Hobara, H., Grabowski, A. M., et al. (2017). Elite long jumpers with below the knee prostheses approach the board slower, but take-off more effectively than non-amputee athletes. *Sci. Rep.* 7:16058. doi: 10.1038/s41598-017-16383-5

Wong, C., Zhang, Z., Lo, B. P. L., and Yang, G.-Z. (2015). Wearable Sensing for Solid Biomechanics. *IEEE Sens. J.* 15, 2747–2760. doi: 10.1109/JSEN.2015.2393883

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., Van Beijnum, B.-J. F., et al. (2018). Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors. *Front. Physiol.* 9:218. doi: 10.3389/fphys.2018.00218

Yang, E. C.-Y., and Mao, M.-H. (2015). 3D analysis system for estimating intersegmental forces and moments exerted on human lower limbs during walking motion. *Measurement* 73, 171–179. doi: 10.1016/j.measurement.2015.05.020

Check for updates

Tibial Acceleration-Based Prediction of Maximal Vertical Loading Rate During Overground Running: A Machine Learning Approach

Rud Derie[1]*, Pieter Robberechts[2]*, Pieter Van den Berghe[1], Joeri Gerlo[1], Dirk De Clercq[1], Veerle Segers[1] and Jesse Davis[2]

[1] Department of Movement and Sports Sciences, Ghent University, Ghent, Belgium, [2] Department of Computer Science, KU Leuven, Leuven, Belgium

Ground reaction forces are often used by sport scientists and clinicians to analyze the mechanical risk-factors of running related injuries or athletic performance during a running analysis. An interesting ground reaction force-derived variable to track is the maximal vertical instantaneous loading rate (VILR). This impact characteristic is traditionally derived from a fixed force platform, but wearable inertial sensors nowadays might approximate its magnitude while running outside the lab. The time-discrete axial peak tibial acceleration (APTA) has been proposed as a good surrogate that can be measured using wearable accelerometers in the field. This paper explores the hypothesis that applying machine learning to time continuous data (generated from bilateral tri-axial shin mounted accelerometers) would result in a more accurate estimation of the VILR. Therefore, the purpose of this study was to evaluate the performance of accelerometer-based predictions of the VILR with various machine learning models trained on data of 93 rearfoot runners. A subject-dependent gradient boosted regression trees (XGB) model provided the most accurate estimates (mean absolute error: 5.39 ± 2.04 BW·s^{-1}, mean absolute percentage error: 6.08%). A similar subject-independent model had a mean absolute error of 12.41 ± 7.90 BW·s^{-1} (mean absolute percentage error: 11.09%). All of our models had a stronger correlation with the VILR than the APTA ($p < 0.01$), indicating that multiple 3D acceleration features in a learning setting showed the highest accuracy in predicting the lab-based impact loading compared to APTA.

Keywords: running biomechanics, impact loading, tibial shock, machine learning, wearable sensor, gait analysis

INTRODUCTION

Ground reaction forces are relevant parameters for running analysis (Pohl et al., 2009; Crowell and Davis, 2011; Van Der Worp et al., 2016; Clark et al., 2017). They partially describe the center of mass' state of motion during running and are often used by sport scientists and clinicians to analyze the mechanical risk-factors of running related injuries (Bredeweg et al., 2013; Napier et al., 2018) and/or athletic performance (Preece et al., 2019).

A commonly used ground reaction force-derived variable is the maximal vertical instantaneous loading rate (VILR), which is calculated as the maximal slope of the rising vertical ground reaction force – time curve (Ueda et al., 2016). VILR has been used to characterize the impact (i.e., high rate of force development due to the rapid deceleration of all body segments during the foot-ground collision) during running (Gerritsen et al., 1995). This measure could discriminate groups of rearfoot runners with a history of stress fractures (Van Der Worp et al., 2016) and plantar fasciitis (Pohl et al., 2009). Consequently, VILR has been considered clinically relevant and has been a main outcome variable in gait retraining studies targeting runners with high VILR (Crowell and Davis, 2011; Clansey et al., 2014; Willy et al., 2016).

Ground reaction forces are traditionally measured using fixed force platforms or instrumented treadmills (Ueda et al., 2016). Unfortunately, measurements with force platforms are laboratory-based and require both expensive equipment and extensive post-processing. These factors limit the potential of monitoring in-field running biomechanics, whereas wearable inertial measurement units can accommodate this by predicting running gait parameters outside the laboratory (Falbriard et al., 2018; Wouda et al., 2018). In this respect, an ambulatory low-cost accelerometer was proposed as a potential surrogate candidate to estimate VILR when force platforms are not available (Ngoh et al., 2018). Previous research has identified a moderate to good correlation (range of $r_{mean} = 0.64–0.84$) between the axial peak tibial acceleration (APTA) captured by a skin-mounted accelerometer at the tibia and VILR (Laughton et al., 2003; Pohl et al., 2009; Greenhalgh et al., 2012; Zhang et al., 2016; Van den Berghe et al., 2019). Therefore, using APTA as a surrogate measure for VILR seems justifiable (Sheerin et al., 2019).

However, the APTA is based on a single, basic feature (i.e., the peak value) of the time-continuous 1D tibial acceleration signal. Consequently, a large amount of data is neglected, which may lead to missing important information. A combination of multiple features of the 3D tibial acceleration signals, possibly including complex and higher-order ones, may result in a more accurate predictor of VILR than only considering APTA. Hence, a performant computational model that extracts relevant features and effectively copes with any non-linear relationships (between the features of the tibial acceleration signals and the target VILR) is desired. In that way, machine learning techniques could help to analyze continuous time-series data without pre-selecting discrete variables. Holzreiter and Köhle (1993) introduced the use of neural networks to assess gait patterns in locomotion biomechanics. Recently more advanced machine learning techniques have been used to detect pathologic gait-patterns (Williams et al., 2015; Zeng et al., 2016), fatigue (Janssen et al., 2011; Op De Beéck et al., 2018) as well as classifying gender, performance-level (Clermont et al., 2018) and age-related running patterns (Fukuchi et al., 2011).

To gain a better understanding of the relationship between the external load and potential injury risk in overground running, a more accurate estimation of the athlete's impact loading is an essential methodological prerequisite. The screening of runners on impact intensity could be more accurate by estimating VILR by means of a machine-learned model instead of relying on the APTA only. Consequently, this study proposes and evaluates the performance (e.g., predictive accuracy, calculation time, diagnostic ability) of an inertial sensor-based method to estimate the runner's VILR based on bilateral 3D shin-mounted accelerometer data using a machine learning approach. It was hypothesized that the incorporation of these extracted features into a set of machine-learned models would result in stronger predictive and diagnostic capacities than considering APTA only.

MATERIALS AND METHODS

Ethics Statement and Participants

Ninety three subjects engaged in recreational as well as competitive running (55 men and 38 women) were recruited from the local community. Runners were included if they were free of running-related injuries and ran at least 15 km per week (**Table 1**). All subjects signed an informed consent prior to the testing. Approval for the study was obtained from the ethical committee of the Ghent University Hospital (2015/0864).

Protocol and Setup

All runners were equipped with a backpack/tablet system to measure the tibial accelerations (Van den Berghe et al., 2019). Two tri-axial accelerometers (LIS331, Sparfkun, Colorado, United States; 1000 Hz/axis), were as tight as tolerable strapped with sports tape on the antero-medial side of both tibias, 8 cm above the malleolus medialis (Laughton et al., 2003; Clansey et al., 2014). The axis of each accelerometer was orientated in a way that the vertical axis of the accelerometer coincided with the longitudinal axis of the concerned tibia. The skin around the lower leg was pre-stretched with sports tape to improve the rigid coupling between the accelerometers and the tibia (Clansey et al., 2014; Van den Berghe et al., 2019). Data collection took place during two different projects, but with an exact same measurement setup.

The first cohort consisted of 13 subjects who were asked to run on a 30 m instrumented running track at multiple running speeds (2.55 ms^{-1}, 3.20 ms^{-1}, 5.10 ms^{-1}, and preferred running speed). All subjects were habitual rearfoot strikers and were provided with the same standardized neutral distance running shoe (Li Ning Magne, ARHF041). The second cohort consisted of 80 runners running at 3.20 m·s^{-1}. Subjects were not pre-selected on their habitual footstrike pattern and received no verbal instruction about the desired footfall pattern. They wore

---

TABLE 1 | Characteristics of the subjects.

	Men		Women	
	Mean	**SD**	**Mean**	**SD**
Age (Yrs.)	35.9	9.2	34.6	10.8
Body height (m)	1.79	0.07	1.67	0.06
Body mass (kg)	76.5	10.2	60.6	7.3
Training volume (km/week)	36.4	16.9	27.9	11.0
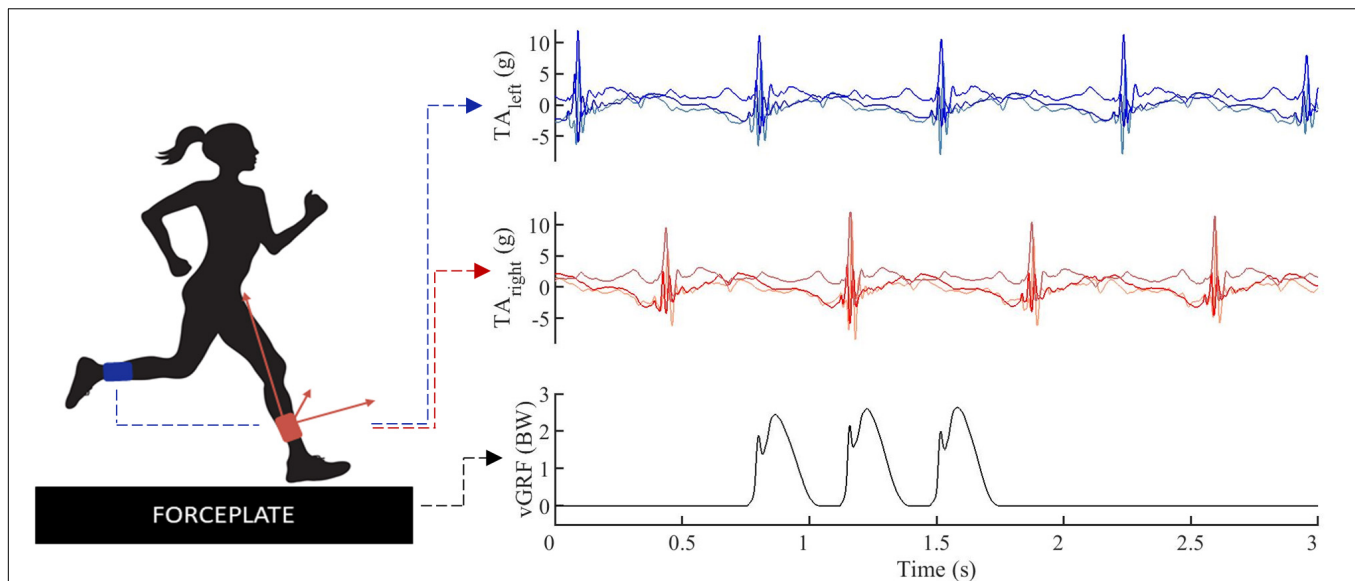
FIGURE 1 | Data example, tri-axial.1pc accelerations (TA) were simultaneously captured for the left (blue) and right (red) lower leg, vertical ground reaction forces (vGRF) were synchronized in time (black).

their regular training shoes. In both cohorts running speed was controlled by timing gates. Recorded trials were discarded and the runners received verbal feedback if their running speed was not within a 0.2 m·s^{-1} of the targeted speed. Ground reaction forces were measured at 1000 Hz by two built-in force platforms (2 and 1.2 m, AMTI, Watertown, MA, United States). Accelerometer and force data were synchronized in time (**Figure 1**) by means of an infrared impulse sent from the motion capture system. The pulse was captured by an infrared sensor attached to the backpack system. For a more detailed description of this synchronization protocol we refer to Van den Berghe et al. (2019).

Data Processing
Example Construction and Data Preprocessing
Ground reaction force data were filtered using a zero-lag second-order low-pass Butterworth filter with a cutoff frequency of 60 Hz. VILR was calculated as the maximal value of the first derivative of the vertical ground reaction force component following initial contact (vertical ground reaction forces exceeding a 5N threshold) (Ueda et al., 2016). This was subsequently normalized to the subject's body weight. The acceleration signals were filtered in order to separate the linear acceleration from the gravity component and remove high-frequency noise using the approach of van Hees et al. (2013). The filtering settings were selected using a tuning procedure where 2/3 of the data was used to train a model and 1/3 to evaluate the model. First, to find a sensible range for the parameters, a manual exploration was performed using Chebyshev (type I and type II) and Butterworth filters with settings derived from related research. Subsequently, a grid search of Butterworth filters [(0.2, 1.0; step = 0.2)×(40.0, 70.0; step = 5)] was applied to the acceleration signals and the filter which resulted in the best performance on the evaluation set was selected, which

was a second-order band-pass filter with cutoff frequencies of 0.8 and 45 Hz (**Figure 2**).

We extracted individual strides by splitting the collected signals at the take-off events of the opposite feet. This guarantees that each window contains the part of the acceleration signal that is relevant for determining the VILR. Next, we mirrored the data from the right and left leg, such that each of these strides starts with the right leg making ground contact. This procedure effectively doubled the amount of training data.

Each of the 93 subjects completed on average 16 trials (range: 6 to 67 trials), with each trial containing 2.67 strides on average. In total, 23 trials were removed from the data set due to clear errors in measured ground reaction forces and/or tibial accelerations. This resulted in 4037 examples in total.

Feature Construction
A large set of features consisting of three broad categories was considered: (1) auto-generated statistical features of the 3D acceleration waveforms, (2) trial-specific features, and (3) subject-describing features (**Figure 2**).

Auto-generated statistical features
First, from the tri-axial filtered acceleration signals of both feet, we extracted the window between the initial ground contact event and the event where the vertical acceleration component reaches 0 g. Next, we calculated a comprehensive set of time-series features from these windows using the TsFresh Python package (Christ et al., 2018). The extracted features include both basic characteristics of the signals (e.g., mean, maximum, number of peaks, timing of peak values) and more complex features (e.g., continuous wavelet coefficients, coefficients of an autoregressive model, the time reversal symmetry statistic, Fourier coefficients). We refer to the TsFresh paper (Christ et al., 2018) for a full description and list of features.

Data Preprocessing and Feature Engineering

FIGURE 2 | Data preprocessing and feature engineering part of the machine learning pipeline. First, the raw acceleration signals were filtered using a Butterworth bandpass filter. The optimal filter configuration was determined by training multiple models, using different filter configurations. The configuration which enabled the most accurate predictions was used henceforth. Second, feature engineering was used to derive a lower-dimensional representation of the data. The generated features were a combination of automatically generated statistical features and manually crafted domain-specific features. The set of automatically generated features was reduced using a univariate feature selection technique.

The FRESH procedure (Christ et al., 2017) was used for feature selection. First, this procedure evaluates the influence of every feature on the target (VILR) using a univariate test (i.e., Kendall rank test for real-valued features and Kolmogorov-Smirnov for binary features) and computes the *p*-value. So, it tests whether the feature and the target are not statistically independent. Subsequently, the Benjamini-Yekutieli procedure was carried out to control for the false discovery rate. This procedure reduced the set of auto-generated features to 1662.

Trial-specific features

Running speed, derived from timing gates (Van den Berghe et al., 2019), and ground contact time, derived from tibial accelerations, were included as trial-specific features for each stride. Because the ground contact time cannot be inferred directly from the tibial acceleration signals, we modeled this as a separate prediction problem. Specifically, we solved the related task of predicting the timings of the initial contact and toe off gait events. The ground contact time can then be inferred from the time difference between both events. Due to the interrelations between both gait events (e.g., a toe off event follows 160 to 350 ms after an initial contact event), we framed this as a structured prediction task. In this framework, a function between the acceleration profile and a sequence of initial contact and toe off timings was learned. Specifically, a deep structured recurrent neural network architecture was used. The neural network component of the model used the raw acceleration signals, the jerk (first order derivative of acceleration signals), roll ($\arctan(a_y \cdot a_z)$) and pitch ($\arctan\left(-a_x\sqrt{a_y^2 + a_z^2}\right)$) of both legs to infer the likelihood of a gait event happening for each

sample. Subsequently, the structured component consisted of a constrained peak detection algorithm on the likelihood function that finds the most likely combination of initial contact and toe off timings. Both components were optimized jointly. For a detailed description of this model, see Robberechts et al. (2019).

Subject-describing features

Third, the body weight and the shoe type were included. The weight of each subject is a logical feature to consider since the loading rate is expressed as a function of body weight. Furthermore, earlier research has found that footwear properties may affect VILR, even with similar foot-strike patterns (Kulmala et al., 2018). When testing the second cohort ($n = 80$), the subjects reported their shoe brand and type. The shoe's properties were verified through online databases (running shoes guru, solereview, runner's world, manufacturer's website, etc.) and subsequently categorized as being neutral, stabilization or racing flats.

Learning Approach

We considered two different learning settings, each learned on different subsets of the data (**Figure 3**):

Subject-independent model

This setting trained a model using the data from all runners except for one. The model was then evaluated on the trials from the one held-aside runner. That is, at training time the model has no access to any data about the runner for whom predictions will be made. As such, this setup estimates the model's accuracy when making predictions for new runners for whom there is no available data, which is interesting in practice. Moreover, the model remains valid if a runner adapts his running style.

FIGURE 3 | Model selection, training and evaluation part of the machine learning pipeline. Two different learning settings were considered, differing in how the data were split into training and test sets. In the subject-dependent setting, we trained a model on the data of one specific runner, including all trials except one. Subsequently, the model was evaluated on the data of the one held-aside trial. In the subject-independent setting, we trained a model on the data of all runners except for one and evaluated the model on all data of the one held-aside runner. This procedure was repeated for each trial and subject, such that we obtained performance metrics for each fold. Last, the average MAE and R^2 score per subject were reported.

Subject-dependent model

This setting trained a unique personalized model for each subject using only data from that subject. This model would work well if the relationship between the tibial acceleration and the VILR is unique to each subject.

For both settings, we compared the performance of three regression techniques: (1) Linear Regression with Elastic Net regularization (EN), (2) Linear Regression with Least Absolute Shrinkage and Selection Operator regularization (LASSO) and (3) Gradient Boosted Regression Trees (XGB). We used the implementations available in scikit-learn (Pedregosa et al., 2012) for the first two models. For the third regression technique, we used the XGBoost Python package (Chen and Guestrin, 2016).

All models were trained and evaluated in a leave-one-out cross-validation analysis. The subject independent model was iteratively trained on all but one of the subjects to be evaluated on the remaining subject. Similarly, the subject-dependent model

is trained on all but one trial of the same subject to be evaluated on the remaining trial. This procedure was repeated for all possible subjects and trials, and the mean accuracy across all folds is reported. As such, this procedure determines the average performance of the models on a group level.

Model Evaluation and Statistical Analysis

The model's accuracy was assessed using both the mean absolute error (MAE) and the coefficient of determination (R^2 score). The MAE was calculated as the absolute difference between the force platform based VILR and the machine learning predicted VILR. It measures the average magnitude of the errors in the same unit as the VILR and is therefore an easily interpretable measure for the quality of a model. This metric is mainly useful to compare across two models and for domain experts that have insight into the range of VILR values and the magnitude of acceptable errors. The R^2 score was computed as $R^2 = 1 - \frac{\sum_i y_i - f_i}{\sum_i y_i - \bar{y}}$, where y_i

are the force plate based VILR values and f_i are the machine learning predicted values. It has the advantage of being scale-free, thereby indicating how a model performs compared to a constant baseline.

The number of trials completed by each runner varies substantially. In order to avoid that one runner has an excessively large influence on the accuracy of our models, we computed the global MAE and R^2 score in a two-step procedure. First, the average MAE and R^2 score were calculated over all strides of that runner. Second, the global metrics were then calculated as the average values of these metrics over all runners that completed at least ten trials. This helps prevent the results from being unduly influenced by a single trial or a single runner.

Additionally, we considered two baseline models: a first model that always predicts a runner's average VILR for the corresponding landing foot; and a second linear regression model that only includes the APTA as a covariate.

Repeated measures analysis of variance (ANOVA) was used to examine the effect of various learning settings and regression techniques on the estimated VILR. *Post hoc* testing was conducted using a Tuckey HSD test on the relative errors. Additionally, Cohen's d_{rm} effect sizes (Lakens, 2013) were computed for the differences in MAE between each machine learning model and the APTA baseline model. We refer to effect sizes as small ($d \leq 0.2$), medium ($0.2 < d \leq 0.8$) and large ($d > 0.8$) as suggested by Cohen (2013). Statistical analysis was done in Python using the *SciPy* (ANOVA) and *statsmodels* (Tuckey HSD) libraries, with the significance level set at $p = 0.05$.

To assess the diagnostic ability of each model, it was opted to express the model accuracy in the proportion of correct classifications of high impact runners at a common running speed. Because a cut-off for high impact running at the speed of 3.2 m·s^{-1} was lacking, those runners with a mean VILR within the highest 33% of our database were selected. The diagnostic ability of the models was assessed by calculating their sensitivity and specificity. Sensitivity is the proportion of runners who are correctly categorized as having a high VILR among those who truly have a high VILR. Similarly, specificity is the proportion of runners who are correctly categorized as not having a high VILR among all runners who truly do not have a high VILR. The Receiver Operating Characteristic curves were constructed to demonstrate the trade-off between both metrics using various cut-off values for the predictions.

RESULTS

Predictive Performance of the Machine Learning Models

Table 2 summarizes the predictive performance (MAE and R^2 scores) of all learned models. In terms of regression techniques, XGB consistently outperformed the other learners ($p < 0.05$; in all but the subject-independent model with subject-describing features setting). Therefore, the results of the XGB learner is reported in the remainder of this section. The differences between the different learning settings were all statistically significant

TABLE 2 | Mean absolute error (MAE) ±SD, coefficient of determination R^2 scores and effect sizes of MAE's versus the axial peak tibial acceleration (APTA) baseline for the estimation of the vertical instantaneous loading rate (VILR) by three different regression models.

Model	MAE [BW·s^{-1}]	R^2	d_{rm}	Effect size
Subject-independent (without subject-describing features)				
APTA	21.07 ± 8.13	0.6027	/	
LASSO	13.13 ± 8.79	0.7789	0.3576	Medium
EN	12.91 ± 7.73	0.7811	0.3749	Medium
XGB	12.71 ± 7.57	0.7397	0.4187	Medium
Subject-independent (with subject-describing features)				
APTA	18.68 ± 8.44	0.6090	/	
LASSO	12.75 ± 9.01	0.7682	0.3468	Medium
EN	12.48 ± 8.28	0.7713	0.3707	Medium
XGB	12.41 ± 7.90	0.7741	0.4061	Medium
Subject-dependent				
APTA	7.39 ± 4.03	0.8500	/	
LASSO	7.50 ± 3.45	0.8657	0.0168	Small
EN	7.36 ± 3.40	0.9124	0.0719	Small
XGB	5.39 ± 2.04	0.9461	0.2900	Medium

Linear Regression with Elastic Net regularization (EN), Linear Regression with Least Absolute Shrinkage and Selection Operator regularization (LASSO), and Gradient Boosted Regression Trees (XGB) in the subject-independent and subject-dependent learning settings.

($p < 0.05$). A subject-independent model without subject-describing features resulted in the least accurate estimations of VILR (MAE: 12.71 ± 7.57 BW·s^{-1}; R^2: 0.7397). Including the subject's weight and shoe type improved the subject-independent model (MAE: 12.41 ± 7.90 BW·s^{-1}; R^2: 0.7741). Training a unique model for each subject further improved the predictions by a significant margin (MAE: 5.39 ± 2.04 BW·s^{-1}; R^2: 0.9461; $p < 0.01$).

Predictive Performance of the Single Metric Linear Regression Models

Table 3 shows the predictive performance of linear models that include a single feature in the subject-independent model learning setting. For comparison purposes was the predictive performance of the subject-independent XGB model added as well. Notwithstanding the moderate correlation between the APTA and the VILR, 32 of the extracted features had a higher predictive accuracy than the currently used proxy. Of these 32 features, the mean over the absolute differences between

TABLE 3 | Mean absolute error (MAE) ±SD and coefficient of determination R^2 scores for the estimation of the VILR by linear regression models using a single variable in the subject-independent model (SIM) learning setting.

Statistical model	MAE	R^2
APTA	21.07 ± 8.13	0.60
Standard deviation on linear trend	18.06 ± 7.28	0.67
Mean over the absolute differences between subsequent acceleration values	17.47 ± 7.98	0.71
SIM XGB model	12.41 ± 7.90	0.77

subsequent values of the vertical acceleration signal had the highest correlation with the VILR. A comprehensive overview of all 32 features was made available (**Supplementary Table A**). The previously discussed regression models that combine multiple of these features still outperform these single-feature models by a large margin.

Diagnostic Ability

The models' ability to identify runners with a high VILR is shown in **Figure 4**. With an area under the curve of 0.92, the subject-independent model XGB had a stronger diagnostic ability than the APTA which has an area under the curve of only 0.82.

Figure 5 shows cumulatively the percentage of predictions for which the relative error is below a threshold. The subject-independent model outperformed both baselines by a significant margin. However, the predicted VILR has still an error larger than 25% for 12% of the samples in the test set. The subject-dependent fails for only 3% of the examples.

Computing Time

The mean calculation time for each prediction was 142 ms (2.3 GHz Intel Core i5), of which the majority (140 ms) is spent on estimating the ground contact time. Meaning a prediction of the VILR can be made within one foot contact (160 - 350 ms).

DISCUSSION

The overall aim of this study was to predict the VILR during overground running by creating performant machine learning

FIGURE 4 | The Receiver Operating Characteristic curve reflects the ability of the subject-independent model XGB (SIM) and APTA models to identify runners with a high VILR. The sensitivity was plotted in function of the false positive rate (1 – specificity). The subject-independent model XGB model had a stronger diagnostic ability than the APTA.

models. Advanced signal processing was used to identify time-discrete features of the 3D acceleration waveforms that discriminate between subtle changes in running biomechanics. Machine-learned models were subsequently built to estimate the VILR and the performance (predictive accuracy, diagnostic ability) of those models were compared to a traditional approach. Two other machine learning techniques not discussed in this study were attempted, but gave unsatisfactory results. First, a data-driven deep recurrent neural network would require much more data than available to learn the complex relations between the tibial acceleration signals and VILR. Second, dynamic time warping was used as a tool for gait-curve matching, incorrectly assuming that runners with similar acceleration profiles have a similar VILR. Moreover, the feature engineering approach is preferable, since the learned models are interpretable (to a certain extend) and have a much lower computational cost.

The findings point out that applying machine learning to multiple 3D tibial acceleration features results in a more accurate prediction of the VILR than the frequently used APTA, which is a single time-discrete variable of tibial acceleration. Additionally, this prediction can be made in real-time, because the data pre-processing (i.e., filtering and feature construction) and prediction requires less calculation time than the typical duration of a single foot contact (~250 ms).

Overall, the XGB models systematically outperformed the other learners, suggesting that the XGB model can cope more effectively with the large number of features or that the relationship among the features and target are non-linear (Hepp et al., 2016).

From a machine learning setting perspective, building a subject-dependent model resulted in the most accurate predictions compared to the subject-independent models. The difference in predictive performance between the subject-independent model and subject-dependent model may partially be explained by the fact that all runners of the second cohort wore their own habitual running footwear, which might influence the measured impact loading. This assumption is further reinforced by the fact that the performance of the subject-independent model can be further improved by incorporating subject-describing features (body weight and shoe type). However, the phenotypical variability and choice of footwear can only partly explain the differences in accuracy between a subject-dependent and independent model. Although all runners ran in a similar environment, the ranked order of variable importance for predicting the VILR is unique for each runner in a subject-dependent learned model. Moreover, we observed a large asymmetry between the average VILR for most subject's left and right legs, suggesting that the subject-dependent models could be further improved by building separate models for both legs. However, in our study not mirroring the data resulted in a worse predictive accuracy due to the limited amount of data available for each subject.

The better predictive performance for a subject-dependent model compared to a subject-independent model is in line with previous findings described by Wouda et al. (2018) and Ahamed et al. (2019). However, our subject-independent model is more practical toward real-world applications. It is applicable
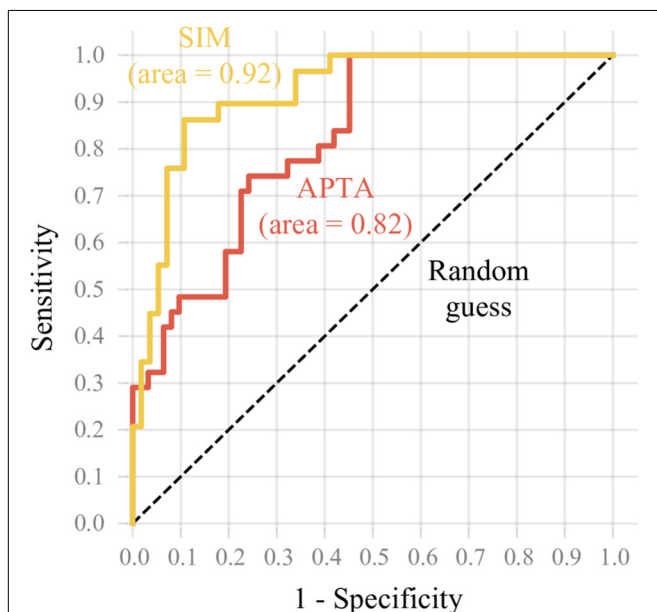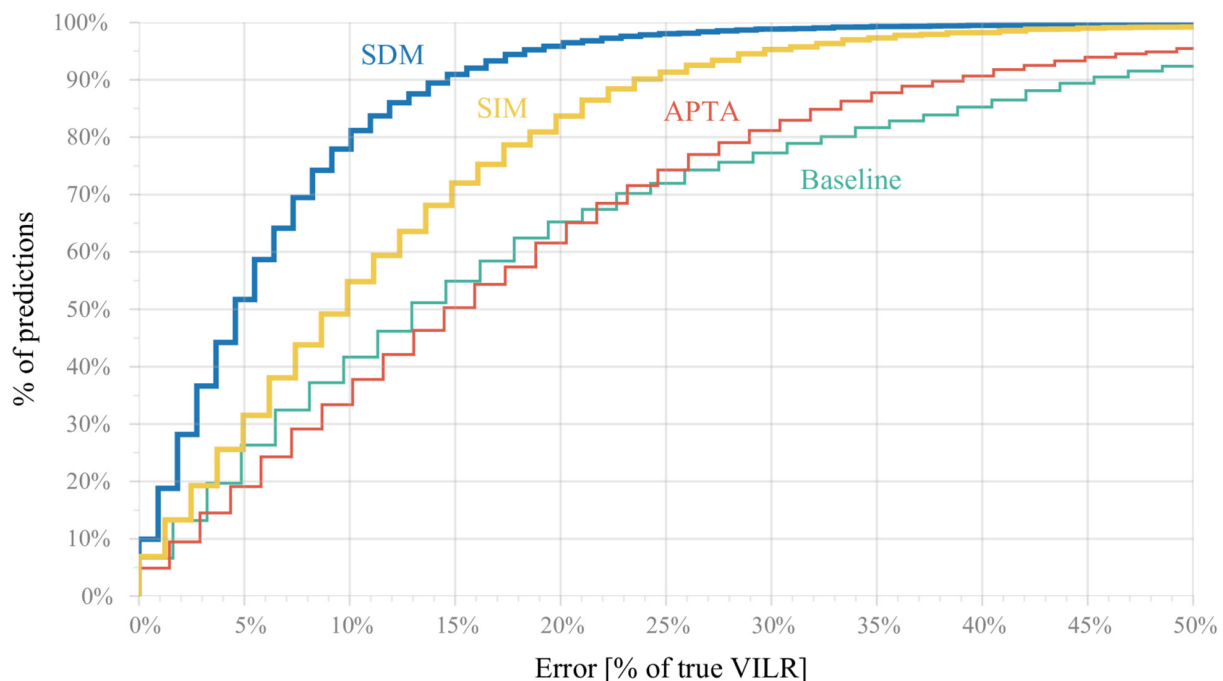
FIGURE 5 | Cumulative percentage (*y*-axis) of predictions for which the relative error is below a threshold (*x*-axis). For example, a relative error of at most 20% on the true VILR can be achieved for 97% of all predictions using the subject-dependent model (SDM), 83% of all strides using the subject-independent model (SIM), 66% using the APTA and 65% by predicting a runner's average max VILR for the corresponding landing foot (baseline model). The subject-independent model outperformed the baseline and APTA by a significant margin. Similarly, the subject-dependent model outperformed all others, but is less applicable in practice.

to any runner, regardless of whether prior data is available about the respective runner, which makes this approach generalizable over different subjects. Supporting our hypothesis, the subject-independent XGB model still outperformed the linear APTA model in terms of prediction accuracy and diagnostic ability.

By incorporating multiple running speeds we were able to create a machine learning algorithm that is capable of making accurate predictions across a broad range of running speeds, making it more usable in practice. As a consequence of this design choice, we observe relatively high R^2 scores for these models in comparison with previous research that considered a single running speed (Laughton et al., 2003; Pohl et al., 2009; Greenhalgh et al., 2012; Zhang et al., 2016) due to the restricted range effect (Bland and Altman, 2011): the inclusion of multiple speeds increases the range of the maximal VILR and makes it easier to see the global trend. However, this applies to all models discussed here and therefore does not affect the inter-model differences. For comparison, the evaluation metrics for all models trained on exclusively the most frequent running speed of 3.2 m·s^{-1} are provided as **Supplementary Table B**.

The VILR was predicted accurately, using a broad range of variables derived from filtered 3D accelerations. In order to screen runners on their VILR at a common training speed of 3.2 m·s^{-1} (e.g., identifying runners with a high VILR, during a simple overground running test without the need of an expensive force plate) the classification of runners on impact intensity is preferably done by estimating VILR by means of a machine-learned model instead of relying on the APTA only.

Because VILR is the maximum increase in acceleration of the lower extremity and of the rest of the body during stance (Clark et al., 2017), the predictive accuracy may be further improved by adding trunk acceleration to the accelerometer-derived input data.

This study has several limitations. Firstly, we trained the models only on habitual rearfoot strikers. Since machine learning can only be used to memorize patterns that are present in the training data, the trained models can only be applied to other rearfoot strikers and our findings do not necessarily generalize to other foot strike patterns. Secondly, all data was recorded in a laboratory environment. Previous research identified significant variations in APTA or contact time among different running surfaces (Tessutti et al., 2012; Boey et al., 2017). Hence, the findings should be transferred with caution to running on other surfaces.

CONCLUSION

This study proposes an advanced method to predict VILR during overground running by using only tri-axial shin mounted accelerometers derived data and an XGB machine learning approach. These algorithms, which incorporate time-continuous variables, are able to predict the VILR more accurately than currently possible using a time-discrete variable (e.g., APTA). Since these algorithms do not require significant computational power, they could be implemented on wearables worn by the

runner in order to screen, monitor or provide biofeedback on the predicted VILR whilst running overground.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available upon request to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ghent University Hospital Ethical Committee. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RD, PR, PB, DC, VS, and JD conceived, designed, and coordinated the study. RD, JG, and PB collected original data. PR and JD developed the machine learning algorithms. RD, PR, PB, and JG participated in data analysis. RD and PR developed the figures, initially drafted the manuscript, and the other authors provided useful suggestions in the preparation of the final manuscript.

All authors reviewed the manuscript and gave approval for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00033/full#supplementary-material

REFERENCES

Ahamed, N. U., Kobsar, D., Benson, L. C., Clermont, C. A., Osis, S. T., and Ferber, R. (2019). Subject-specific and group-based running pattern classification using a single wearable sensor. *J. Biomech.* 84, 227–233. doi: 10.1016/j.jbiomech.2019.01.001

Bland, J. M., and Altman, D. G. (2011). Correlation in restricted ranges of data. *BMJ* 342:d556. doi: 10.1136/bmj.d556

Boey, H., Aeles, J., Schütte, K., and Vanwanseele, B. (2017). The effect of three surface conditions, speed and running experience on vertical acceleration of the tibia during running. *Sports Biomech.* 16, 166–176. doi: 10.1080/14763141.2016.1212918

Bredeweg, S. W., Kluitenberg, B., Bessem, B., and Buist, I. (2013). Differences in kinetic variables between injured and noninjured novice runners: a prospective cohort study. *J. Sci. Med. Sport* 16, 205–210. doi: 10.1016/j.jsams.2012.08.002

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, (San Francisco, CA: ACM Press), 785–794. doi: 10.1145/2939672.2939785

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307, 72–77. doi: 10.1016/j.neucom.2018.03.067

Christ, M., Kempa-Liehr, A. W., and Feindt, M. (2017). "Distributed and parallel time series feature extraction for industrial big data applications," in *Proceedings of the Asian Machine Learning Conference (ACML) 2016, Workshop on Learning on Big Data (WLBD)*, Hamilton.

Clansey, A. C., Hanlon, M., Wallace, E. S., Nevill, A., and Lake, M. J. (2014). Influence of Tibial shock feedback training on impact loading and running economy. *Med. Sci. Sports Exerc.* 46, 973–981. doi: 10.1249/MSS.0000000000000182

Clark, K. P., Ryan, L. J., and Weyand, P. G. (2017). A general relationship links gait mechanics and running ground reaction forces. *J. Exp. Biol.* 220, 247–258. doi: 10.1242/jeb.138057

Clermont, C. A., Benson, L. C., Osis, S. T., Kobsar, D., and Ferber, R. (2018). Running patterns for male and female competitive and recreational runners based on accelerometer data. *J. Sports Sci.* 37, 204–211. doi: 10.1080/02640414.2018.1488518

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Abingdon: Routledge.

Crowell, H. P., and Davis, I. S. (2011). Gait retraining to reduce lower extremity loading in runners. *Clin. Biomech.* 26, 78–83. doi: 10.1016/j.clinbiomech.2010.09.003

Falbriard, M., Meyer, F., Mariani, B., Millet, G. P., and Aminian, K. (2018). Accurate estimation of running temporal parameters using foot-worn inertial sensors. *Front. Physiol.* 9:610. doi: 10.3389/fphys.2018.00610

Fukuchi, R. K., Eskofier, B. M., Duarte, M., and Ferber, R. (2011). Support vector machines for detecting age-related changes in running kinematics. *J. Biomech.* 44, 540–542. doi: 10.1016/j.jbiomech.2010.09.031

Gerritsen, K. G. M., van den Bogert, A. J., and Nigg, B. M. (1995). Direct dynamics simulation of the impact phase in heel-toe running. *J. Biomech.* 28, 661–668. doi: 10.1016/0021-9290(94)00127-P

Greenhalgh, A., Sinclair, J., Protheroe, L., and Chockalingam, N. (2012). Predicting impact shock magnitude: which ground reaction force variable should we use? *ISSN Int. J. Sports Sci. Eng.* 06, 1750–9823.

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression – a comparison between gradient boosting and the lasso. *Methods Inf. Med.* 55, 422–430. doi: 10.3414/ME16-01-0033

Holzreiter, S. H., and Köhle, M. E. (1993). Assessment of gait patterns using neural networks. *J. Biomech.* 26, 645–651. doi: 10.1016/0021-9290(93)90028-D

Janssen, D., Schöllhorn, W. I., Newell, K. M., Jäger, J. M., Rost, F., and Vehof, K. (2011). Diagnosing fatigue in gait patterns by support vector machines and self-organizing maps. *Hum. Mov. Sci.* 30, 966–975. doi: 10.1016/j.humov.2010.08.010

Kulmala, J.-P., Kosonen, J., Nurminen, J., and Avela, J. (2018). Running in highly cushioned shoes increases leg stiffness and amplifies impact loading. *Sci. Rep.* 8:17496. doi: 10.1038/s41598-018-35980-35986

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* 4:863. doi: 10.3389/fpsyg.2013.00863

Laughton, C. A., Davis, I. M., and Hamill, J. (2003). Effect of strike pattern and orthotic intervention on tibial shock during running. *J. Appl. Biomech.* 19, 153–168. doi: 10.1123/jab.19.2.153

Napier, C., MacLean, C. L., Maurer, J., Taunton, J. E., and Hunt, M. A. (2018). Kinetic risk factors of running-related injuries in female recreational runners. *Scand. J. Med. Sci. Sports* 28, 2164–2172. doi: 10.1111/sms.13228

Ngoh, K. J.-H., Gouwanda, D., Gopalai, A. A., and Chong, Y. Z. (2018). Estimation of vertical ground reaction force during running using neural network model and uniaxial accelerometer. *J. Biomech.* 76, 269–273. doi: 10.1016/j.jbiomech.2018.06.006

Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., and Davis, J. (2018). "Fatigue prediction in outdoor runners via machine learning and sensor fusion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, (London: ACM Press), 606–615. doi: 10.1145/3219819.3219864

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: machine learning in python. *J. Machine Learn. Res.* 12, 2825–2830.

Pohl, M. B., Hamill, J., and Davis, I. S. (2009). Biomechanical and anatomic factors associated with a history of plantar fasciitis in female runners. *Clin. J. Sport Med. Off. J. Can. Acad. Sport Med.* 19, 372–376. doi: 10.1097/JSM.0b013e3181b8c270

Preece, S. J., Bramah, C., and Mason, D. (2019). The biomechanical characteristics of high-performance endurance running. *Eur. J. Sport Sci.* 19, 784–792. doi: 10.1080/17461391.2018.1554707

Robberechts, P., Derie, R., den Berghe, P. V., Gerlo, J., De Clercq, D., Segers, V., et al. (2019). Gait event detection in tibial acceleration profiles: a structured learning approach. *arXiv* [Preprint] Available at: http://arxiv.org/abs/1910.13372 (accessed October 30, 2019).

Sheerin, K. R., Reid, D., and Besier, T. F. (2019). The measurement of tibial acceleration in runners—A review of the factors that can affect tibial acceleration during running and evidence-based guidelines for its use. *Gait Posture* 67, 12–24. doi: 10.1016/j.gaitpost.2018.09.017

Tessutti, V., Ribeiro, A. P., Trombini-Souza, F., and Sacco, I. C. N. (2012). Attenuation of foot pressure during running on four different surfaces: asphalt, concrete, rubber, and natural grass. *J. Sports Sci.* 30, 1545–1550. doi: 10.1080/02640414.2012.713975

Ueda, T., Hobara, H., Kobayashi, Y., Heldoorn, T. A., Mochimaru, M., and Mizoguchi, H. (2016). Comparison of 3 methods for computing loading rate during running. *Int. J. Sports Med.* 37, 1087–1090. doi: 10.1055/s-0042-107248

Van den Berghe, P., Six, J., Gerlo, J., Leman, M., and De Clercq, D. (2019). Validity and reliability of peak tibial accelerations as real-time measure of impact loading during over-ground rearfoot running at different speeds. *J. Biomech.* 86, 238–242. doi: 10.1016/j.jbiomech.2019.01.039

Van Der Worp, H., Vrielink, J. W., and Bredeweg, S. W. (2016). Do runners who suffer injuries have higher vertical ground reaction forces than those who remain injury-free? A systematic review and meta-analysis. *Br. J. Sports Med.* 50, 450–457. doi: 10.1136/bjsports-2015-094924

van Hees, V. T., Gorzelniak, L., Dean León, E. C., Eder, M., Pias, M., Taherian, S., et al. (2013). Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One* 8:e61691. doi: 10.1371/journal.pone.0061691

Williams, G., Lai, D., Schache, A., and Morris, M. E. (2015). Classification of gait disorders following traumatic brain injury. *J. Head Trauma Rehabil.* 30, E13–E23. doi: 10.1097/HTR.0000000000000038

Willy, R. W., Buchenic, L., Rogacki, K., Ackerman, J., Schmidt, A., and Willson, J. D. (2016). In-field gait retraining and mobile monitoring to address running biomechanics associated with tibial stress fracture: in-field gait retraining and monitoring. *Scand. J. Med. Sci. Sports* 26, 197–205. doi: 10.1111/sms.12413

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., van Beijnum, B.-J. F., et al. (2018). Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors. *Front. Physiol.* 9:218. doi: 10.3389/fphys.2018.00218

Zeng, W., Liu, F., Wang, Q., Wang, Y., Ma, L., and Zhang, Y. (2016). Parkinson's disease classification using gait analysis via deterministic learning. *Neurosci. Lett.* 633, 268–278. doi: 10.1016/j.neulet.2016.09.043

Zhang, J. H., An, W. W., Au, I. P. H., Chen, T. L., and Cheung, R. T. H. (2016). Comparison of the correlations between impact loading rates and peak accelerations measured at two different body sites: intra- and inter-subject analysis. *Gait Posture* 46, 53–56. doi: 10.1016/j.gaitpost.2016.02.002

Check for
updates

Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network

Marion Mundt[1]*, Arnd Koeppe[1], Sina David[2], Tom Witter[1], Franz Bamer[1], Wolfgang Potthast[2] and Bernd Markert[1]

[1] Institute of General Mechanics, RWTH Aachen University, Aachen, Germany, [2] Institute of Biomechanics and Orthopeadics, German Sport University Cologne, Cologne, Germany

Enhancement of activity is one major topic related to the aging society. Therefore, it is necessary to understand people's motion and identify possible risk factors during activity. Technology can be used to monitor motion patterns during daily life. Especially the use of artificial intelligence combined with wearable sensors can simplify measurement systems and might at some point replace the standard motion capturing using optical measurement technologies. Therefore, this study aims to analyze the estimation of 3D joint angles and joint moments of the lower limbs based on IMU data using a feedforward neural network. The dataset summarizes optical motion capture data of former studies and additional newly collected IMU data. Based on the optical data, the acceleration and angular rate of inertial sensors was simulated. The data was augmented by simulating different sensor positions and orientations. In this study, gait analysis was undertaken with 30 participants using a conventional motion capture set-up based on an optoelectronic system and force plates in parallel with a custom IMU system consisting of five sensors. A mean correlation coefficient of 0.85 for the joint angles and 0.95 for the joint moments was achieved. The RMSE for the joint angle prediction was smaller than $4.8°$ and the nRMSE for the joint moment prediction was below 13.0%. Especially in the sagittal motion plane good results could be achieved. As the measured dataset is rather small, data was synthesized to complement the measured data. The enlargement of the dataset improved the prediction of the joint angles. While size did not affect the joint moment prediction, the addition of noise to the dataset resulted in an improved prediction accuracy. This indicates that research on appropriate augmentation techniques for biomechanical data is useful to further improve machine learning applications.

Keywords: machine learning, artificial neural networks, wearable sensors, inertial sensors, motion analysis, data simulation

1. INTRODUCTION

Motion analysis, especially gait, in real-world environments gains more and more relevance in today's society. Since people are aging and want to retain their mobility, it is important to early detect abnormal gait patterns in order to prevent them from falling. To achieve this, the improvement of ambulatory motion analysis systems is relevant (Mundt et al., 2019a). Systems that are capable of determining motion kinematics and kinetics without expensive equipment and with less expert knowledge required will drastically increase the availability of motion analysis to a wider range of people. By providing wearable easy-to-use systems in daily life, risky motion patterns (e.g., in gait) might be identified before a major injury occurs or the onset of gait related diseases (Kobsar and Ferber, 2018; Majumder et al., 2019).

Gait is one of the main tasks of mobility. Baker et al. (2016) established four reasons for gait analysis: to diagnose a disease or injury, to assess the severity of a disease or injury, to monitor the progress of a disease or injury and to predict the outcome of an intervention. In all cases, long term or frequent monitoring of a person during daily life is desirable, thus allowing to identify any progression of a disease. To bring motion analysis into daily life, wearable sensors—especially inertial measurement units (IMUs)—have become increasingly popular (Caldas et al., 2017; Jarchi et al., 2018).

To extract joint angles from IMU data, the orientation of each sensor in a global reference system needs to be determined and a sensor-to-segment alignment performed. The most popular sensor fusion techniques for IMU-based motion analysis systems are (extended) Kalman filters or complementary filters (Gui et al., 2015). These filters fuse the signals of each single sensor of the IMU to determine its orientation. Either the data of the accelerometer and gyroscope only (Gui et al., 2015) or additionally the magnetometer data is used to identify the sensor orientation in a global reference system (Sabatini, 2006). The use of a magnetometer for the estimation of sensor orientation can be seen as a major limitation because magnetometers are highly susceptible to local disturbances in the magnetic field (de Vries et al., 2009; Teufl et al., 2018). Different attempts have been made either correcting magnetic disturbances or omitting the use of magnetometers at all (Ligorio and Sabatini, 2016; Teufl et al., 2018). However, another major issue of the commonly used approach is the (mal-)alignment of the sensor axes to physiological meaningful segment and rotation axes that define the anatomical model (Picerno, 2017; Robert-Lachaine et al., 2017; Mundt et al., 2019d). Several approaches have been suggested to overcome this problem: calibration postures or movements (Favre et al., 2009; Ferrari et al., 2010; Palermo et al., 2014), anatomical calibrations (Picerno et al., 2008; Bisi et al., 2015), post-trial calibration procedures (Hamacher et al., 2014; Li and Zhang, 2014) and more recently machine learning approaches (Zimmermann et al., 2018). While the use of calibration postures and movements will always be prone to errors because they are dependent on the execution of the subject (Seel et al., 2014; Picerno, 2017; Robert-Lachaine et al., 2017), the post-trial alignment prohibits fast data analysis. Therefore, the use of machine learning algorithms or the exploitation

of kinematic constraints seems to be most promising. The most recent advancements of the kinematic-constraint-based approaches (Laidig et al., 2017, 2019; Müller et al., 2017; Nowka et al., 2019) have not been evaluated on gait analysis. Seel et al. (2014) evaluated the knee and ankle joint sagittal plane angle achieving deviations to the gold standard of less then $1°$. Machine learning approaches have been undertaken by Findlow et al. (2008), Goulermas et al. (2008) achieving a mean correlation of about 0.70 for the sagittal plane joint angles. In recent work, we predicted joint angles based on simulated IMU data during gait achieving an accuracy higher than 0.86 (Mundt et al., 2019c).

Different approaches to determine the ground reaction force have been suggested and were systematically reviewed recently (Shahabpoor and Pavic, 2017). They concluded that the use of kinematic data as inputs reveals the highest practicality although showing lower accuracy than force plates. Additionally, the authors noted the limited validation of these methods for long-term measurements in real-life environment. This indicates that further research in this direction is useful, and if the aim is the evaluation of joint moments, a direct approach to determine these quantities might be advantageous. Different research has been undertaken in this direction, but less frequently. Ardestani et al. (2014) used a wavelet neural network to predict the 3D hip joint moments, the sagittal knee joint moment and the plantar flexion and eversion moment of the ankle joint during gait using GRF and EMG data as inputs. They reported normalized root-mean-squared errors of <20% and correlation coefficients ranging from 0.84 to 0.96. Johnson et al. (2018, 2019) used pre-trained convolutional neural networks for the prediction of the GRF and the knee joint moment during walking, running and sidestepping based on marker trajectories. They achieved a mean correlation higher than 0.85 for the knee joint moments and GRF. Analyzing normal gait, Hahn and O'Keefe (2008) estimated the sagittal plane lower limb joint moments based on demographic, anthropometric, kinematic, and EMG data. They achieved a coefficient of determination higher than 0.88, but they did not split their test set subject-wise, hence, data from subjects in the test set was also present in the training set. This leads to an improved accuracy (Saeb et al., 2017). Wouda et al. (2018) used inertial sensor data to determine the vertical GRF and the sagittal knee kinematics. For the joint angle prediction the correlation coefficient was larger than 0.83, and for the GRF larger than 0.90. In previous work, we used either marker trajectories or joint angles and the GRF as input data to predict all joint moments of the lower limbs during side stepping achieving a mean correlation higher than 0.86 (Mundt et al., 2019b). In a recent study, we used simulated IMU data to predict the joint moments during gait achieving a similar accuracy (Mundt et al., 2019c).

Despite the already good results, machine learning approaches have one important requirement: large datasets. These are - due to the novelty of the system - not openly available from IMU sensors. To overcome the lack of a large amount of data, their synthesizing is one reasonable solution (Young et al., 2011; Brunner et al., 2015; Zimmermann et al., 2018; Mundt et al., 2019c). Young et al. (2011) was the first who simulated IMU data from existing optical motion capture data to enlarge a

dataset for pose estimation. This approach was taken a step further and validated by Brunner et al. (2015) and Zimmermann et al. (2018). In previous work, we simulated IMU data from optical data as well, but only validated the simulation based on a single participant (Mundt et al., 2019c). In this study, the validation of the simulation is continued and IMU data that was simulated based on optoelectronic data as well as measured IMU data is used as input parameters to train fully-connected feedforward neural networks. To be independent of a homogeneous magnetic field, the magnetometer data is not considered as input but the 3D angular rates and linear accelerations only. The major advantages of the proposed method are that the anatomical model is implicitly learned, hence no calibration postures or movements are necessary, and that joint kinematics and kinetics can be determined. We aim to predict the joint angles and joint moments of the lower limbs during gait and hypothesize that the use of combined simulated and measured data will achieve a higher accuracy than the use of measured data only. Furthermore, we hypothesize that the additional noise in measured data caused by soft tissue movements will decrease the prediction accuracy. We aim to provide a first step into the direction of in-field gait analysis based on IMUs and artificial intelligence.

2. MATERIALS AND METHODS

An overview on the workflow of the proposed methodology is given in **Figure 1**.

2.1. Gold Standard Approach

All data used in this study was collected at the German Sport University Cologne. The studies were approved by the Ethical Committee of the German Sport University Cologne and all participants provided their informed written consent. The motion was recorded using an optoelectronic motion capture system (VICON™, MX F40, Oxford, UK, 100–125 Hz) and two force plates (Kistler Instrumente AG, Winterthur, Switzerland, 1,000 Hz). In all studies, the participants were equipped with 28 reflecting markers that were attached to bony landmarks as depicted in detail in Mundt et al. (2019d) to create a rigid body model of the lower limbs. The marker trajectories and GRF were filtered using a zero-lag second-order low-pass Butterworth filter with a cut-off frequency of 6 Hz (Robertson et al., 2013) prior to calculating the joint angles and joint moments of the lower limbs with an anatomical landmark scaled model (Lund et al., 2015) using the AnyBody Modeling System™ (Version 6.0, AnyBody Technology, Aalborg, Denmark). First, the kinematics are calculated using an overdetermined kinematic solver to optimize the markers using a least-squares approach. Afterwards, the models joint parameters are fitted to the subject-specific parameters before calculating the kinetics. All data was segmented into consistent sequences of 101 frames. For the kinematic data, full gait cycles were extracted based on an implementation of the foot contact algorithm proposed by Maiwald et al. (2009). For the joint moments a threshold-based segmentation of the stance phase was applied based on the force plate data. The joint moments were normalized to body height and weight of the participant.

2.2. Machine Learning Method
2.2.1. Data Simulation

To derive the simulated IMU data, first, the anatomical coordinate systems of the biomechanical model need to be set up, because these coordinate systems are translated and rotated to match possible sensor positions before the derivatives are calculated to display the acceleration and angular rate.

The joint origins and segment coordinate systems of the hip, knee and ankle joint are calculated based on the marker trajectories. The marker set is displayed in **Figure 2**. The joint centers for pelvis and ankles are based on the recommendations of the International Society of Biomechanics (ISB) (Wu et al., 2002). The hip joint center is defined as per (Harrington et al., 2007). The definition of the knee joint center is based on Pennock and Clark (1990). After this step, five coordinate systems, one for the pelvis, two for the thighs and shanks, respectively, are set up. For ease of calculation, the coordinate systems are transformed to quaternions (Solà, 2017), denoted by $\mathbf{q}_{seg}$. For this purpose, the Hamilton convention is used:

$$i^2 = -1, \ j^2 = -1, \ k^2 = -1 \text{ and } ijk = -1 \quad (1)$$

with i, j and k displaying the imaginary units of the quaternion. Any quaternion Q can thus be defined as:

$$Q = q_0 + iq_1 + jq_2 + kq_3, \quad (2)$$

with q_0 being the scalar part of the quaternion and $iq_1 + jq_2 + kq_3$ being the vector part. The quaternion can be interpreted as vector $\mathbf{q}$ in $\mathbb{R}^4$, which is defined as:

$$\mathbf{q} = \begin{pmatrix} \cos(\theta) \\ \mathbf{u}\sin(\theta) \end{pmatrix}, \qquad \|\mathbf{u}\| = 1, \quad (3)$$

where $\mathbf{u} = u_x i + u_y j + u_z k$ is a unit vector describing the rotation axis and θ is a scalar describing the rotation angle.

In the following step, the anatomical coordinate systems are translated and rotated to match possible initial sensor positions and orientations. The rotation between the segment and sensor orientation can be described by $\mathbf{q}_\phi$. The quaternion describing the orientation of the sensor in global space is calculated by:

$$\mathbf{q}_{sensor} = \mathbf{q}_{seg} \otimes \mathbf{q}_\phi, \quad (4)$$

with $\mathbf{q}_{sensor}$ describing the sensor orientation, $\mathbf{q}_{seg}$ describing the segment coordinate systems orientation in a global reference frame and $\mathbf{q}_\phi$ describing the global quaternion rotation. The translation $\mathbf{x}$ of the segment coordinate system to the sensor position $\hat{\mathbf{x}}$ is performed by:

$$\hat{\mathbf{x}} = \mathbf{x}_0 + (\mathbf{q}_{seg} \otimes \mathbf{x} \otimes \mathbf{q}_{seg}^*), \quad (5)$$

FIGURE 1 | Overview of the methods applied. To get the ground truth information on the joint angles and joint moments of the lower limbs, the gold standard approach using an optical motion capture system and force plates to collect the data is used. Based on this data, inverse dynamics simulations are undertaken to calculate the joint angles and joint moments. Using the proposed ML method, inertial data (angular rate ω and acceleration $\mathbf{a}$) is simulated from the optical data and used as inputs for an artificial neural network. Based on the ground truth joint angles and moments, the network learns the connection between the input and output data. The method is validated using an IMU system based on five sensors that are placed consistently with the simulated data.

with $\mathbf{x}$ and $\hat{\mathbf{x}}$ being pure quaternions, with their components x_0 and $\hat{x}_0 = 0$ yielding the sensor position in a global reference system and $\mathbf{q}_{seg}^*$ denoting the conjugate of $\mathbf{q}_{seg}$. The original position, the joint center, is defined by $\mathbf{x}_0$.

In the following step, the angular velocity ω of each sensor can be calculated as the numerical quaternion derivative of the sensor orientation $\mathbf{q}_{sensor}$. For ease of readability, the subscript is omitted in the following. All quaternions $\mathbf{q}$ display the sensor's orientation. For two consecutive orientations $\mathbf{q}_k$ and $\mathbf{q}_{k+1}$, the local rotation $\Delta \mathbf{q}_l$ of each sensor reads:

$$\Delta \mathbf{q}_l = \mathbf{q}_k^* \otimes \mathbf{q}_{k+1}, \qquad (6)$$

which leads to:

$$\omega = \frac{2}{\Delta t} \frac{\Delta \mathbf{q}_{lv}}{\|\Delta \mathbf{q}_{lv}\|} \arctan(\|\Delta \mathbf{q}_{lv}\|, \Delta q_{l0}). \qquad (7)$$

FIGURE 2 | Marker set and sensor placement. The markers in the front are displayed in red, the ones at the back are displayed in blue. The green boxes display the IMU sensors.
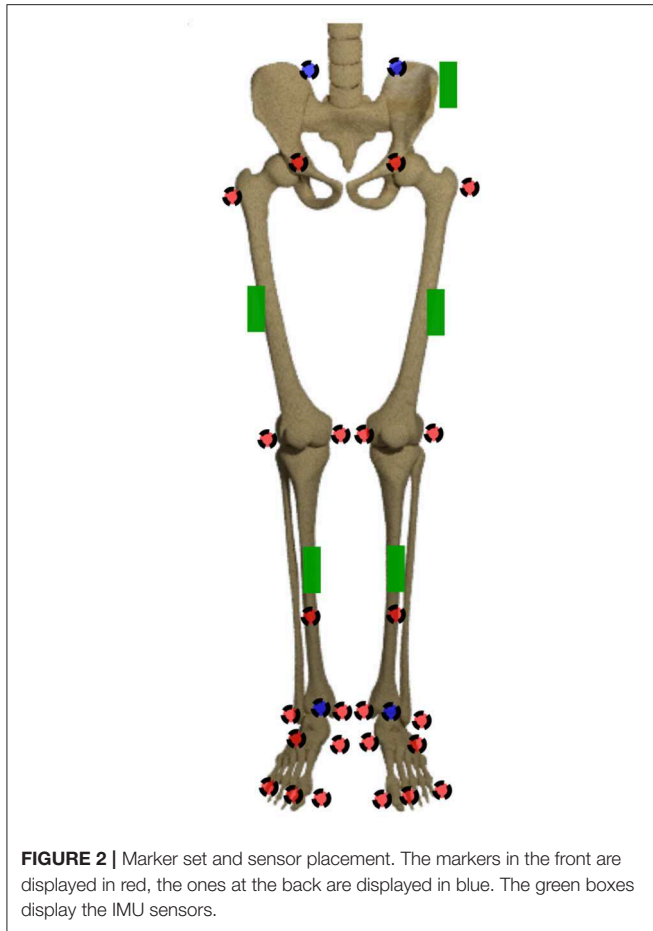
The subscripts v and 0 refer to the vector and the scalar part of the quaternion respectively. For further information on the derivation, see Solà (2017).

The linear acceleration of each sensor is calculated as the second derivative of the origin of the segment coordinate systems. This is achieved by reformulating a Taylor series expansion around $\mathbf{x}_k$ with $k = 2, ..., n-1$ with n being the last time step. For $\mathbf{x}_1$ and $\mathbf{x}_n$ the one-sided forward and backward differences need to be used respectively. The approximation can be improved for $\mathbf{x}_2$ to $\mathbf{x}_{n-1}$ by applying the central differences scheme (Atkinson and Han, 2005). In summary, this yields the following equations for the velocities:

$$\mathbf{v}_1 = \frac{\mathbf{x}_2 - \mathbf{x}_1}{\Delta t}, \qquad \mathbf{v}_k = \frac{\mathbf{x}_{k+1} - \mathbf{x}_{k-1}}{2\Delta t}, \qquad \mathbf{v}_n = \frac{\mathbf{x}_n - \mathbf{x}_{n-1}}{\Delta t}. \tag{8}$$

The same procedure can be applied again to derive the second order differentiation for $\mathbf{a}_k$. Analogously to $\mathbf{v}_k$, the same restrictions hold for the first and last time steps. Thus, the finite difference approximations of the accelerations are:

$$\mathbf{a}_1 = \frac{\mathbf{x}_3 - 2\mathbf{x}_2 + \mathbf{x}_1}{\Delta t^2}, \qquad \mathbf{a}_k = \frac{\mathbf{x}_{k+1} - 2\mathbf{x}_k + \mathbf{x}_{k-1}}{\Delta t^2},$$
$$\mathbf{a}_n = \frac{\mathbf{x}_n - 2\mathbf{x}_{n-1} - \mathbf{x}_{n-2}}{\Delta t^2}. \tag{9}$$

To transform the numerical derivatives into the actual sensor readings the different signs of gravity and motion need to be considered to define the acceleration of the sensor in the global reference system $\mathbf{a}_g$:

$$\mathbf{a}_g = -\mathbf{a} + \mathbf{g}. \tag{10}$$

The different signs are caused by the working principle of accelerometers that are used in inertial measurement units. Accelerometers are based on the inertial force of a small mass acting upon a piezoelectric element (Elwenspoek and Wiegerink, 2001). Thus, the gravitational acceleration directly translates to the sensor reading while the acceleration of the sensor origin results in an inertial force in the opposite direction of the segment acceleration. This means that the sign of a needs to be inverted while the sign of $\mathbf{g}$ remains unchanged. To describe the sensor readings in its local coordinate system the following transformation is necessary:

$$\mathbf{a}_l = \mathbf{q}^*_{sensor} \otimes \mathbf{a}_g \otimes \mathbf{q}_{sensor}, \tag{11}$$

with $\mathbf{a}_l$ displaying the linear acceleration of a sensor. The acceleration $\mathbf{a}_g$ and $\mathbf{a}_l$ are pure quaternions, with their components a_{g0} and $a_{l0} = 0$.

As the sensor is assumed to be a rigid body, its local position and orientation can be exactly described by six degrees of freedom, three translations and three rotations, described by the translation vector $\mathbf{x}$ (see Equation 5) and the rotation vector $\mathbf{q}_\phi$ (see Equation 4). In order to optimize these quantities, a vector $\mathbf{z} = [x_1, x_2, x_3, q_{\phi 1}, q_{\phi 2}, q_{\phi 3}]$ is defined. We fit the values using the following objective function:

$$\Theta(\mathbf{z}) = \sum_{n_t=1}^{N_t} (\boldsymbol{\omega}_{n_t}^{(m)} - \boldsymbol{\omega}_{n_t}^{(s)}(\mathbf{z}))^T (\boldsymbol{\omega}_{n_t}^{(m)} - \boldsymbol{\omega}_{n_t}^{(s)}(\mathbf{z}))$$
$$+ \sum_{n_t=1}^{N_t} (\mathbf{a}_{n_t}^{(m)} - \mathbf{a}_{n_t}^{(s)}(\mathbf{z}))^T (\mathbf{a}_{n_t}^{(m)} - \mathbf{a}_{n_t}^{(s)}(\mathbf{z})), \tag{12}$$

subject to,

$$x_{min} \leq x_i \leq x_{max}, \qquad i = 1, 2, 3 \tag{13}$$
$$0 \leq \phi_i \leq \frac{\pi}{2}, \qquad i = 1, 2, 3. \tag{14}$$

In this formulation, $\boldsymbol{\omega}_{n_t}^{(m)}$ and $\boldsymbol{\omega}_{n_t}^{(s)}(\mathbf{z})$ denote the angular rates in the three dimensional space. The superscripts (m) and (s) describe the measured and simulated values. Equivalently, $\mathbf{a}_{n_t}^{(m)}$ and $\mathbf{a}_{n_t}^{(s)}(\mathbf{z})$ denote the acceleration in three dimensions. We additionally defined the minimum and maximum allowed deviation of the positions x_{min} and x_{max} to be ± 50 mm as well as the maximum allowed orientation deviation $\frac{\pi}{2}$. The constrained optimization problem was solved using the interior-point algorithm (Byrd et al., 1999), which is implemented in

MATLAB. The defined constraints do not allow for an arbitrary sensor positioning but for a compensation for positioning and orientation errors in the range specified by the constraints. For this purpose, the sensor-to-segment-assignment needs to be consistent.

2.2.2. Neural Network Implementation

The python library Tensorflow (Abadi et al., 2015) was used to implement a fully-connected feedforward neural network (Koeppe et al., 2019). Artificial neural networks work as universal function approximators. Instead of explicitly programming the solution of one specific task, they learn from existing (big) data. Artificial neural networks have been inspired by the structure of the human brain, consisting of single neurons that add up to layers to increase the capacity of the network. Using multiple (hidden) layers with a specified number of neurons, the capacity of the network can be adapted (Mundt et al., 2019c). Fully-connected neural networks need time-normalized data as inputs, hence, only an offline analysis can be performed.

Different networks were trained for the prediction of joint kinematics and the joint kinetics based on different datasets. The first one is a collection of optical motion capture data of gait trials previously collected at the German Sport University Cologne. The dataset comprised 93 participants (38 female, 39.9 (18–75) years, 72.6 (47.1–97.6) kg, 1.73 (1.54–1.98) m, BMI 24.3 (17.5–31.6) kg m^{-2}). A number of 24 participants underwent knee arthroplasty 1.8 ± 0.4 years post-surgery prior to gait analysis (Komnik et al., 2015). The optical data collected in this study was additionally added to the dataset as well.

After validation of the simulated data, the neural network was trained using the accelerations and angular rates of the five sensors depicted in **Figure 2** as input data, which resulted in 30 inputs. One sensor was placed at the pelvis, one on each thigh and one on each shank. The sensors were not aligned to the segments, because the dataset is supposed to cover the complete range of orientations and positions due to the data simulation. Thereby, the neural network can learn to handle the differences. A kinematic model was trained to predict the 18 joint angles of the lower limbs, while a kinetic model was trained to predict the 18 joint moments of the lower limbs. Because we use a fully-connected feedforward neural network, no time-dependencies can be covered by the neural network (Goodfellow et al., 2016). Therefore, all data was time normalized and unrolled before being input to the network. This resulted in an input layer of size $30 \times 101 = 3{,}030$ and an output layer of $18 \times 101 = 1{,}818$. For the analysis, only the nine angles/moments of the foot touching the ground are evaluated.

In a first step, all simulated IMU data was used to determine the best network architecture and hyperparameters for the application using a 5-fold cross-validation. Therefore, one fixed test set was split from the complete dataset as well as five different validation sets. The split was undertaken randomly ensuring that no overlapping between the sets occurred (cf. **Figure 3**). A grid search was conducted to optimize the architectures and hyperparameters.

2.3. Validation

2.3.1. Experimental Set-Up

Thirty healthy subjects (12 female, 28.1 ± 6.0 years, 72.3 ± 12.7 kg, 1.77 ± 0.07 m) participated in this study that was approved by the Ethical Committee of the German Sport University Cologne. All participants provided their informed written consent. Each subject performed 10 level walking trials at five different speeds: 0.8 m s^{-1}, 1.1 m s^{-1}, 1.4 m s^{-1}, 1.7 m s^{-1} and 2.0 m s^{-1} ±10% on a 5 m walkway. According to the set-up of all previous experimental investigations, each participant was equipped with 28 retro-reflective markers to capture the motion by 12 infrared cameras (125 Hz, VICON^{TM}, MX F40, Oxford, UK). Simultaneously, the participants were equipped with five sensors of a custom low cost IMU system (100 Hz, TinyCircuits, Akron, OH, USA) with an associated microcontroller (Atmel ATmega328P) and a WIFI-board (Atmel ATWINC1500). An Android application was developed to collect the data on a smartphone (Mundt et al., 2018b). The marker set and sensor placement are displayed in **Figure 2**. The sensors were only roughly aligned to the segments but a consistent sensor-to-segment assignment was used. The data of seven subjects was excluded from this study due to connectivity issues, hence, data of 23 participants is presented.

2.3.2. Data Synchronization

The synchronization of the IMU system and the optoelectronic system cannot be performed automatically. Therefore, a synchronization algorithm was developed. For this purpose, the simulated medio-lateral acceleration of the pelvis was used. An average position and orientation estimation of the pelvis sensor was chosen. For the actual synchronization an optimization problem was defined. We obtained the minimization problem with the following mean-square objective function Υ:

$$\Upsilon(\delta) = \sum_{n_t=1}^{N_t} (a_{n_t}^{(m)} - a_{n_t}^{(s)})^2. \tag{15}$$

Here, $a_{n_t}^{(m)}$ and $a_{n_t}^{(s)}$ denote the measured and simulated acceleration of the pelvis in the medio-lateral direction. The value δ is the distance between the first local maximum peak in the measured and simulated data (cf. **Figure 4**). The start value for the optimization was chosen based on the output of the optical motion tracking system. The optimization problem is iteratively solved using the Nelder-Mead Simplex method (Lagarias et al., 1998) already implemented in MATLAB. After synchronization, the optical motion capture data and the inertial sensor data was segmented into steps as described in section 2.1.

2.3.3. Simulation Framework

First, the simulation framework was validated. For this purpose the optimum position and orientation of each sensor were determined for each trial. Hence the sensors were fixed once for each subject during the complete experiment, the best estimation was determined for each subject based on the root-mean-squared error. Afterwards, all data was simulated again based on the optimized values. Thereby, a valid solution representing the
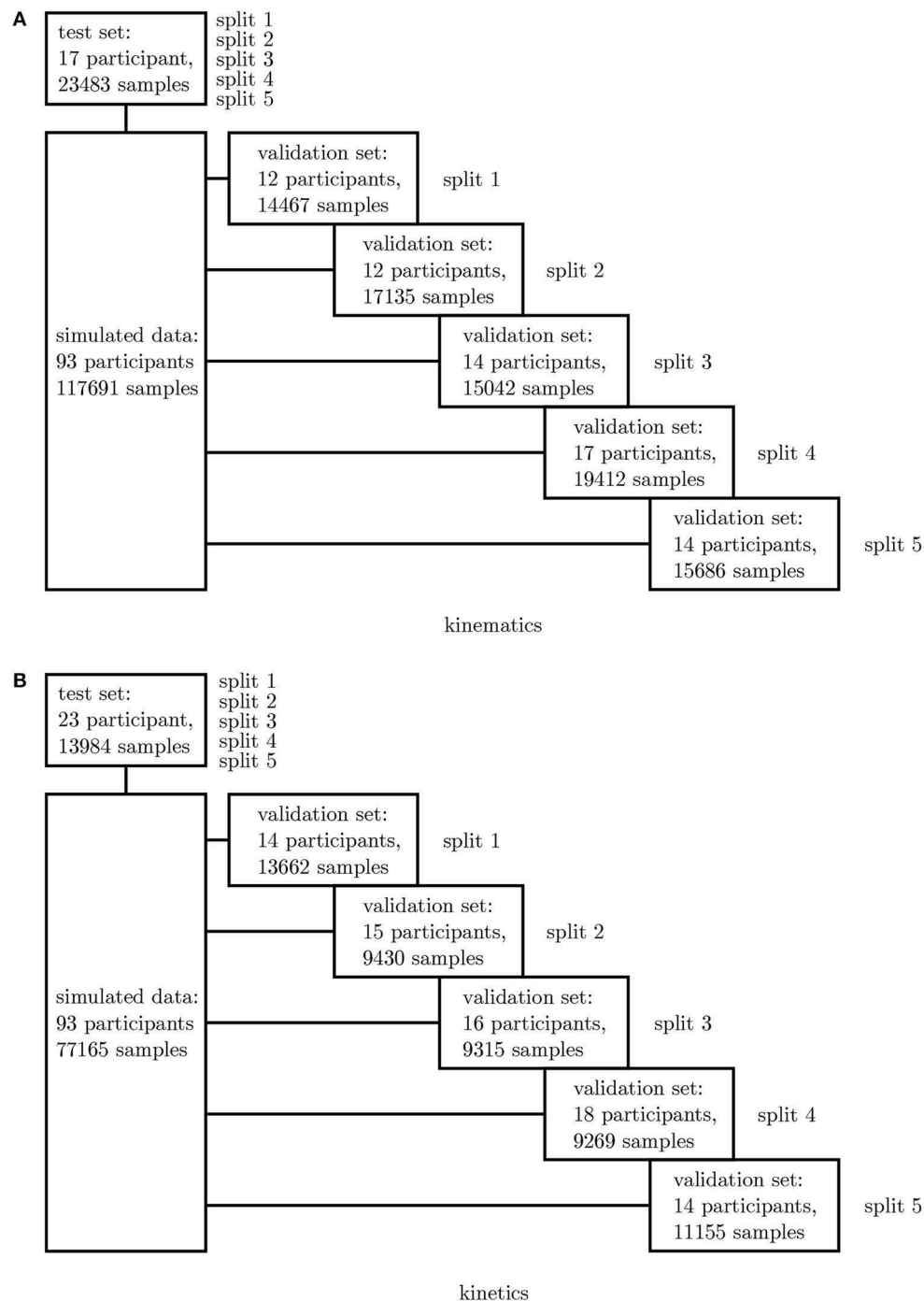
FIGURE 3 | Overview of the 5-fold cross-validation process. The dataset for the kinematics **(A)** and kinetics **(B)** differ and were treated separately.

differences in placement during the experiments was found. This procedure resulted in 23 (one per subject) optimized initial values. This information was used to generate a large simulated dataset based on the optical data of the former studies. All sensor positions and orientations found during the experiments were covered.

2.3.4. Neural Network Application
The inertial sensor data was used to validate the simulation framework presented. Afterwards, the neural network application was verified on the measured data. For this purpose, a leave-one-out cross-validation (Arlot and Celisse, 2010) was performed to enable the performance analysis of
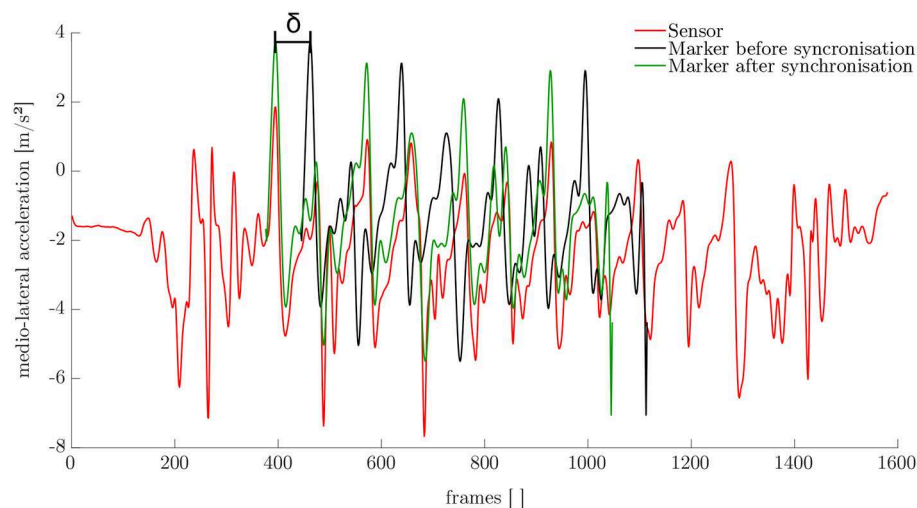
FIGURE 4 | Results of the synchronization based on the medio-lateral acceleration of the pelvis.

the neural network on single subjects. Two different scenarios were investigated: (1) all experimental data—besides one subject—was used for training purposes and (2) all experimental data—besides one subject—plus the simulated data was used for training. Since the best architecture and hyperparameters have been found in the first step, no further validation set is necessary. The left-out subject served as test set (cf. **Figure 5**).

2.4. Data Analysis

First, the results of the data simulation are presented. Afterwards, the results of the 5-fold cross-validation and the grid search to find the optimum neural network parameters are displayed. Finally, the results of the leave-one-out cross-validation based on measured data only and on the combined—measured and simulated—data are presented. To evaluate the simulation and prediction accuracy, the correlation coefficient was calculated. Furthermore, the RMSE was determined for the joint angles and the nRMSE (normalized RMSE to the range of the data) for the joint moments. A paired t-test and the effect size were calculated on the RMSE/nRMSE values. Additionally, the maximum joint angles and joint moments were calculated to evaluate the performance on this scalar parameter. An ANOVA and *post-hoc* t-test with Bonferroni correction as well as the effect size were calculated on the prediction of the maximum joint angles and joint moments. For each subject one mean step was considered.

3. RESULTS

3.1. Neural Network Parameters

The best parameters for the neural network were evaluated based on the simulated dataset. An initial learning rate $\alpha = 10^{-4}$ and an increasing batch size of 16-32-32-64 samples during the four phased training schedule performed best for both the kinematic and kinetic model. For the kinematic model,

two hidden layers with 4,000 and 6,000 neurones, a dropout rate of 20% and 12,500 training steps per phase revealed the highest accuracy. For the kinematic model, two hidden layers with 6,000 and 4,000 neurones, a dropout rate of 40% and a number of 15,000 training steps per phase showed the best results.

3.2. Data Simulation

The simulation of the data for all sensors was based on one fixed sensor position and orientation for each subject. The mean RMSE between the measured and simulated data is displayed in **Figure 6**. With an increase in gait velocity, the RMSE increased while the correlation coefficient decreased. The simulated data of the pelvis sensor achieved the highest accuracy ($r_{pelvis} = 0.95 \pm 0.08$), while the accuracy for the sensors of the legs is slightly lower ($r_{right\,thigh} = 0.88 \pm 0.12$, $r_{left\,thigh} = 0.91 \pm 0.08$, $r_{right\,shank} = 0.91 \pm 0.11$, $r_{left\,shank} = 0.92 \pm 0.10$).

3.3. Five-Fold Cross-Validation

To find the best model architecture and optimize the hyperparameters, a 5-fold cross-validation was undertaken using the simulated data of all subjects. The results for the kinematic and kinetic model are displayed in **Figure 7**. For both models the mean correlation coefficient was very similar on new test data (kinetics: 0.86, kinematics: 0.87). The prediction of the knee joint frontal plane angle and the transverse moment showed the weakest correlations, while the prediction of the joint moments showed the highest accuracy in all planes for the hip joint (>0.91) and the joint angle prediction in all joints in the sagittal plane (>0.95). Additionally, the kinetic predictions showed less outliers than the kinematic predictions. Over all, the RMSE of the joint angle prediction was smaller than $6.0°$ for all joints and motion planes with a mean value of $4.1°$ and the nRMSE of the joint moment prediction was smaller than 25.5% with a mean value of 15.5%.

FIGURE 5 | Overview of the leave-one-out validation process. Kinematics and kinetics were treated separately.
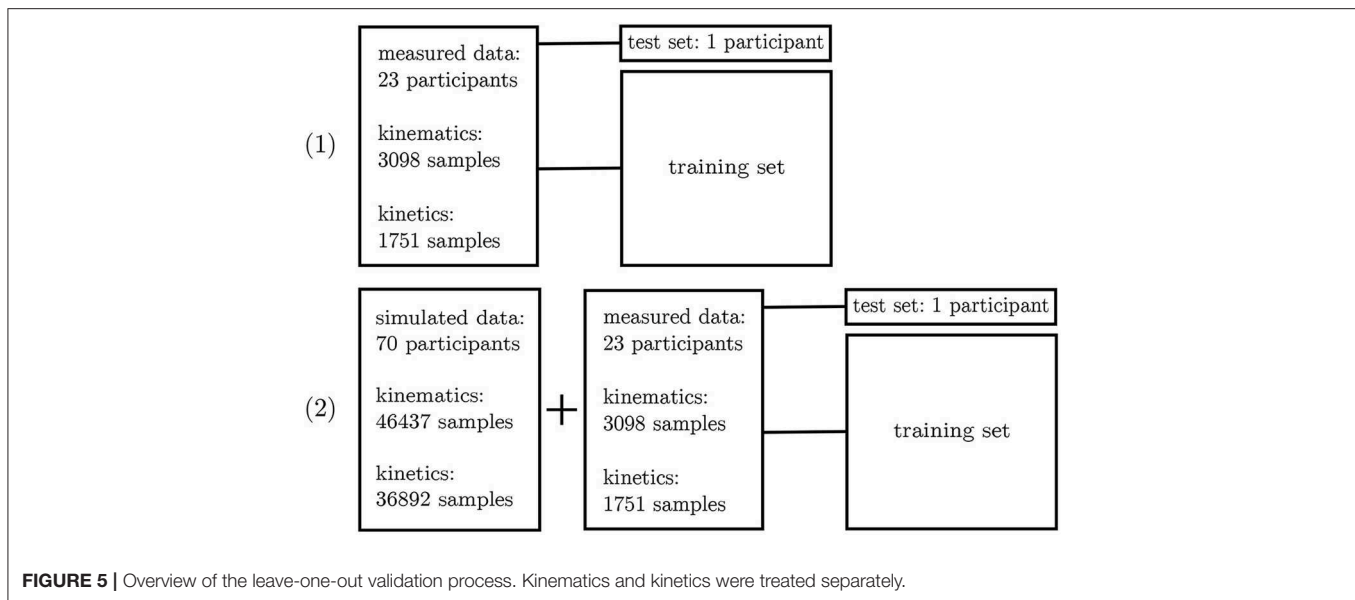
3.4. Leave-One-Out Cross-Validation

The leave-one-out cross-validation shows the performance of the model for each subject when trained on all other subjects' measured and simulated data. As the hyperparameters were fixed from the 5-fold cross-validation, no further validation set was necessary. There were only small differences in the correlation of the predicted and measured data using measured data only or the combined data set (cf. **Figure 8**). This finding was supported by the results of the *t*-test: there were differences between the two kinematic models in the prediction of all sagittal joint angles and the frontal hip joint angles ($r_{hip\ sagittal} < 0.001$, $r_{hip\ frontal} < 0.001$, $r_{knee\ sagittal} < 0.001$, $r_{ankle\ sagittal} = 0.043$). For the kinetic model, differences in all hip joint moments and the ankle joint sagittal moment were found ($r_{hip\ sagittal} < 0.001$, $r_{hip\ frontal} < 0.001$, $r_{hip\ transverse} < 0.001$, $r_{ankle\ sagittal} = 0.006$). Additionally, the correlation coefficient showed distinct differences between the motion planes: the prediction accuracy of the hip joint angle in the transverse plane, the knee joint angle in the frontal plane and the ankle joint angle in the frontal and transverse plane was lower than in the other planes. The prediction of the joint moments was more accurate although there were some subjects showing lower correlation coefficients in some features. The same behavior could be found when analyzing the distribution of the results (cf. **Figure 9**). Those parameters with weaker correlations showed a wider spread and more outliers in the distribution of the RMSE/nRMSE and correlation coefficient. The mean correlation of the models was $r_{kinematic\ measured} = 0.85$, $r_{kinematic\ combined} = 0.89$, $r_{kinetic\ measured} = 0.95$ and $r_{kinetic\ combined} = 0.95$. The mean error was $RMSE_{kinematic\ measured} = 4.8°$, $RMSE_{kinematic\ combined} = 4.3°$, $nRMSE_{kinetic\ measured} = 13.0\%$ and $nRMSE_{kinetic\ combined} = 11.6\%$.

Compared to the model used for the 5-fold cross-validation that was based on simulated data only, the accuracy was similar for the kinematic model and even higher for the kinetic model. With regard to **Figures 7**, **9**, the mean accuracy was similar for

FIGURE 6 | Root-mean-squared error between the measured and simulated data exemplarily displayed for the pelvis sensor. With an increasing gait velocity the simulation error increases. Some trials show outliers with larger errors during slow walking.
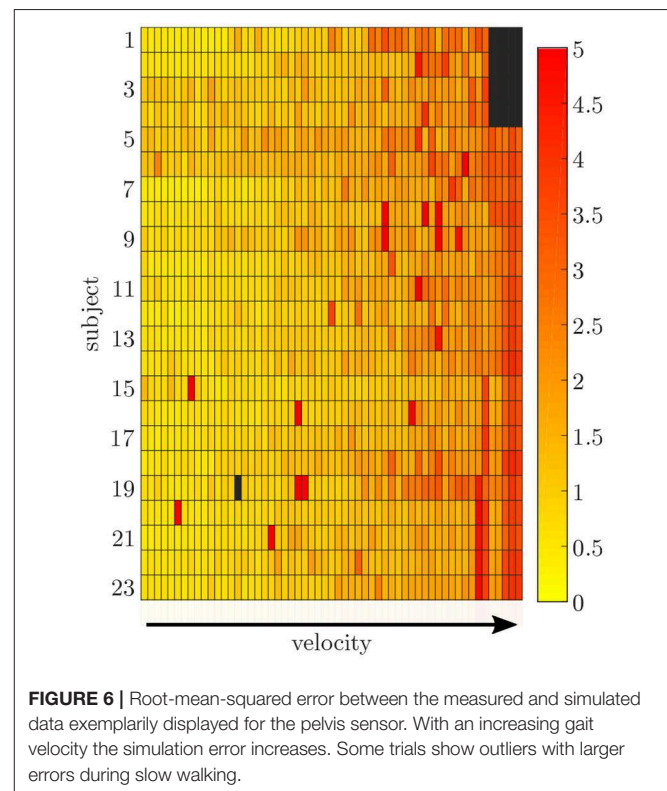
the cross-validation and the leave-one-out validation, but the number of outliers was decreased.

The ANOVA showed significant differences in the maxima between the predicted and measured joint angles and moments. The *post-hoc t*-test indicated significant differences between both the measured and predicted and the two predicted values. The prediction of the peak joint moments showed more significant
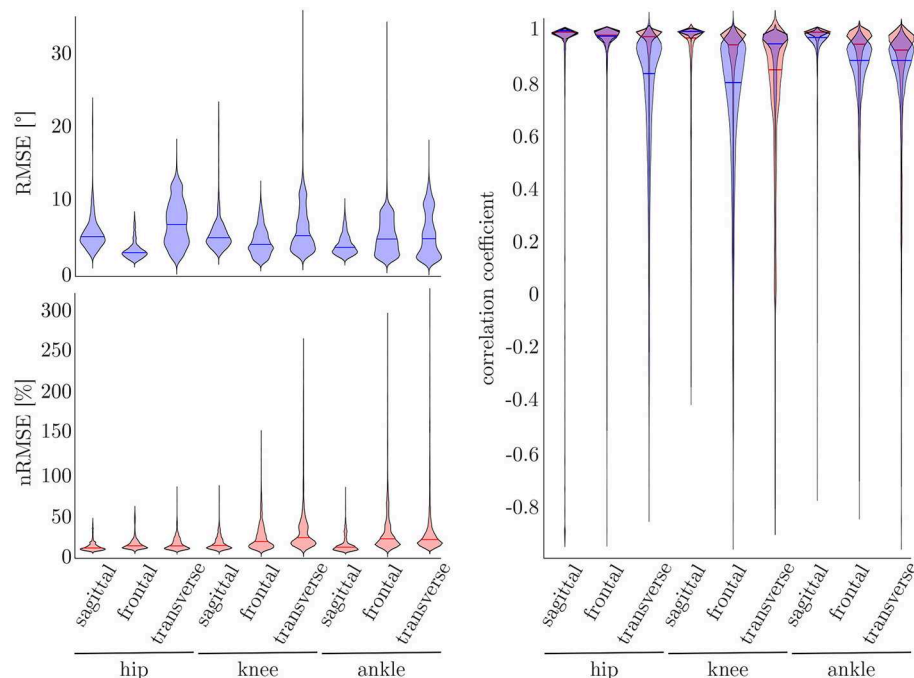
FIGURE 7 | On the right, the distribution of the correlation coefficient for the kinematic (blue) and kinetic (red) model is displayed. Additionally, on top, the distribution of the RMSE for the kinematic and on the bottom the distribution of the nRMSE for the kinetic model can be found. The violin's width displays how much data is accumulated, while the height shows the range of the distribution. The horizontal line indicates the median value of the distribution.

differences than the prediction of the peak joint angles (cf. **Table 1**). The trends of the statistical analysis can also be seen in **Figures 10**, **11**.

4. DISCUSSION

4.1. Data Processing

The data processing was one major challenge in this study because there was no possibility to synchronize both measurement systems automatically. The developed approach based on optimization might lead to errors, especially because gait is a cyclic motion. Sequences were not filtered for outliers, which might also cause outliers in the prediction. We decided not to remove outliers from the dataset to minimize the pre-processing on the data and observe the networks' capability to handle this data. The simulated data can represent the measured IMU data well, showing good correlations when using a fixed sensor position and orientation for the calculation of angular rate and linear acceleration for each subject. Higher gait velocities cause larger deviations between the measured and simulated data (cf. **Figure 6**), which can be attributed to soft tissue movements that cause noise in the measured data that is not included in the simulated data. The optical markers placed on bony landmarks are the basis for the simulated IMU data, while the physical IMU sensors are placed on the body as displayed in **Figure 2**. Hence, the markers and sensors experience different soft tissue movements that correlate with the gait velocity and increase the error. However, the mean correlation coefficients indicate an overall good accuracy. The correlation found by Young

et al. (2011), Brunner et al. (2015) and Zimmermann et al. (2018), who proposed frameworks for simulating IMU data, is comparable to the results presented in this study. Young et al. (2011) and Brunner et al. (2015) tested their simulator for leg swinging and single rigid body movements achieving very good correlation coefficients ($r > 0.97$). Since during this motion no impact occurs which causes soft tissue movements the results are better than the ones achieved in this study. Zimmermann et al. (2018) evaluated their simulation approach on a pure rigid body motion ($r > 0.97$) and during gait ($r_{acc} > 0.57$ and $r_{gyr} > 0.93$). These results support the explanation that the impact causes soft tissue movements during gait and limits the comparability between simulated and measured data. However, the results of this study are slightly better than in the study of Zimmermann et al. (2018).

4.2. Cross-Validation

The kind of validation strategy chosen can highly influence the results (Little et al., 2017; Saeb et al., 2017). We aimed to find the best model parameters and hyperparameters using a 5-fold cross-validation based on the simulated dataset only. Thereby, it was possible to exclude a fixed test set of a representative size as well as a randomly chosen validation set that covers most gait patterns. We ensured, that no data of any subject was part of more than one subset to avoid bias (Saeb et al., 2017). Simultaneously, it was possible to undertake a grid search on the best parameters and hyperparameters in a reasonable time frame. Afterwards, we performed a leave-one-out cross-validation on the new data collected in this study. This led to 23 training runs
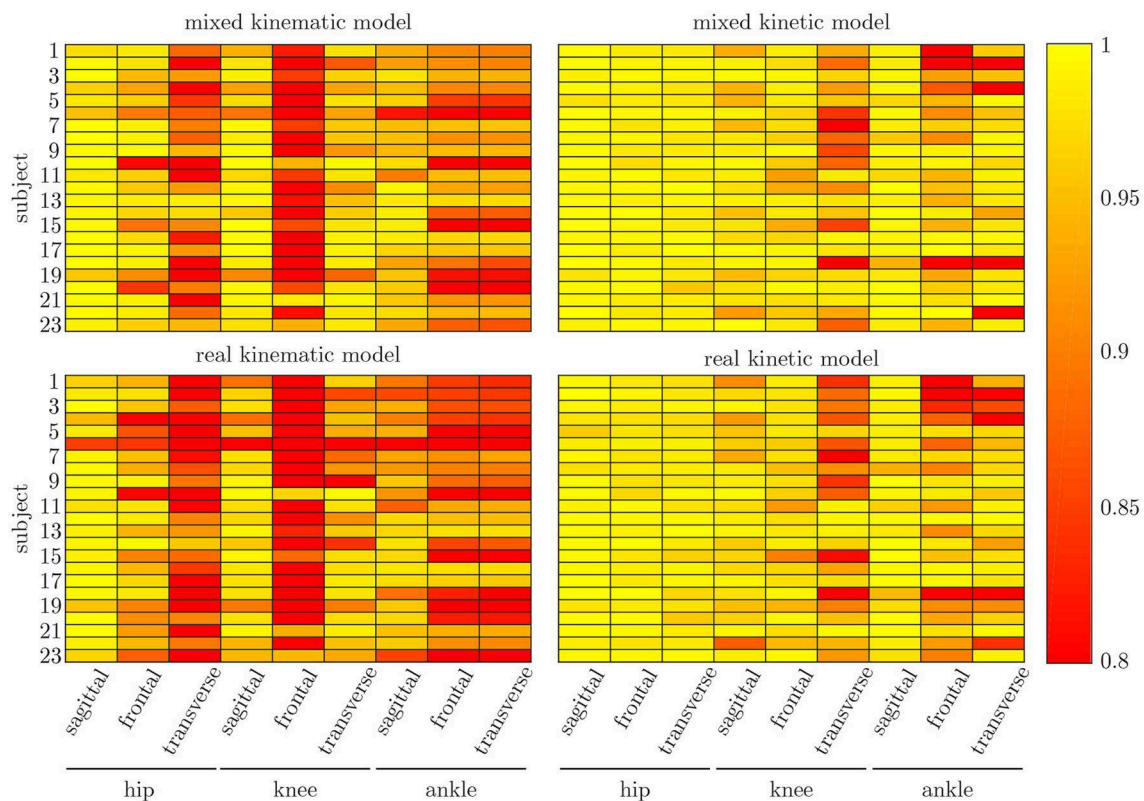
FIGURE 8 | Mean correlation coefficient for each joint and motion plane of each subject in the test set. On top, the results for the combined input data are displayed, while on bottom, the results of the model using measured data only are depicted. There are only small differences between both models, while there are distinct differences in the different motion planes and between subjects.

per model. Thereby we aimed to analyze the prediction accuracy on single subjects. Using this two-stage validation approach, it is possible to use as much data as possible for training the models, because there is no validation set necessary as the hyperparameters were fixed. As a side effect of the 5-fold cross-validation, it is possible to additionally compare the results of measured and combined input data to only simulated input data. However, this approach might also cause a suboptimal accuracy on the measured data, because the network architecture and hyperparameters were tuned to optimize the prediction on the simulated dataset only, which is larger than the measured dataset and the combined one.

4.3. Kinematics and Kinetics

The lower accuracy of the prediction of the kinematics indicates that it is a more difficult task for the neural network to predict the joint angles than the joint moments. This might be attributed to the closer physical relationship of acceleration and (normalized) joint moments. Additionally, the joint angles do not start at a value around zero, which leads to a more difficult initial value problem than for the joint moments. Therefore, the prediction of the kinematics profits from an enlarged dataset, which can be seen in the increased prediction accuracy from measured over combined to simulated data. In contrast, the kinetics prediction seems to improve with additional noise in the input

data instead of the larger dataset. This can be seen in the increased prediction accuracy in the combined and measured dataset compared to the simulated data only, which does not include the larger soft tissue movements the sensors experience in faster walking. Soft tissue movements also affect the calculation of joint angles and joint moments, which is a limitation in every motion analysis. One disadvantage of the simulated IMU data is that it does not include the same soft tissue movements as the marker trajectories.

Both, the kinematic and the kinetic model, are not able to cover the complete variance of the measured data (cf. **Figures 10**, **11**). This might be improved by further increasing the dataset and the noise of the inputs. Therefore, research on data augmentation should be further emphasized. The higher variance in the results of the cross-validation models compared to the leave-one-out model might be attributed to the dataset. The dataset used for the cross-validation includes participants with larger demographic differences as well as knee arthroplasty patients while the leave-one-out dataset comprises young participants without any impairment only. For the cross-validation, one test set was split from the complete dataset, while for the leave-one-out validation only the participants of this study served as test set (cf. **Figures 4**, **5**). Additionally, **Figure 7** displays each single trial while in **Figure 9** the mean results of each participant are displayed.

FIGURE 9 | On the right, the distribution of the RMSE and the correlation coefficient for the kinematic data is displayed. On the left, the distribution of the nRMSE and the correlation coefficient for the kinetic model can be found. The results for the measured data as inputs is displayed in red, while the results for the combined data inputs are displayed in blue. The violin's width displays how much data is accumulated, while the height shows the range of the distribution. The horizontal line indicates the median value of the distribution.

Comparing the results of this study to the literature is difficult, because this is—to the authors knowledge—the first time that IMU data was used to predict the joint angles and moments in all three motion planes. Especially the analysis of kinematic parameters using machine learning is not well investigated so far. As displayed in **Figure 8**, the correlation coefficient is larger than 0.8 in the sagittal plane for all subjects regarding the joint angles and even higher in all motion planes regarding the joint moments. Only Findlow et al. (2008) used an approach based on neural networks to predict joint angles. We achieved a higher accuracy in our study, which is probably caused by the larger dataset, more sensors involved and an improved computing power and algorithms compared to their study undertaken in 2008. Another approach is the use of kinematic constraints to determine joint angles from IMU data. Based on different joints, this approach reveals very good results (Müller et al., 2017; Laidig et al., 2019; Nowka et al., 2019). Nevertheless, it was not analyzed recently for gait analysis. Seel et al. (2014) achieved already good results when analyzing the sagittal knee and ankle joint angles with an mean RMSE of 3.3° and 1.6°, respectively. These results are slightly better than our results with 4.62° and 2.42°. It might be possible to improve the accuracy of the proposed method when also specializing on single joint angles or adding additional sensors to the model. In a previous study, we could achieve an error smaller than 2.5° in all joints and motion planes, when using simulated data only and additional data for the feet sensors (Mundt et al., 2019c). Zihajehzadeh and Park (2017) used a more common approach for the joint angle estimation based on an adapted Kalman Filter that does

not use magnetometer data for the orientation estimation. They achieved RMSE values smaller than 3.5° for the three sagittal plane angles and the hip adduction/abduction during walking. Teufl et al. (2018) also investigated the use of magnetometer free joint angle estimation. Their method achieved mean RMSE values of <2.3° for all joints and motion planes. This results is very promising, although it needs to be considered that the biomechanical model was set up using optical motion capture data. In a previous study, we could show that the differences in joint angle estimation is mainly based on the definition of the rotation axes used by the IMU systems (Mundt et al., 2019d), what we aimed to overcome with the neural network approach that implicitly learns the biomechanical model during the training process.

There is more research on estimating joint kinetics, but none was undertaken using IMU sensors as input data to predict all 3D lower limb joint moments. In one of our previous studies, we used joint angles as input parameters to predict joint moments (Mundt et al., 2018a). We achieved slightly better results than in this study. In another study using simulated IMU data, the joint moment prediction resulted in an nRMSE of 12.16%, which is slightly lower than in this study although using additional sensors on the feet (Mundt et al., 2019c). This further supports the hypothesis that more noise in the data is favorable for the joint moment prediction and that it might be useful to investigate the relevant features for the neural network. For this purpose, Horst et al. (2019) suggested to use the Layer-Wise Relevance Propagation technique. Another approach might be the use of principle component analysis to

TABLE 1 | Results of the statistical analysis of the peak prediction.

	Joint angles					
	Measured vs. real		Measured vs. combined		Real vs. combined	
	p	d	*p*	d	*p*	d
Hip sagittal	<0.001*	1.867	<0.001*	1.392	0.004*	0.676
Hip frontal	<0.001*	1.293	<0.001*	0.948	<0.001*	0.924
Hip transverse	0.085	0.377	0.226	0.260	0.080	0.383
Knee sagittal	<0.001*	2.580	<0.001*	1.683	<0.001*	1.103
Knee frontal	0.324	0.210	0.566	0.121	0.061	0.412
Knee transverse	0.123	0.335	0.104	0.354	0.724	0.074
Ankle sagittal	0.005*	0.645	0.018	0.536	0.013*	0.561
Ankle frontal	0.412	0.174	0.211	0.269	0.141	0.319
Ankle transverse	0.035*	0.470	0.174	0.293	0.006*	0.638

	Joint moments					
	Measured vs. real		Measured vs. combined		Real vs. combined	
	p	d	*p*	d	*p*	d
Hip sagittal	<0.001*	1.659	<0.001*	1.033	<0.001*	1.452
Hip frontal	<0.001*	2.283	<0.001*	1.850	<0.001*	1.036
Hip transverse	0.004*	0.674	0.010*	0.592	0.021	0.517
Knee sagittal	<0.001*	0.854	0.004*	0.671	0.011*	0.576
Knee frontal	<0.001*	1.193	<0.001*	0.961	0.003*	0.690
Knee transverse	<0.001*	0.977	0.002*	0.735	0.031	0.482
Ankle sagittal	<0.001*	1.296	<0.001*	0.901	0.001*	0.839
Ankle frontal	0.041	0.452	0.108	0.350	0.078	0.386
Ankle transverse	0.038	0.460	0.079	0.384	0.087	0.373

*Significant results are indicated by *.

FIGURE 10 | Overview of the mean and standard deviation of the joint moments of the 23 subjects.
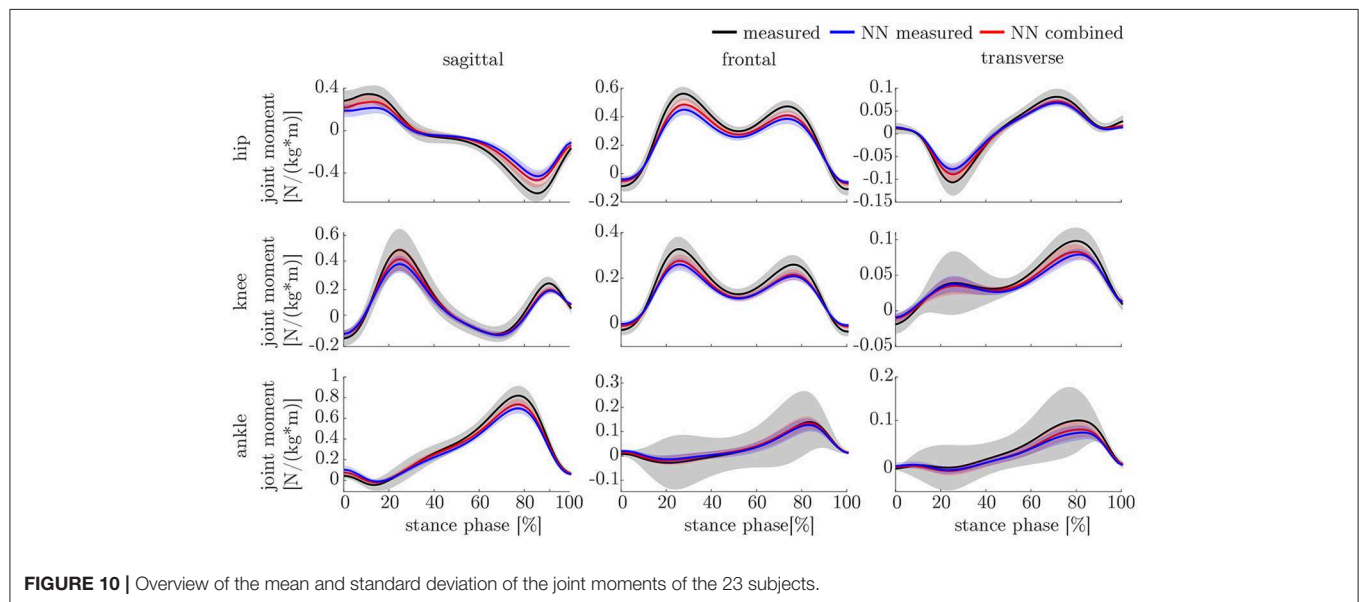
analyze the sensitivity of inputs and outputs (Ardestani et al., 2015).

In future work, it might be useful to investigate a two-staged approach: first, predict the joint angles from IMU data and second, use the estimated joint angles to predict the joint moments. However, for this approach the joint angle estimation needs to be further improved. It might also be conceivable to take this approach the other way round, using joint moments

FIGURE 11 | Overview of the mean and standard deviation of the joint angles of the 23 subjects.
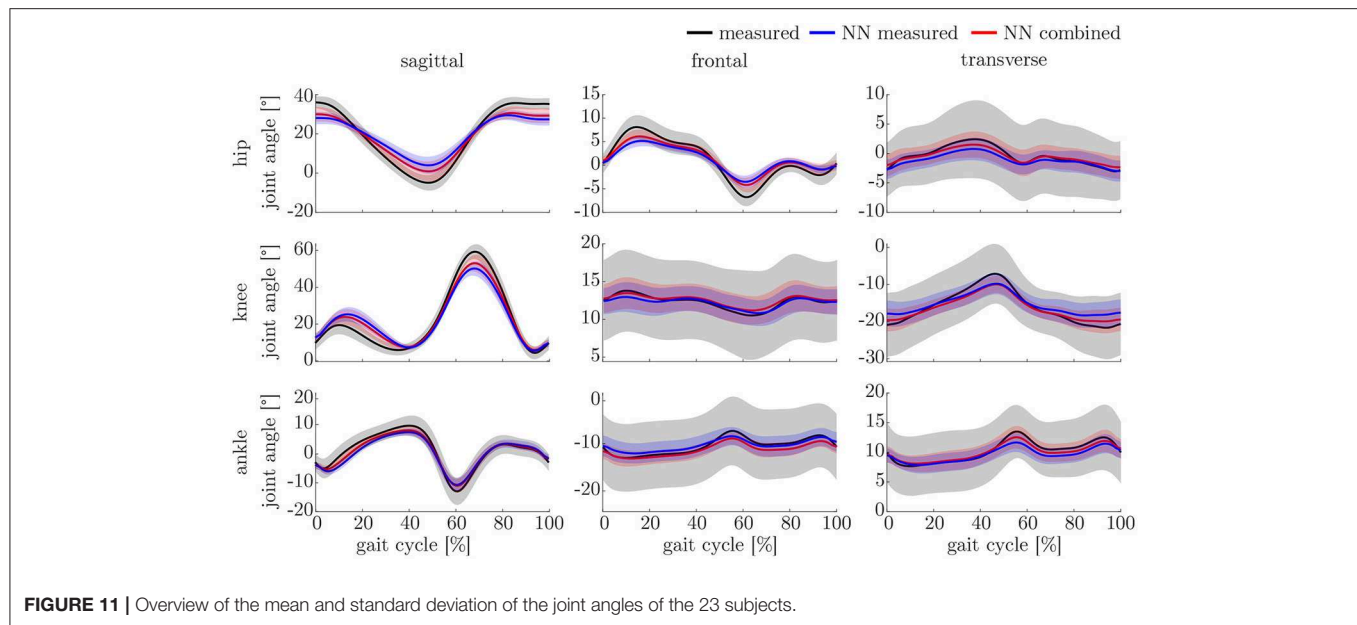
as input data to predict joint angles, because the joint moments show better results so far. It might also be feasible to add (estimated) joint angles or joint moments to the IMU input data for further improvement. Additionally, the choice of another kind of artificial neural network, e.g., long short-term memory (LSTM) or convolutional neural networks (CNN), might be suitable for the underlying task. Especially due to the high number of inputs (30 features times 101 time frames) these neural networks might outperform the fully-connected feedforward neural network, that was used in this study. While a fully-connected feedforward neural network uses flattened data (no time dependency) as inputs, LSTMs and CNNs preserve the time dependency. Thereby, it might be easier for these networks to extract the most relevant features from the data (Goodfellow et al., 2016). In this study, we analyzed short sequences of motion only. During these sequences, no gyroscope drift could be observed. For future research, to bring this method further toward application, this aspect needs to be considered. Another sensor system might overcome this limitation. We also only analyzed straight walking. Most probably, this method can also be applied to more diverging motion, when this motion is present in the training dataset. It might even lead to an improved accuracy, when using a dataset showing more variance (Mundt et al., 2019b). Further analysis on the relevant features for the neural network to predict the joint angles and moments will be valuable to maybe reduce the number of sensors necessary for the prediction and thereby decreasing the complexity of the model. Further validation of the method with a larger amount of measured data should be undertaken.

5. CONCLUSION

This study analyzed the ability of a fully-connected feedforward neural network to predict joint angles and joint moments of

the lower limbs based on IMU data. Our hypothesis, that simulated data can support the learning of the neural network can be accepted for the joint angle prediction while it can only be partly accepted for the prediction of joint moments. Our second hypothesis, that noise in the input data decreases the prediction accuracy can be rejected. For the kinetic prediction, the noise attributed to soft tissue movements improves the prediction accuracy and seems to be more important than the size of the dataset. The prediction of the joint angles is not affected by noise. Therefore, it needs to be evaluated if the prediction can be further improved using a simulated dataset containing soft tissue movement induced noise in the input data. Thereby, the measured data might be better represented and the learning of the neural network improved. Nevertheless, the results already demonstrate the high potential of the approach and support further research on neural networks in gait analysis. Besides the aforementioned data augmentation, different kind of neural networks (e.g., recurrent or convolutional neural networks) should be investigated on the task in future work. Thereby, data that is not time normalized could be used, hence the gait velocity could be included in the data. For the analysis of clinically relevant parameters, it might also be suitable to train patient-specific models to achieve a higher accuracy (Saeb et al., 2017).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission der Deutschen Sporthochschule

Köln. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MM and SD carried out the experiment. MM, AK, and TW undertook the data analysis. MM and FB wrote the manuscript.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems*. Available online at: tensorflow.org (accessed August 31, 2019).

Ardestani, M. M., Moazen, M., and Jin, Z. (2015). Sensitivity analysis of human lower extremity joint moments due to changes in joint kinematics. *Med. Eng. Phys.* 37, 165–174. doi: 10.1016/j.medengphy.2014.11.012

Ardestani, M. M., Zhang, X., Wang, L., Lian, Q., Liu, Y., He, J., et al. (2014). Human lower extremity joint moment prediction: a wavelet neural network approach. *Exp. Syst. Appl.* 41, 4422–4433. doi: 10.1016/j.eswa.2013.11.003

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.* 4, 40–79. doi: 10.1214/09-SS054

Atkinson, K., and Han W. (2009). *Theoretical Numerical Analysis: A Functional Analysis Framework*. New York, NY: Springer–Verlag.

Baker, R., Esquenazi, A., Benedetti, M. G., and Desloovere, K. (2016). Gait analysis: clinical facts. *Eur. J. Phys. Rehabil. Med.* 52, 560–574.

Bisi, M. C., Stagni, R., Caroselli, A., and Cappello, A. (2015). Anatomical calibration for wearable motion capture systems: video calibrated anatomical system technique. *Med. Eng. Phys.* 37, 813–819. doi: 10.1016/j.medengphy.2015.05.013

Brunner, T., Lauffenburger, J. P., Changey, S., and Basset, M. (2015). Magnetometer-augmented IMU simulator: in-depth elaboration. *Sensors* 15, 5293–5310. doi: 10.3390/s150305293

Byrd, R., Hribar, M., and Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optimiz.* 9, 877–900.

Caldas, R., Mundt, M., Potthast, W., Buarque de Lima Neto, F., and Markert, B. (2017). A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait Post.* 57, 204–210. doi: 10.1016/j.gaitpost.2017.06.019

de Vries, W. H. K., Veeger, H. E. J., Baten, C. T. M., and van der Helm, F. C. T. (2009). Magnetic distortion in motion labs, implications for validating inertial magnetic sensors. *Gait Post.* 29, 535–541. doi: 10.1016/j.gaitpost.2008.12.004

Elwenspoek, M., and Wiegerink, R. (2001). "Mechanical microsensors," in *Microtechnology and MEMS* (Berlin; Heidelberg: Springer-Verlag), 295.

Favre, J., Aissaoui, R., Jolles, B. M., de Guise, J. A., and Aminian, K. (2009). Functional calibration procedure for 3D knee joint angle description using inertial sensors. *J. Biomechan.* 42, 2330–2335. doi: 10.1016/j.jbiomech.2009.06.025

Ferrari, A., Cutti, A. G., Garofalo, P., Raggi, M., Heijboer, M., Cappello, A., et al. (2010). First *in vivo* assessment of "outwalk": a novel protocol for clinical gait analysis based on inertial and magnetic sensors. *Med. Biol. Eng. Comput.* 48, 1–15. doi: 10.1007/s11517-009-0544-y

Findlow, A. H., Goulermas, J. Y., Nester, C. J., Howard, D., and Kenney, L. P. J. (2008). Predicting lower limb joint kinematics using wearable motion sensors. *Gait Post.* 28, 120–126. doi: 10.1016/j.gaitpost.2007.11.001

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Goulermas, J. Y., Findlow, A. H., Nester, C. J., Liatsis, P., Zeng, X. J., Kenney, L. P., et al. (2008). An instance-based algorithm with auxiliary similarity information for the estimation of gait kinematics from wearable sensors. *IEEE Trans. Neural Netw.* 19, 1574–1582. doi: 10.1109/TNN.2008.2000808

Gui, P., Tang, L., and Mukhopadhyay, S. (2015). "MEMS based IMU for tilting measurement: comparison of complementary and kalman filter based data

fusion," *Proceedings of the 2015 10th IEEE Conference on Industrial Electronics and Applications, ICIEA 2015* (Auckland), 2004–2009.

Hahn, M. E., and O'Keefe, K. B. (2008). A neural network model for estimation of net joint moments during normal gait. *J. Musculoskel. Res.* 11, 117–126. doi: 10.1142/S0218957708002036

Hamacher, D., Hamacher, D., Taylor, W. R., Singh, N. B., and Schega, L. (2014). Towards clinical application: repetitive sensor position re-calibration for improved reliability of gait parameters. *Gait Post.* 39, 1146–1148. doi: 10.1016/j.gaitpost.2014.01.020

Harrington, M. E., Zavatsky, A. B., Lawson, S. E. M., Yuan, Z., and Theologis, T. N. (2007). Prediction of the hip joint centre in adults, children, and patients with cerebral palsy based on magnetic resonance imaging. *J. Biomechan.* 40, 595–602. doi: 10.1016/j.jbiomech.2006.02.003

Horst, F., Lapuschkin, S., Samek, W., Müller, K. R., and Schöllhorn, W. I. (2019). Explaining the unique nature of individual gait patterns with deep learning. *Sci. Rep.* 9:2391. doi: 10.1038/s41598-019-38748-8

Jarchi, D., Pope, J., Lee, T. K., Tamjidi, L., Mirzaei, A., and Sanei, S. (2018). A review on accelerometry-based gait analysis and emerging clinical applications. *IEEE Rev. Biomed. Eng.* 11, 177–194. doi: 10.1109/RBME.2018.2807182

Johnson, W. R., Member, S., Alderson, J., Lloyd, D., and Mian, A. (2019). Predicting athlete ground reaction forces and moments from spatio-temporal driven CNN models. *IEEE Trans. Biomed. Eng.* 66, 689–694. doi: 10.1109/TBME.2018.2854632

Johnson, W. R., Mian, A., Lloyd, D. G., and Alderson, J. A. (2018). On-field player workload exposure and knee injury risk monitoring via deep learning. *J. Biomechan.* 93, 185–193. doi: 10.1016/j.jbiomech.2019.07.002

Kobsar, D. and Ferber, R. (2018). Wearable sensor data to track subject-specific movement patterns related to clinical outcomes using a machine learning approach. *Sensors* 18:E2828. doi: 10.3390/s18092828

Koeppe, A., Bamer, F., and Markert, B. (2019). An efficient monte carlo strategy for elasto-plastic structures based on recurrent neural networks. *Acta Mechan.* 230, 3279–3293. doi: 10.1007/s00707-019-02436-5

Komnik, I., Weiss, S., Fantini Pagani, C., and Potthast, W. (2015). Motion analysis of patients after knee arthroplasty during activities of daily living - a systematic review. *Gait Post.* 41, 370–377. doi: 10.1016/j.gaitpost.2015.01.019

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optimiz.* 9, 112–147.

Laidig, D., Lehmann, D., Begin, M.-A., and Seel, T. (2019). "Magnetometer-free realtime inertial motion tracking by exploitation of kinematic constraints in 2-DoF joints," in *41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Berlin), 1233–1238.

Laidig, D., Müller, P., and Seel, T. (2017). Automatic anatomical calibration for IMU- based elbow angle measurement in disturbed magnetic fields. *Curr. Direct. Biomed. Eng.* 3, 167–170. doi: 10.1515/cdbme-2017-0035

Li, Q., and Zhang, J. T. (2014). Post-trial anatomical frame alignment procedure for comparison of 3D joint angle measurement from magnetic/inertial measurement units and camera-based systems. *Physiol. Measure.* 35, 2255–2268. doi: 10.1088/0967-3334/35/11/2255

Ligorio, G., and Sabatini, A. M. (2016). Dealing with magnetic disturbances in human motion capture: a survey of techniques. *Micromachines* 7:43. doi: 10.3390/mi7030043

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., et al. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience* 6, 1–6. doi: 10.1093/gigascience/gix020

Lund, M. E., Andersen, M. S., de Zee, M., and Rasmussen, J. (2015). Scaling of musculoskeletal models from static and dynamic trials. *Int. Biomechan.* 2, 1–11. doi: 10.1080/23335432.2014.993706

Maiwald, C., Sterzing, T., Mayer, T., and Milani, T. (2009). Detecting foot-to-ground contact from kinematic data in running. *Footwear Sci.* 1, 111–118. doi: 10.1080/19424280903133938

Majumder, S., Mondal, T., and Deen, M. J. (2019). A simple, low-cost and efficient gait analyzer for wearable healthcare applications. *IEEE Sensors J.* 19, 2320–2329. doi: 10.1109/JSEN.2018.2885207

Müller, P., Begin, M. A., Schauer, T., and Seel, T. (2017). Alignment-free, self-calibrating elbow angles measurement using inertial sensors. *IEEE J. Biomed. Health Inform.* 21, 312–319. doi: 10.1109/JBHI.2016.2639537

Mundt, M., Batista, J. P., Markert, B., Bollheimer, C., and Laurentius, T. (2019a). Walking with rollator: a systematic review of gait parameters in older persons. *Eur. Rev. Aging Phys. Act.* 16, 1–9. doi: 10.1186/s11556-019-0222-5

Mundt, M., David, S., Koeppe, A., Bamer, F., Markert, B., and Potthast, W. (2019b). Intelligent prediction of kinetic parameters during cutting manoeuvres. *Med. Biol. Eng. Comput.* 57, 1833–1841. doi: 10.1007/s11517-019-02000-2

Mundt, M., Koeppe, A., Bamer, F., Potthast, W., and Markert, B. (2018a). "Prediction of joint kinetics based on joint kinematics using neural networks," in *36th Conference of the International Society of Biomechanics in Sports* (Auckland, NZ), 7–10.

Mundt, M., Koeppe, A., David, S., Bamer, F., Potthast, W., and Markert, B. (2019c). Prediction of ground reaction forces and joint moments during gait. *Med. Eng. Phys.* 17:75. doi: 10.3390/s17010075

Mundt, M., Thomsen, W., Bamer, F., and Makert, B. (2018b). Determination of gait parameters in real-world environment using low-cost inertial sensors. *PAMM* 18:e201800014. doi: 10.1002/pamm.201800014

Mundt, M., Thomsen, W., David, S., Dupré, T., Bamer, F., Potthast, W., et al. (2019d). Assessment of the measurement accuracy of inertial sensors during different tasks of daily living. *J. Biomechan.* 84, 81–86. doi: 10.1016/j.jbiomech.2018.12.023

Nowka, D., Kok, M., and Seel, T. (2019). "On motions that allow for identification of hinge joint axes from kinematic constraints and 6D IMU data," *2019 18th European Control Conference, ECC 2019* (Naples), 4325–4331.

Palermo, E., Rossi, S., Marini, F., Patanè, F., and Cappa, P. (2014). Experimental evaluation of accuracy and repeatability of a novel body-to-sensor calibration procedure for inertial sensor-based gait analysis. *Measure. J. Int. Measure. Confeder.* 52, 145–155. doi: 10.1016/j.measurement.2014.03.004

Pennock, G. R., and Clark, K. J. (1990). An anatomy-based coordinate system for the description of the kinematic displacements in the human knee. *J. Biomechan.* 23, 1209–1218.

Picerno, P. (2017). 25 years of lower limb joint kinematics by using inertial and magnetic sensors: a review of methodological approaches. *Gait Post.* 51, 239–246. doi: 10.1016/j.gaitpost.2016.11.008

Picerno, P., Cereatti, A., and Cappozzo, A. (2008). Joint kinematics estimate using wearable inertial and magnetic sensing modules. *Gait Post.* 28, 588–595. doi: 10.1016/j.gaitpost.2008.04.003

Robert-Lachaine, X., Mecheri, H., Larue, C., and Plamondon, A. (2017). Accuracy and repeatability of single-pose calibration of inertial measurement units for whole-body motion analysis. *Gait Post.* 54, 80–86. doi: 10.1016/j.gaitpost.2017.02.029

Robertson, G., Caldwell, G., Hamill, J., Kamen, G., and Whittlesey, S. (2013). *Research Methods in Biomechanics*, 2nd Edn. Champaign, IL: Human Kinetics.

Sabatini, A. M. (2006). Quaternion-based extended kalman filter for determining orientation by inertial and magnetic sensing. *IEEE Transact. Biomed. Eng.* 53, 1346–1356. doi: 10.1109/TBME.2006.875664

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., and Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience* 6, 1–9. doi: 10.1093/gigascience/gix019

Seel, T., Raisch, J., and Schauer, T. (2014). IMU-based joint angle measurement for gait analysis. *Sensors* 14, 6891–6909. doi: 10.3390/s140406891

Shahabpoor, E., and Pavic, A. (2017). Measurement of walking ground reactions in real-life environments: a systematic review of techniques and technologies. *Sensors* 17:2085. doi: 10.3390/s17092085

Solà, J. (2017). Quaternion kinematics for the error-state Kalman filter. *arXiv:1711.02508.*

Teufl, W., Miezal, M., Taetz, B., Fröhlich, M., and Bleser, G. (2018). Validity, test-retest reliability and long-term stability of magnetometer free inertial sensor based 3D joint kinematics. *Sensors* 18:1980. doi: 10.3390/s18071980

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., van Beijnum, B.-J. F., et al. (2018). Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors. *Front. Physiol.* 9:218. doi: 10.3389/fphys.2018.00218

Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., et al. (2002). ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion–part I: ankle, hip, and spine. *J. Biomechan.* 35, 543–548. doi: 10.1016/S0021-9290(01)00222-6

Young, A. D., Ling, M. J., and Arvind, D. K. (2011). "IMUSim: a simulation environment for inertial sensing algorithm design and evaluation," *Proceedings of the 10th ACMIEEE International Conference on Information Processing in Sensor Networks* (Chicago, IL), 199–210.

Zihajehzadeh, S., and Park, E. J. (2017). "A novel biomechanical model-aided IMU/UWB fusion for magnetometer-free lower body motion capture. *IEEE Trans. Syst. Man Cybernet. Syst.* 47, 927–938. doi: 10.1109/TSMC.2016.2521823

Zimmermann, T., Taetz, B., and Bleser, G. (2018). IMU-to-segment assignment and orientation alignment for the lower body using deep learning. *Sensors* 18, 1–35. doi: 10.3390/s18010302

A Novel Hybrid Deep Neural Network to Predict Pre-impact Fall for Older People Based on Wearable Inertial Sensors

Xiaoqun Yu[1†], Hai Qiu[2†] and Shuping Xiong[1]*

[1] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [2] CETHIK Group Corporation Research Institute, Hangzhou, China

Falls in the elderly is a major public health concern due to its high prevalence, serious consequences and heavy burden on the society. Many falls in older people happen within a very short time, which makes it difficult to predict a fall before it occurs and then to provide protection for the person who is falling. The primary objective of this study was to develop deep neural networks for predicting a fall during its initiation and descending but before the body impacts to the ground so that a safety mechanism can be enabled to prevent fall-related injuries. We divided the falling process into three stages (non-fall, pre-impact fall and fall) and developed deep neutral networks to perform three-class classification. Three deep learning models, convolutional neural network (CNN), long short term memory (LSTM), and a novel hybrid model integrating both convolution and long short term memory (ConvLSTM) were proposed and evaluated on a large public dataset of various falls and activities of daily living (ADL) acquired with wearable inertial sensors (accelerometer and gyroscope). Fivefold cross validation results showed that the hybrid ConvLSTM model had mean sensitivities of 93.15, 93.78, and 96.00% for non-fall, pre-impact fall and fall, respectively, which were higher than both LSTM (except the fall class) and CNN models. ConvLSTM model also showed higher specificities for all three classes (96.59, 94.49, and 98.69%) than LSTM and CNN models. In addition, latency test on a microcontroller unit showed that ConvLSTM model had a short latency of 1.06 ms, which was much lower than LSTM model (3.15 ms) and comparable with CNN model (0.77 ms). High prediction accuracy (especially for pre-impact fall) and low latency on the microboard indicated that the proposed hybrid ConvLSTM model outperformed both LSTM and CNN models. These findings suggest that our proposed novel hybrid ConvLSTM model has great potential to be embedded into wearable inertial sensor-based systems to predict pre-impact fall in real-time so that protective devices could be triggered in time to prevent fall-related injuries for older people.

Keywords: fall risk, pre-impact fall, deep neural network, machine learning, inertial sensor

INTRODUCTION

Falls are a major safety concern for the older people. Annual fall rates range from 30% among those aged over 65 years old to 50% for those over 85 (Rubenstein, 2006). Due to the high prevalence, falls are the leading cause of both fatal and non-fatal injuries among the older people (Bergen, 2016). The annual medical costs for falls of the older adults have been estimated at $31.3 billion in United States since 2015 (Burns et al., 2016). Fall-related injuries are considered as "Global Burden of Disease" by the World Health Organization (Murray et al., 2001). Aside from the physical injury, falls can also cause post-fall syndrome such as fear of falling and depression among the elderly (Fleming and Brayne, 2008; Qiu and Xiong, 2015). Therefore, effective fall prevention is critical to mitigate the negative consequences of falls for the older people.

Much work has been done on developing context-aware systems and wearable devices for post-fall detection so that timely medical assistance can be initiated for the older fallers to avoid losses caused by "long-lie" (Özdemir and Barshan, 2014; Yang et al., 2016). However, this approach is reactive since injuries from impact falls have happened already. Recently, researchers have shifted their efforts to a proactive approach-fall prevention, which is performed through fall risk assessment and intervention where the older individuals with high fall risks can be screened out earlier and then treated with appropriate interventions to reduce the risks of future falls (Choi et al., 2017; Qiu et al., 2018). However, the developed fall risk assessment tools and fall intervention programs are mainly focused on predicting and reducing the overall risk of falling in a long period (typically 1 year or more), not for the sudden falls. Many falls in the elderly happen suddenly and are difficult to prevent due to the complex multifactorial nature of falls and inevitably increased fall risks with the elderly as their physical and cognitive abilities deteriorate.

Pre-impact fall prediction can overcome aforementioned limitations of post-fall detection and overall long-term fall risk assessment and intervention. Pre-impact fall refers to a stage after the fall initiation but before the body-ground impact (Hu and Qu, 2016). Therefore, this method can predict sudden falls before the body hits against the ground (e.g., pre-impact), which make it possible to timely activate on-demand fall protection systems such as wearable airbags to prevent fall-related injuries. Because of very short period of falling (around 800 ms) and various types of falls (Sucerquia et al., 2017; Tao and Yun, 2017), to predict the fall before the ground impact accurately under different scenarios is very challenging and worthy of research investigation. Some researchers have recently attempted to tackle this challenge using different approaches (Lee et al., 2015; Sabatini et al., 2016; Li M. et al., 2018; Zhong et al., 2018; Ahn et al., 2019). In general, wearable sensors or environmental cameras were utilized and simple threshold-based algorithms were developed to predict pre-impact falls using some selected fall indicators related to human motions. Even though threshold-based algorithms are easy to implement due to simple structure and low computation cost, the thresholds are highly dependent on the certain types of falls (e.g., forward fall, backward fall) and the tested subjects,

which can not fit well for other fall types (lateral fall, vertical fall, etc.) and different older individuals in the real-world. In other words, threshold-based algorithms lack the generalizability and thus are difficult for practical applications. A few studies utilized conventional machine learning methods such as Support Vector Machine and Fisher Discriminant Analysis to predict pre-impact falls (Aziz et al., 2014; Liang et al., 2018; Wu et al., 2019). Tested by small amount of data from very limited types of simulated falls ($\leq$7), they reported good prediction accuracy and reasonable lead time. However, conventional machine learning methods heavily rely on hand-crafted features, which are usually shallow and restricted by human domain knowledge (Wang et al., 2019). Therefore, these approaches generated undermined prediction performance on complex and various falls in the real world as researchers have reported at least 15 common fall types and 19 activities of daily living (ADL; Sucerquia et al., 2017; Tao and Yun, 2017).

Very recently, with the fast advancement of deep learning and computing hardware, a few studies explored deep neural network based algorithms for pre-impact fall prediction. Li et al. (2019) applied convolutional neural network (CNN) on RGB image data recorded by Kinect for pre-impact fall prediction during gait rehabilitation training. Even though they achieved a prediction accuracy of 100% within 0.5 s after a fall initiation, they only tested the model on one type of fall and normal walking. Tao and Yun (2017) proposed a long short term memory (LSTM) model using skeleton data captured by Kinect to predict pre-impact fall. The developed model showed high sensitivity (91.7%) but relatively low specificity (75%), indicating that the model could recognize most of pre-impact falls but with high false alarm rate. Both high sensitivity and specificity are essential for the practical applications. In addition, this method is only restricted in home environment due to the limitations of stationary settings that Kinect cameras often suffer from. Torti et al. (2018) applied an overlapping sliding window segmentation technique to label falling process into three stages (non-fall, pre-impact fall or alert, and fall) and utilized a LSTM model to perform three-class classification based on the SisFall dataset (Sucerquia et al., 2017). They achieved high classification accuracy on fall (98.7%) but lower accuracy on non-fall (88.4%) and pre-impact fall (91.1%), which showed that their algorithm missed ∼9% pre-impact falls and misclassified many non-fall activities as other two classes (most of instances are labeled as non-fall activities in the SisFall dataset due to rarity of fall incidents). Furthermore, both studies only applied one deep learning model-LSTM, comparisons with other deep learning structures were not conducted.

This study aims to develop deep learning algorithms for predicting pre-impact fall in real-time so that a safety mechanism can be enabled to prevent fall induced injuries. A novel hybrid deep neural network which integrates CNN and LSTM architectures was proposed and evaluated on SisFall, a large public dataset of various falls and ADL acquired with accelerometer and gyroscope sensors. We also compared our proposed hybrid model with CNN and LSTM models in terms of model accuracy, latency and learning curve, which could provide more insights about the characteristics of different deep learning models in predicting pre-impact falls. The developed hybrid

model is expected to be embedded into wearable inertial sensor based systems, which would be promising to predict pre-impact fall in real-time so that the protective device could be triggered in time to prevent fall-induced injuries for older people.

MATERIALS AND METHODS

Dataset and Labeling

SisFall, a fall and movement dataset with various falls and ADLs acquired with wearable inertial sensors of accelerometer and gyroscope at a frequency of 200 Hz (Sucerquia et al., 2017), was selected for developing and evaluating deep learning algorithms due to two major reasons. First, it is a publicly available dataset which consists of 15 fall types, 19 ADLs and 38 subjects, including the largest amount of data in terms of number of subjects and number of activities (Musci et al., 2018) when compared with other public datasets such as MobiFall (Vavoulas et al., 2014) and UMAFall (Casilari et al., 2017). Second, the protocol is validated by a medical staff and there are 15 older subjects out of total 38 subjects in the SisFall dataset. Thus, the data pattern in SisFall

dataset should be close to the real-life ADLs and fall scenarios of the older people.

To be consistent with the earlier studies, we adopted the same criteria as Musci et al. (2018) for labeling data associated with three classes of events.

1. Non-fall: the time interval when the person is performing ADLs.
2. Pre-impact fall or alert: the time interval in which the person is transiting from a controlled state to a dangerous state which may lead to a fall.
3. Fall: the time interval when the person is experiencing a state transition that leads to a fall.

One representative diagram for three classes of events is illustrated in **Figure 1**, which shows the 3-axis acceleration data of a forward fall while walking due to a slip. The last part of data is removed for labeling because it is the state after the fall incident.

Design of Model Architecture

In this study, three models were applied to perform the classification. These models are a CNN model, a LSTM model

FIGURE 1 | Illustration of labeling three classes during a fall. The beginning period is labeled as non-fall and the blue and orange areas indicate pre-impact fall and fall, respectively; the remainder of the sequence is removed for the labeling.
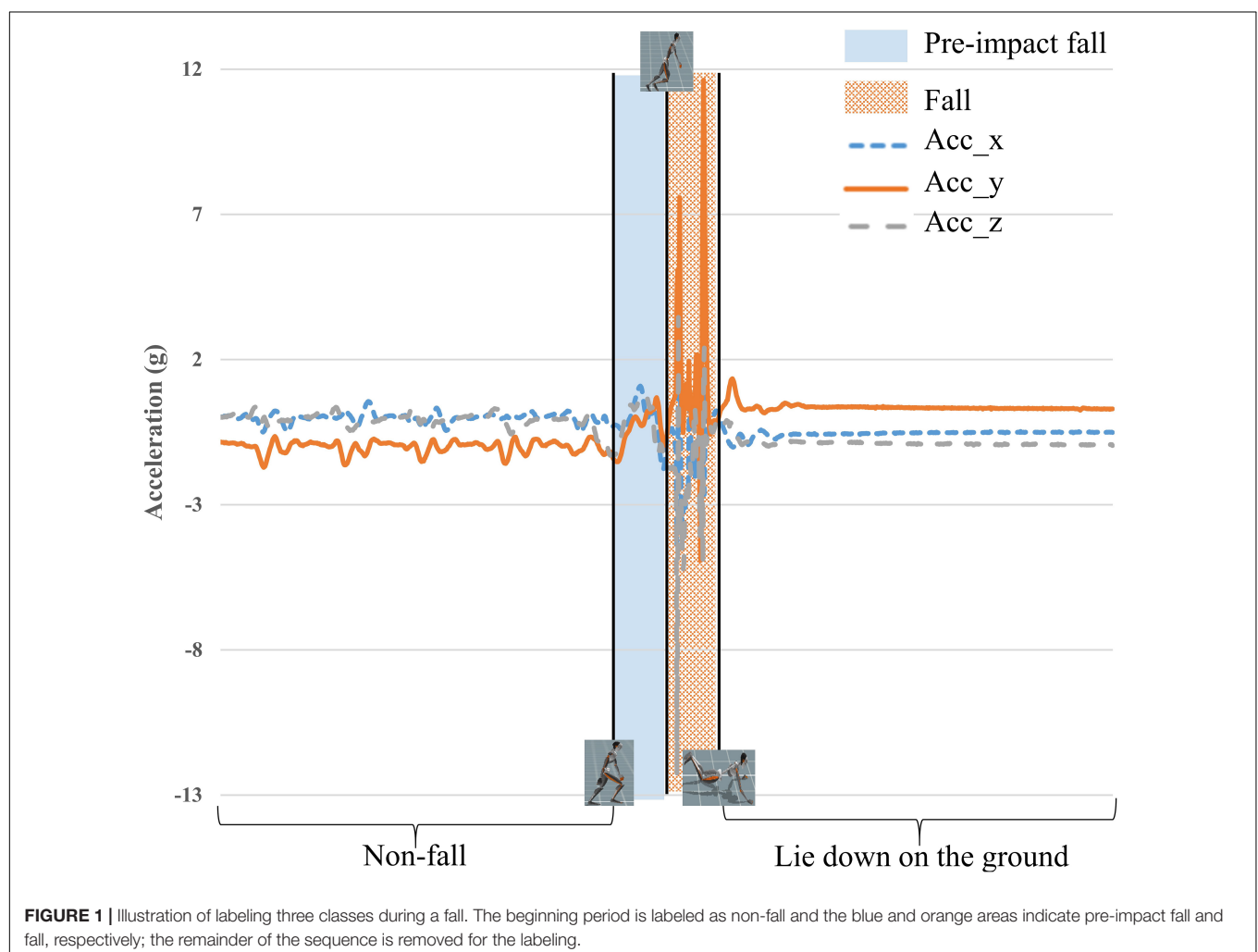
and our proposed hybrid ConvLSTM model. As shown in **Table 1**, the CNN model consists of three convolutional blocks and two fully connection layers. Each convolutional block includes convolutional operation, batch normalization, relu and max pooling. The LSTM model follows the similar design as

Musci et al. (2018), which consists of LSTM cells, relu, dropout and fully connected layers.

The architecture of our proposed ConvLSTM model mainly combines convolutional and recurrent layers. The specific structure of ConvLSTM was determined by the hyperparameter tuning. For this task, we mainly considered three levels of the width (output channels in each convolutional and LSTM layer), two different numbers of layers for both convolutional and LSTM structures, and two levels of dropout (probability of a neuron to be ignored during training). **Table 2** summarizes the results of hyperparameter tuning experiments on one training-testing split.

As shown in **Figure 2**, the finalized ConvLSTM structure after hyperparameter tuning consists of four convolutional blocks and two LSTM cells with dropout operations. Each convolutional block contains operations of convolution, batch normalization, relu, and max pooling. The convolutional layers act as feature extractors and provide abstract representations of the input sensor data in feature maps. They could capture short-term dependencies (spatial relationship) of the data. The recurrent layers deal with the long-term temporal dynamics of the activation of the feature maps and identify useful features over time domain in sequential data. More importantly, this structure could integrate advantages of CNN and LSTM on accuracy and efficiency. In the CNN, features are extracted and then used as inputs of fully connected network for classification.

TABLE 1 | The design of CNN model.

Type	Operations	Filter shape	Input size
Conv1	conv	3 × 64	256 × 6
	batchNorm		
	relu		
	max pooling	3 × 64	
Conv2	conv	3 × 64	127 × 64
	batchNorm		
	relu		
	max pooling	3 × 64	
Conv3	conv	3 × 64	62 × 64
	batchNorm		
	relu		
	max pooling	3 × 64	
FC1	fully connection	1920 × 512	1 × 1920
FC2	fully connection	512 × 3	1 × 512
Softmax	softmax	Classifier	1 × 3

TABLE 2 | Results of hyperparameter tuning for the structure of ConvLSTM model.

No.	Width	No. of Conv layers	No. of LSTM layers	Dropout	Sensitivity (%)		
					Non-fall	Pre-impact Fall	Fall
1	32	2	2	0.5	88.99	93.31	96.31
2	32	2	2	0.8	91.49	93.31	96.31
3	32	2	4	0.5	91.64	91.21	96.77
4	32	2	4	0.8	92.84	90.79	96.31
5	32	4	2	0.5	92.41	89.12	96.77
6	32	4	2	0.8	90.51	93.72	96.77
7	32	4	4	0.5	94.84	89.54	94.47
8	32	4	4	0.8	91.28	90.38	95.85
9	64	2	2	0.5	90.93	91.63	97.70
10	64	2	2	0.8	91.65	92.89	97.24
11	64	2	4	0.5	88.54	92.05	98.16
12	64	2	4	0.8	85.78	93.51	97.24
13*	64	4	2	0.5	92.30	93.30	95.86
14	64	4	2	0.8	90.18	91.63	96.77
15	64	4	4	0.5	91.47	89.94	95.85
16	64	4	4	0.8	90.22	89.54	96.31
17	128	2	2	0.5	91.73	93.31	93.55
18	128	2	2	0.8	93.77	88.28	95.85
19	128	2	4	0.5	90.58	92.89	96.31
20	128	2	4	0.8	92.10	90.79	98.16
21	128	4	2	0.5	90.45	94.98	96.31
22	128	4	2	0.8	90.75	91.63	99.08
23	128	4	4	0.5	88.85	94.56	95.85
24	128	4	4	0.8	88.53	89.96	97.24

*Finalized structure of ConvLSTM model.

FIGURE 2 | The architecture design of hybrid ConvLSTM model.
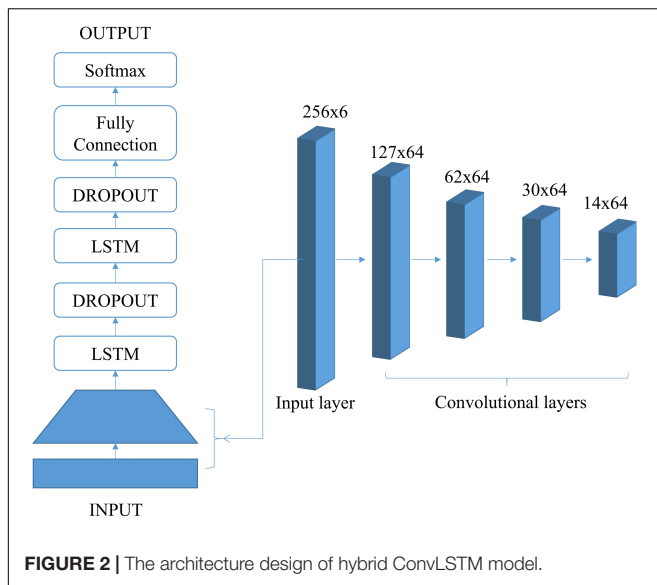
However, it ignores long-term temporal relationships in the time sequence, which is important for identifying actions or behaviors. On the contrary, the LSTM uses the memory cell to learn long-term temporal dependencies for the time-series data. However, it is time consuming for running LSTM model due to its complex structure. In the ConvLSTM, CNN layers extract features from the raw data and send to LSTM layers for identifying temporal relationships, which could save time for computing when compared with LSTM model. It is expected that ConvLSTM will outperform both CNN and LSTM models for predicting different fall stages since it can capture both short-term and long-term dependencies of the motion data.

Model Training

The architectures described in section "Design of Model Architecture" were implemented using the PyTorch library on a computer running Window 10 (64-bit). The models were trained and tested on this computer, equipped with a 3.6 GHz CPU i7-7700, 16GB RAM, and an Nvidia GTX 1080Ti GPU card. Considering the practical applications in the future, we also implemented the models on a microcontroller unit, Jetson Nano (Nvidia, 2019) which runs in Ubuntu 18.04 and equipped with a 64-bit Quad-core ARM A57 at 1.43 GHz CPU, 4GB RAM, and 128-core NVIDIA Maxwell at 921 MHz GPU. During the training, the input data has six dimensions including three-axis accelerometer and three-axis gyroscope. The batch size is 64 and the total epoch is 200. The learning rate is set as 0.0005 and the loss function uses focal loss (Lin et al., 2017).

In order to assess the generalizability of proposed models and prevent overfitting on one specific train/test split, fivefold cross validation was used. There are 23 young and 15 older subjects in SisFall dataset. In our experiment, older subjects were randomly divided into five groups and each group included three older subjects. Young subjects were also randomly divided into five groups in where three groups had five subjects and remaining two groups have four subjects. Each group of older subjects would be

randomly combined with one group of young subjects as onefold. Therefore, there were total fivefold for the dataset. Each fold would be the test set and the rest fourfold would be considered as the training set. The ratio between the training and test set was around 80%/20%. By this splitting, we could prevent the same subject appearing in both the training and test sets and maintain the homogeneity among folds at the same time.

All experiments were implemented for 200 epochs and all general hyper-parameters were set exactly same among three deep learning models for a fair comparison. In order to balance classification accuracy of three classes but without losing our focus on the pre-impact fall, the results of the epochs whose summation sensitivity for three classes are within top three and summation sensitivity is the highest for non-fall and pre-impact fall were used for averaging the fivefold cross-validation results. Because the accuracy can be biased by the majority class when the dataset is highly imbalanced, sensitivity instead of accuracy was used as the criteria to determine the best model (Bekkar et al., 2013).

Torti et al. (2018) sets baseline for our study because they also performed three-class classification (non-fall, pre-impact fall, fall) based on the SisFall dataset.

RESULTS

Classification Performance

The classification performance is represented by different metrics including sensitivity, specificity and accuracy, which are calculated by equations 1, 2, and 3, respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad (3)$$

where TP (True Positives) of non-fall is all non-fall instances that are correctly classified as non-fall class; FN (False Negatives) of non-fall is all non-fall instances that are not correctly classified as non-fall class; TN (True Negatives) of non-fall is all instances of other two classes are not classified as non-fall class; FP (False Positives) of non-fall is all instances of other two classes are wrongly classified as non-fall class. To find the four terms for other two classes, we could replace non-fall with pre-impact fall or fall.

Table 3 summarizes the classification performances of three deep learning models along with the baseline study. The results showed that CNN model had the poorest performance with the mean accuracies of 90.01, 91.51, and 98.38% for non-fall, pre-impact fall and fall, respectively. LSTM model demonstrated higher accuracies (91.59, 93.98, and 97.52%) than CNN, and our proposed hybrid ConvLSTM model achieved the highest accuracies on all classes (93.22, 94.48, and 98.66%). With respect to the sensitivity, the results showed that ConvLSTM model

TABLE 3 | Classification results of three deep learning models on the test dataset.

	Class	CNN	LSTM	ConvLSTM	Torti et al., 2018
Sensitivity (%)	Non-fall	89.90	91.50	93.15	88.39
	Pre-impact fall	90.33	91.48	93.78	91.08
	Fall	93.76	96.22	96.00	98.73
Specificity (%)	Non-fall	95.05	95.93	96.59	97.85
	Pre-impact fall	91.52	94.00	94.49	90.77
	Fall	98.42	97.54	98.69	97.93
Accuracy (%)	Non-fall	90.01	91.59	93.22	93.12
	Pre-impact fall	91.51	93.98	94.48	90.93
	Fall	98.38	97.52	98.66	98.33

had the mean sensitivities of 93.15, 93.78, and 96.00% for non-fall, pre-impact fall and fall, respectively, which were higher than CNN (89.90, 90.33, and 93.76%) and LSTM models (91.50, 91.48, and 96.22%) except the fall class. For the specificity, the ConvLSTM model had the mean specificities of 96.59, 94.49, and 98.69% for non-fall, pre-impact fall and fall, respectively, which

were higher than both LSTM (95.93, 94.00, and 97.54%) and CNN models (95.05, 91.52, and 98.42%).

Learning Curve

Figure 3 presents the representative learning curves of three deep learning models on the same training set. All three models

FIGURE 3 | Learning curves of CNN **(A)**, LSTM **(B)**, and ConvLSTM **(C)** models on the training dataset.
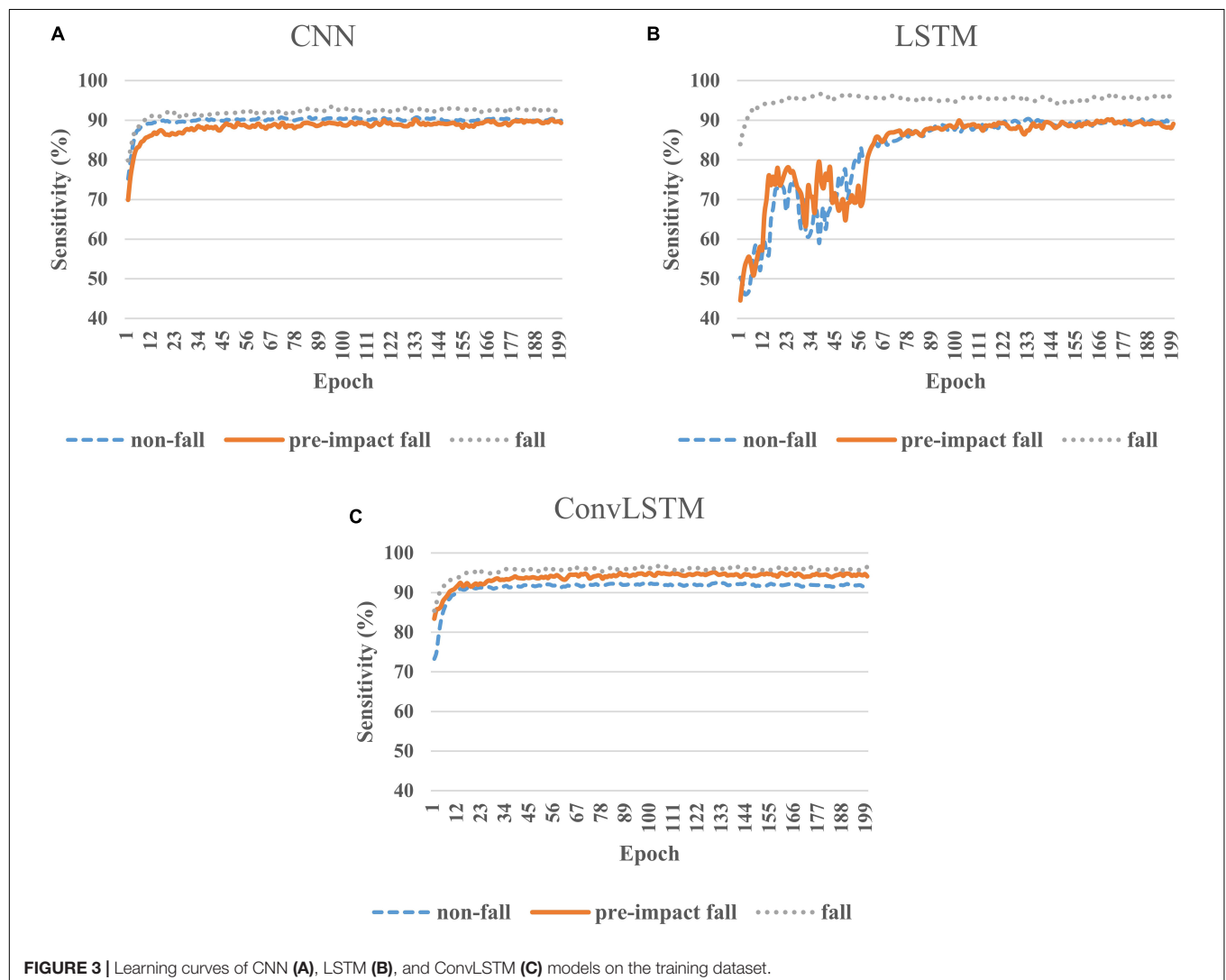
converged after certain number of epochs. Both CNN and ConvLSTM models can quickly learn and achieve the stable status (**Figures 3A,C**) while LSTM model needs more time to train (**Figure 3B**). In terms of training sensitivity, the performance of CNN was similar to LSTM on non-fall and pre-impact fall classes. The sensitivities of both models on these two classes fluctuated around 90%; while for the fall class, LSTM model was obviously better than CNN model (**Figures 3A,B**). For ConvLSTM model, the learning curves on all three classes were above 90%, especially for the pre-impact fall class (**Figure 3C**).

Figure 4 depicts the representative learning curves of three deep learning models on the same test set. CNN model failed to learn the features of pre-impact fall data well because there was a large fluctuation on sensitivity even at the end of training (**Figure 4A**). **Figure 4B** shows that LSTM model can gradually learn the features of three classes and achieved good sensitivity in the last 50 epochs. Compared to the LSTM model, ConvLSTM model can perform well after only first 20 epochs and maintain the high sensitivity for all three classes until the end of training (**Figure 4C**).

Model Latency

The latencies were evaluated with the same training and test sets among three deep learning models. For the practical applications, only processing time on each instance in the test set was summed and averaged over 200 epochs. Models tested on the computer showed average latencies of 0.61, 0.70, and 0.97 ms for CNN, ConvLSTM, and LSTM models, respectively. Further tests on a microcontroller unit (Nvidia Jetson Nano) showed the averaged latency of ConvLSTM model was 1.06 ms, which was slightly higher or comparable with CNN model (0.77 ms) but much lower than LSTM model (3.15 ms).

DISCUSSION

We developed a hybrid deep learning model (ConvLSTM) that integrates the CNN and LSTM architectures to predict the pre-impact fall from accelerometer and gyroscope sensor data. The performance of this hybrid model was comprehensively compared with CNN and LSTM deep learning models. The experimental results showed that the hybrid ConvLSTM model outperformed CNN and LSTM models in terms of sensitivity, specificity and overall accuracy. The hybrid ConvLSTM model obtained ∼2% higher sensitivities than LSTM and ∼3% higher sensitivities than CNN for all three classes except the fall class. Considering our study aimed to predict the pre-impact fall accurately for preventing fall induced injuries, the high sensitivities for non-fall and pre-impact fall were of significant importance in two perspectives. On the one hand, higher classification sensitivity on non-fall class reflected lower false alarm rate and 2% improvement was very meaningful because dominant instances in the SisFall dataset and real-world scenarios are non-falls or ADLs, and fall instances are very rare. On the other hand, higher classification sensitivity for the pre-impact fall directly indicated the superiority of the ConvLSTM model. In addition, the ConvLSTM model obtained the highest specificities

for non-fall (96.59%), pre-impact fall (94.49%), and fall (98.69%) among three deep learning models. A more detailed investigation showed that although the difference on the specificity between ConvLSTM model and LSTM model was marginal, both models had ∼3 and 2.5% higher specificities on pre-impact fall prediction than CNN model. This result indicated that CNN model had the highest rate of misclassifying other two classes as pre-impact fall.

It is understandable that the hybrid ConvLSTM model outperformed individual CNN or LSTM models. CNN could capture local dependency of human motion data (Zeng et al., 2014). For the given time point, the neighboring accelerometer and gyroscope readings are likely to be correlated. However, this dependency is short-term due to the constraint by the size of convolutional kernels (Li F. et al., 2018). On the contrary, LSTM with memory cells could learn to store and output information based on the training, easing the learning of long-term time dependency of motion data (Hochreiter and Schmidhuber, 1997). Therefore, integration of both short-term and long-term dependencies could enhance the ability to distinguish different fall stages that vary in time span and signal distribution.

Our experimental results indicated that the motion features in the long term were more significant in classifying three fall stages (non-fall, pre-impact fall, fall) than those in the short term. This finding was consistent with those of earlier studies using deep learning approaches for human motion recognition (Yao et al., 2017; Li F. et al., 2018). Long-term motion features were also widely used in the conventional machine learning methods for human movement analysis. For example, Su et al. (2016) achieved high accuracy to distinguish falls from non-falls by extracting twelve time-domain features from angular velocity and angle data into a hierarchical classifier. Similarly, Panahandeh et al. (2013) suggested that long-term features of sensor data such as magnitude-squared discrete Fourier transform coefficient and variance were critical for pedestrian activity classification and gait analysis. Furthermore, researchers reported the classification with an integration of time domain (mean, variance, kurtosis, etc.) and gait temporal features (stride time, stance time, double-limb support, etc.) showed better results to differentiate stroke and other neurological disorders than using them separately (Hsu et al., 2018). Compared with conventional machine learning methods, our proposed deep neural networks can eliminate the need of manually designed motion features and can fully utilize the useful information in the raw data for classification.

Two earlier studies utilized CNN and LSTM alone to predict pre-impact falls (Tao and Yun, 2017; Li et al., 2019). Both studies divided the motion data into non-fall and pre-impact fall, and pre-impact fall included several frames before and after the fall initiation so that they could predict the pre-impact fall. However, the data of remaining part of falling (fall class in the three classifications) was not considered, thus these kinds of simple binary classification models can not be used for predicting the fall class. In addition, both studies only tested classification models on a small dataset with limited types of falls (≤4) and ADLs (≤4). To the best of our knowledge, there was only one published study utilizing the LSTM-based three-class classification model to predict the pre-impact fall based on a large dataset-SisFall (Torti et al., 2018). To compare our proposed deep learning
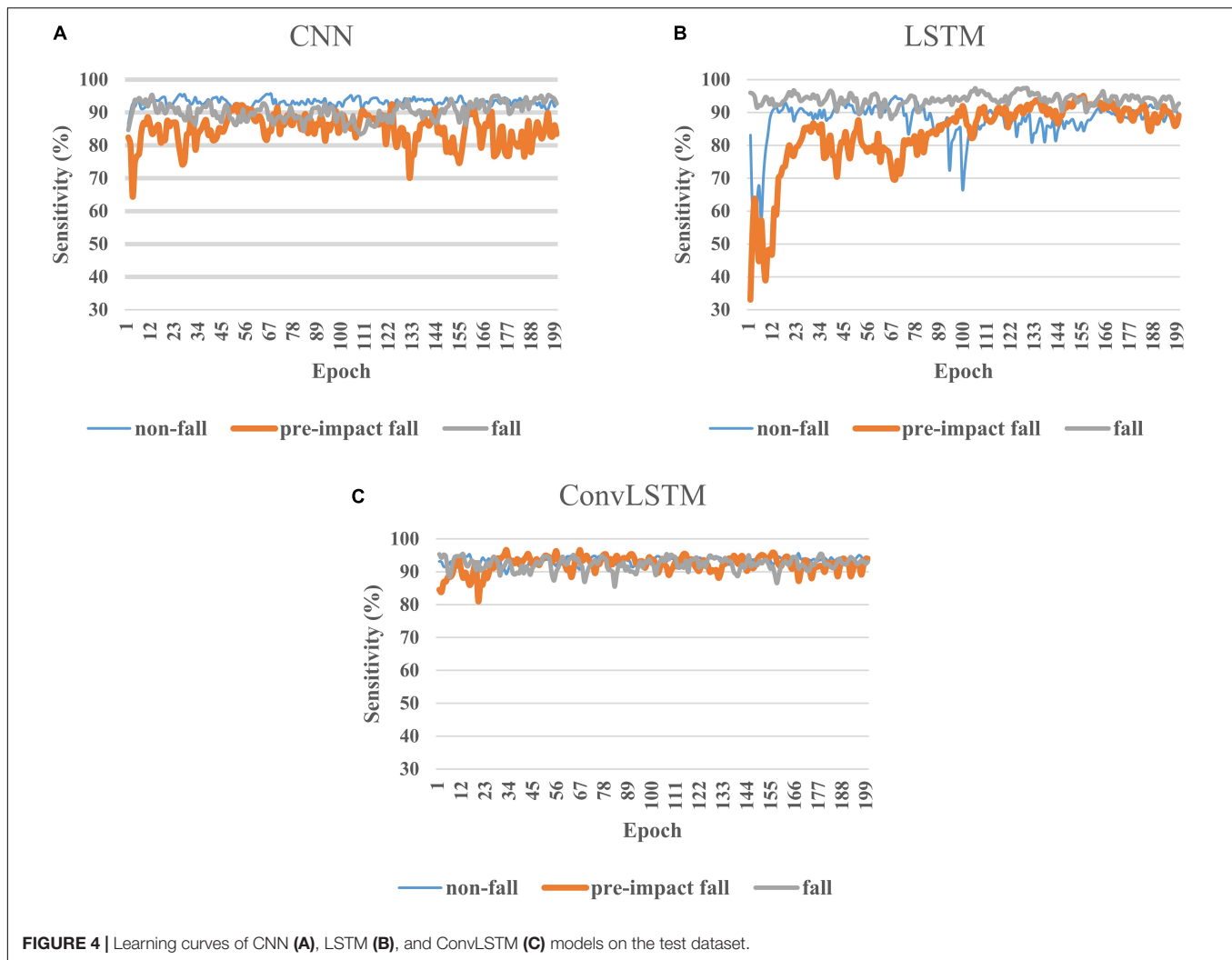
FIGURE 4 | Learning curves of CNN **(A)**, LSTM **(B)**, and ConvLSTM **(C)** models on the test dataset.

models with this baseline study, we also used the SisFall dataset and strictly followed the same criteria for labeling three different classes. Comparison to the benchmark (**Table 3**) showed that our hybrid ConvLSTM model achieved higher sensitivity of ∼5 and 3% for the non-fall and pre-impact fall, respectively, indicating considerably lower false alarm rate but higher true alarm rate for the pre-impact fall prediction. However, our ConvLSTM model had ∼3% lower sensitivity than the benchmark on predicting the fall class. This could be caused by the different strategy we used to choose the best model. We prioritized the high sensitivity on classes of non-fall and pre-impact fall because the primary objective of this study was to predict a fall with a reasonable lead time before the body impacts to the ground rather than detect a fall after it happens. For the specificity, even though there was no considerable difference on classes of non-fall and fall, our ConvLSTM model outperformed the benchmark on the class of pre-impact fall (higher specificity by 3.7%), which demonstrated lower misclassification rate on pre-impact fall prediction.

In terms of latency, LSTM model is time consuming due to its complex structure and difficulty in parallel computing. In

the proposed hybrid ConvLSTM model, the first CNN layers which are appropriate for parallel computation would extract features hierarchically from the raw motion sensor data. The extracted features would be inputted to following LSTM layers for identifying temporal dependencies. Compared with the raw data as the input in LSTM model, these features are in a much lower dimensional space and thus far more concise. Therefore, inserting CNN layers ahead of LSTM layers could save significant amount of time for computation. Interestingly, even tested on a microcontroller unit of the Jetson Nano with the exact same model tested on the computer, the latency of our proposed hybrid model still maintained very short and within 1.1 ms, demonstrating a great potential to implement our developed hybrid model into predicting the pre-impact falls in real-time so that the on-demand fall protection systems (e.g., wearable airbags) can be timely activated to prevent fall-related injuries.

The present study has several limitations worth noting. First, because the SisFall dataset did not provide the video references about the simulated falls and ADLs of each subject, the pre-impact fall and fall intervals of the sensor signal labeled by

authors of the baseline study may not be very consistent. Considering Xsens wearable motion capture system could record motion data and reconstruct graphical videos of human motions synchronously, we will use it to build a new fall dataset and further verify the developed deep learning algorithms. Second, for some types of falls such as a lateral fall, the duration of falling is very short and the time interval of pre-impact fall is too short to specify. Therefore, for these fall cases, pre-impact fall instances may not be labeled reliably due to the much larger width of sliding window. Further analysis on different window sizes could be conducted. Third, the development of the CovnLSTM model was based on the SisFall dataset with simulated falls performed by limited subjects. Caution is thus needed in directly applying this model into practice. Large-scale fall simulations and real-life tests with good protection need to be conducted further. Last but not least, non-fall instances are very dominant in the SisFall dataset compared with instances for other two classes, which induces challenges in training the deep learning models. More scientific techniques such as data argumentation to cope with highly imbalanced data should be explored further.

CONCLUSION

We proposed a hybrid deep learning model (ConvLSTM) which integrates the CNN and LSTM architectures to predict the pre-impact fall for older people based on accelerometer and gyroscope data. The performance of this hybrid model was evaluated on SisFall, a large public dataset of various falls and ADL. We also comprehensively compared the proposed hybrid ConvLSTM model with CNN and LSTM deep learning models in terms of model accuracy, latency and learning curve. Experimental results showed that the hybrid ConvLSTM model obtained both high sensitivities ($>$93%) and specificities ($>$94%) for all three fall stages (non-fall, pre-impact fall and fall), which were higher than LSTM and CNN models. In addition, latency test on a microcontroller unit (Nvidia Jetson Nano) showed that ConvLSTM model had a short latency of 1.06 ms, which was much lower than LSTM model (3.15 ms) and comparable with

CNN model (0.77 ms). High prediction accuracy (especially for pre-impact fall) and low latency on the micro board indicated that the proposed hybrid ConvLSTM model outperformed both LSTM and CNN models. These findings suggested that our proposed novel hybrid ConvLSTM model has great potential to be embedded into wearable inertial sensor-based systems to predict pre-impact fall in real-time so that protective devices could be triggered in time to prevent fall-related injuries for older people.

DATA AVAILABILITY STATEMENT

The pre-processed datasets used in this study for fivefold cross validation are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

SX conceptualized the study, obtained the funding, and reviewed and edited the manuscript. HQ and SX designed the neural networks. XY did the data pre-processing and implemented the training and testing of deep neural networks. XY and HQ wrote the original draft.

FUNDING

ACKNOWLEDGMENTS

We thank the developers of SisFall for publicly sharing the dataset. We also thank the authors of the baseline study (Torti et al., 2018) for sharing the temporal labeling of the SisFall dataset.

REFERENCES

Ahn, S., Kim, J., Koo, B., and Kim, Y. (2019). Evaluation of inertial sensor-based pre-impact fall detection algorithms using public dataset. *Sensors* 19:774. doi: 10.3390/s19040774

Aziz, O., Russell, C. M., Park, E. J., and Robinovitch, S. N. (2014). "The effect of window size and lead time on pre-impact fall detection accuracy using support vector machine analysis of waist mounted inertial sensor data," in *Proceedings of the Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference*, (Piscataway, NJ: IEEE), 30–33.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J. Inform. Eng. Appl.* 3, 27–38.

Bergen, G. (2016). Falls and fall injuries among adults aged $\geq$ 65 years—United States, 2014. *MMWR* 65, 993–998. doi: 10.15585/mmwr.mm 6537a2

Burns, E. R., Stevens, J. A., and Lee, R. (2016). The direct costs of fatal and non-fatal falls among older adults—United States. *J. Saf. Res.* 58, 99–103. doi: 10.1016/j.jsr.2016.05.001

Casilari, E., Santoyo-Ramón, J. A., and Cano-García, J. M. (2017). UMAFall: a multisensor dataset for the research on automatic fall detection. *Proc. Comput. Sci.* 110, 32–39. doi: 10.1016/j.procs.2017.06.110

Choi, S. D., Guo, L., Kang, D., and Xiong, S. (2017). Exergame technology and interactive interventions for elderly fall prevention: a systematic literature review. *Appl. Ergon.* 65, 570–581. doi: 10.1016/j.apergo.2016.10.013

Fleming, J., and Brayne, C. (2008). Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. *BMJ* 337:a2227. doi: 10.1136/bmj.a2227

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Hsu, W.-C., Sugiarto, T., Lin, Y.-J., Yang, F.-C., Lin, Z.-Y., Sun, C.-T., et al. (2018). Multiple-wearable-sensor-based gait classification and analysis in patients with neurological disorders. *Sensors* 18:3397. doi: 10.3390/s18103397

Hu, X., and Qu, X. (2016). Pre-impact fall detection. *Biomed. Eng. Online* 15:61. doi: 10.1186/s12938-016-0194-x

Lee, J. K., Robinovitch, S. N., and Park, E. J. (2015). Inertial sensing-based pre-impact detection of falls involving near-fall scenarios. *IEEE Trans. Neural Syst. Rehabil. Eng.* 23, 258–266. doi: 10.1109/TNSRE.2014.2357806

Li, F., Shirahama, K., Nisar, M., Köping, L., and Grzegorzek, M. (2018). Comparison of feature learning methods for human activity recognition using wearable sensors. *Sensors* 18:E679.

Li, M., Xu, G., He, B., Ma, X., and Xie, J. (2018). Pre-impact fall detection based on a modified zero moment point criterion using data from kinect sensors. *IEEE Sens. J.* 18, 5522–5531. doi: 10.1109/jsen.2018.2833451

Li, S., Xiong, H., and Diao, X. (2019). "Pre-impact fall detection using 3D convolutional neural network," in *Proceedings of the 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, (Piscataway, NJ: IEEE), 1173–1178.

Liang, S., Chu, T., Lin, D., Ning, Y., Li, H., and Zhao, G. (2018). "Pre-impact alarm system for fall detection using MEMS sensors and HMM-based SVM classifier," in *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Piscataway, NJ: IEEE), 4401–4405.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, (Piscataway, NJ: IEEE), 2980–2988.

Murray, C. J., Lopez, A. D., Mathers, C. D., and Stein, C. (2001). *The Global Burden of Disease 2000 Project: Aims, Methods and Data Sources*. Piscataway, NJ: CiteseerX.

Musci, M., De Martini, D., Blago, N., Facchinetti, T., and Piastra, M. (2018). Online fall detection using recurrent neural networks. *arXiv*, arXiv:1804.04976. Available at: https://arxiv.org/abs/1804.04976 (accessed March 28, 2019).

Nvidia (2019). *Jetson Nano Developer Kit [Online]*. Available at: https://developer.nvidia.com/embedded/jetson-nano-developer-kit (accessed September 10, 2019).

Özdemir, A. T., and Barshan, B. (2014). Detecting falls with wearable sensors using machine learning techniques. *Sensors* 14, 10691–10708. doi: 10.3390/s140610691

Panahandeh, G., Mohammadiha, N., Leijon, A., and Händel, P. (2013). Continuous hidden Markov model for pedestrian activity classification and gait analysis. *IEEE Trans. Instrum. Meas.* 62, 1073–1083. doi: 10.1109/tim.2012.2236792

Qiu, H., Rehman, R. Z. U., Yu, X., and Xiong, S. (2018). Application of Wearable inertial sensors and a new test battery for distinguishing retrospective fallers from non-fallers among community-dwelling older people. *Sci. Rep.* 8:16349. doi: 10.1038/s41598-018-34671-6

Qiu, H., and Xiong, S. (2015). Center-of-pressure based postural sway measures: reliability and ability to distinguish between age, fear of falling and fall history. *Int. J. Ind. Ergon.* 47, 37–44. doi: 10.1016/j.ergon.2015.02.004

Rubenstein, L. Z. (2006). Falls in older people: epidemiology, risk factors and strategies for prevention. *Age Ageing* 35, ii37–ii41. doi: 10.1093/ageing/afl084

Sabatini, A. M., Ligorio, G., Mannini, A., Genovese, V., and Pinna, L. (2016). Prior-to- and post-impact fall detection using inertial and barometric altimeter measurements. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 774–783. doi: 10.1109/TNSRE.2015.2460373

Su, Y., Liu, D., and Wu, Y. (2016). "A multi-sensor based pre-impact fall detection system with a hierarchical classifier," in *Proceedings of the Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress*, (Piscataway, NJ: IEEE), 1727–1731.

Sucerquia, A., López, J. D., and Vargas-Bonilla, J. F. (2017). SisFall: a fall and movement dataset. *Sensors* 17:198. doi: 10.3390/s17010198

Tao, X., and Yun, Z. (2017). Fall prediction based on biomechanics equilibrium using kinect. *Int. J. Distrib. Sens. Netw.* 13, 1–9.

Torti, E., Fontanella, A., Musci, M., Blago, N., Pau, D., Leporati, F., et al. (2018). "Embedded real-time fall detection with deep learning on wearable devices," in *Proceedings of the 2018 21st Euromicro Conference on Digital System Design (DSD)*, (Piscataway, NJ: IEEE), 405–412.

Vavoulas, G., Pediaditis, M., Chatzaki, C., Spanakis, E. G., and Tsiknakis, M. (2014). The mobifall dataset: fall detection and classification with a smartphone. *Int. J. Monit. Surveill. Technol. Res.* 2, 44–56. doi: 10.4018/ijmstr.2014010103

Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. (2019). Deep learning for sensor-based activity recognition: a survey. *Pattern Recogn. Lett.* 119, 3–11. doi: 10.1016/j.patrec.2018.02.010

Wu, Y., Su, Y., Feng, R., Yu, N., and Zang, X. (2019). Wearable-sensor-based pre-impact fall detection system with a hierarchical classifier. *Measurement* 140, 283–292. doi: 10.1016/j.measurement.2019.04.002

Yang, L., Ren, Y., and Zhang, W. (2016). 3D depth image analysis for indoor fall detection of elderly people. *Dig. Commun. Netw.* 2, 24–34. doi: 10.1016/j.dcan.2015.12.001

Yao, S., Hu, S., Zhao, Y., Zhang, A., and Abdelzaher, T. (2017). "Deepsense: a unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web: International World Wide Web Conferences Steering Committee*, (Piscataway, NJ: IEEE), 351–360.

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., et al. (2014). "Convolutional neural networks for human activity recognition using mobile sensors," in *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*, (Piscataway, NJ: IEEE), 197–205.

Zhong, Z., Chen, F., Zhai, Q., Fu, Z., Ferreira, J. P., Liu, Y., et al. (2018). "A real-time pre-impact fall detection and protection system," in *Proceedings of the 2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, (Piscataway, NJ: IEEE), 1039–1044.

New Considerations for Collecting Biomechanical Data Using Wearable Sensors: The Effect of Different Running Environments

Lauren C. Benson[1], Christian A. Clermont[1] and Reed Ferber[1,2,3]*

[1] *Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada,* [2] *Running Injury Clinic, Calgary, AB, Canada,* [3] *Faculty of Nursing, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada*

Traditionally, running biomechanics analyses have been conducted using 3D motion capture during treadmill or indoor overground running. However, most runners complete their runs outdoors. Since changes in running terrain have been shown to influence running gait mechanics, the purpose of this study was to use a machine learning approach to objectively determine relevant accelerometer-based features to discriminate between running patterns in different environments and determine the generalizability of observed differences in running patterns. Center of mass accelerations were recorded for recreational runners in treadmill-only ($n = 28$) and sidewalk-only ($n = 25$) environments, and an independent group ($n = 16$) ran in both treadmill and sidewalk environments. A feature selection algorithm was used to develop a training dataset from treadmill-only and sidewalk-only running. A binary support vector machine model was trained to classify treadmill and sidewalk running. Classification accuracy was determined using 10-fold cross-validation of the training dataset and an independent testing dataset from the runners that ran in both environments. Nine features related to the consistency and variability of center of mass accelerations were selected. Specifically, there was greater ratio of vertical acceleration during treadmill running and a greater ratio of anterior-posterior acceleration during sidewalk running in both the training and testing dataset. Step and stride regularity were significantly greater in the treadmill condition for the vertical axis in both the training and testing dataset, and in the medial-lateral axis for the testing dataset. During sidewalk running, there was significantly greater variability in the magnitude of the vertical and anterior-posterior accelerations for both datasets. The classification accuracy based on 10-fold cross-validation of the training dataset ($M = 93.17\%$, SD = 2.43%) was greater than the classification accuracy of the independent testing dataset ($M = 83.81\%$, SD = 3.39%). This approach could be utilized in future analyses to identify relevant differences in running patterns using wearable technology.

Keywords: running, machine learning, classification, treadmill, outdoor

INTRODUCTION

Traditional running biomechanical analysis is confined to treadmill or over-ground indoor running (Simon, 2004). In contrast, most runners complete their runs outdoors (Taunton et al., 2003) and research has shown that machine learning algorithms trained on gait-related features from an accelerometer can distinguish whether a runner is running on concrete, synthetic, woodchip surfaces (Schütte et al., 2016). However, to our knowledge, no study has examined differences in running biomechanics between indoor running, where the speed, surface inclination and available space are often dictated by a treadmill or a small flat runway, and outdoor running, where these features are less controlled. Insights gleaned from biomechanical analyses conducted in less controlled settings may be more applicable to runners who train and compete outdoors.

Limited research has been conducted to compare treadmill to overground running, but has shown that the running biomechanical patterns during treadmill running gait dynamics do not necessarily reflect overground running patterns (Lindsay et al., 2014; Schütte et al., 2016). Moreover, methodological limitations make it difficult to generalize these results. For example, Lindsay et al. (2014) compared treadmill running to overground running on an indoor track and Schütte et al. (2016) performed outdoor investigations on a short, flat and straight course, limiting the ability to generalize the findings to runners outside of the study sample and real-world conditions. Dixon et al. (2019) collected only 8 s of data, from between 2 and 4 running trials, whilst runners ran on a straight 90 m segment of either concrete road, synthetic track, or woodchip trail. Indoor tracks and short, straight and flat runways do not necessarily reflect real-world running conditions, particularly for long-distance runners. Thus, research is needed in order to collect running biomechanical data in a runner's natural environment. Considering that the vast majority of running biomechanical data collected to date have been in controlled laboratory settings, it will be beneficial to understand which biomechanical variables are similar, or dissimilar, to those exhibited during running in real-world environments.

Inertial measurement units (IMUs) are portable devices that can be used to quantify running biomechanical patterns in a runner's natural environment (Norris et al., 2014; Reenalda et al., 2016), yet, these investigations are still rare (Benson et al., 2018a). Running biomechanical analysis using IMUs is commonly conducted by recording 3D center of mass accelerations and extracting features related to the magnitude, consistency and variability of the signal (Henriksen et al., 2004; Moe-Nilssen and Helbostad, 2004; Kobsar et al., 2014; Benson et al., 2018b; Clermont et al., 2018). There remains an absence of an association between joint-level mechanics commonly investigated using laboratory-based motion capture systems and features generated from center of mass accelerations. Thus, a challenge in identifying new methods for collecting biomechanical data using wearable sensors is to identify which accelerometer-based features are relevant for observing running patterns in real world settings.

The purpose of this study was to determine whether running environments could be successfully classified from movement patterns quantified by the use of a single accelerometer, with generalizability to an independent dataset. A secondary objective was to determine which features drive successful classification between treadmill-only and sidewalk-only running. It was expected that key features would quantify the consistency and variability of running patterns, and that the model would be generalizable to an independent set of runners.

MATERIALS AND METHODS

Participants and Equipment

A total of 69 self-identified recreational runners provided informed consent to participate in this study approved by the Ethics Board at the University of Calgary (REB16-1183). Both male and female runners with no running-related injury in the previous 6 months were included. All participants were outfitted with an IMU (Shimmer3 GSR+® ±8 g, Shimmer Inc., Dublin, IE, United States) on the lower back near the center of mass, such that the positive x-axis pointed to the right, the positive y-axis pointed vertically, and the positive z-axis pointed posteriorly. Three-dimensional accelerations were recorded at 201.03 Hz and stored on an SD card. Additionally, a GPS-capable watch (Garmin vivoactive HR, Garmin Inc., Olathe, KS, United States) with a sampling rate of 1 Hz was worn on the preferred wrist. Participants wore their preferred clothes and shoes.

Data Collection

Each participant was included in just one of three protocols, based on weather (i.e., outdoor running only occurred on days with no snow or rain) and availability to attend multiple sessions (**Table 1**). In Protocol 1, 28 participants ran on a level treadmill (Bertec, Columbus, OH, United States) only. The speed was initially set to a speed equal to what the participant self-reported as their typical training pace, and was subsequently adjusted in 0.1 m/s increments until it matched the participant's preferred speed, described as "a pace which you would be comfortable to run for about 45 min and represents a usual, common, or typical pace (Lindsay et al., 2014)." Participants first completed a 5–10 min warmup at this speed. Next, data were recorded as the participants ran at their preferred speed (recorded as the treadmill setting) for 5 min. In Protocol 2, 25 different participants ran outdoors on a concrete sidewalk only. First, participants completed a 5–10 min warmup at their own pace. Then, data were recorded as the participants ran at their preferred running speed (recorded with GPS watch) on a continuous stretch of sidewalk that featured a straightaway, curve and slight incline typical of real-world outdoor running conditions. The sidewalk was 300 m, and the participants paused for 10 s at the turnaround to complete a total of 600 m (**Figure 1**). It was expected that all runners would complete the 600 m course within 5 min (8:20/km pace). In Protocol 3, a different set of 16 participants completed both the treadmill and sidewalk runs on separate days, with the order of days randomized, via a coin flip, for each participant.

Environments	Protocol 1 TM (5 min) only	Protocol 2 S (600 m) only	Protocol 3 TM (5 min) and S (600 m)
Sex	18M, 10F	12M, 13F	8M, 8F
Height, m	1.74 (0.09)	1.73 (0.10)	1.70 (0.09)
Mass, kg	70.5 (10.3)	70.2 (13.0)	67.1 (8.1)
Age, yr	32.2 (13.4)	36.9 (10.1)	31.3 (10.2)
TM speed, m/s	2.78 (0.26)	–	2.75 (0.39)*
S speed, m/s	–	3.24 (0.42)	3.10 (0.60)*

TM, treadmill; S, sidewalk. * Within Protocol 3, TM speed was significantly lower than S Speed, p = 0.001. There were no significant differences between protocols for any variables.
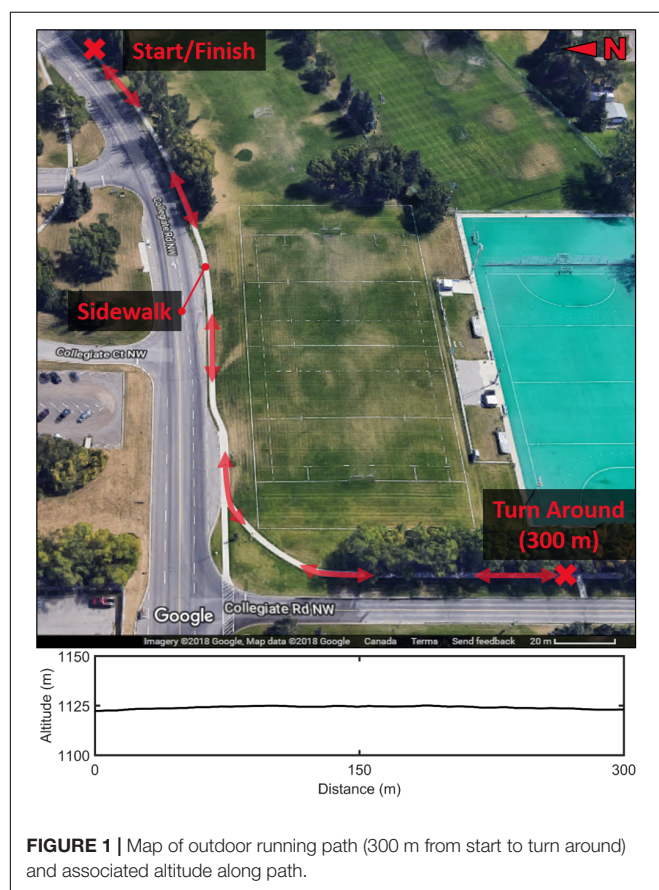
FIGURE 1 | Map of outdoor running path (300 m from start to turn around) and associated altitude along path.

Data Processing

For each run, the accelerometer data were filtered using a 4th-order low-pass Butterworth filter with a cutoff frequency at 10 Hz (Wundersitz et al., 2015), and the first and last 5% of the signal was removed to eliminate effects of starting and stopping. The trimming was applied to each 300 m section of the sidewalk runs, as a complete turnaround is likely not generalizable to real-world running conditions. The acceleration signal was then aligned with gravity (Moe-Nilssen, 1998) and the direction of motion within the horizontal plane (Avvenuti et al., 2013). The signal was segmented into steps (Lee et al., 2010), each step was normalized to 50 data points, and a previously defined

set of 24 features (Moe-Nilssen and Helbostad, 2004; Kobsar et al., 2014; Barden et al., 2016) was extracted from the signal (**Table 2**). These features included the peaks, magnitude (RMS), and ratio of the acceleration in three dimensions, averaged across all steps. Several features related to consistency and variability of the running pattern across all steps and strides. Regularity is the consistency of the stride-to-stride or step-to-step pattern, while symmetry is the difference between step and stride regularity (Barden et al., 2016), and higher values indicate a more consistent gait pattern. Mean running speed was included as a 25th feature for each participant.

Feature Selection

To improve generalizability of classification and to reduce model complexity, a subject-specific forward-sequential feature selection algorithm with a linear discriminant analysis wrapper and 10-fold cross-validation (Chizi and Maimon, 2010; Caby et al., 2011) was applied to the data from Protocols 1 and 2 to identify relevant features, ranked based on their order of selection, for the classification of running environments (**Figure 2**). Only the features selected in at least 10% of

Feature	Independent of axes	AP	ML	VT
Speed*	✓			
Step time CV	✓			
Stride time CV	✓			
RMS tesultant	✓			
Regularity step		✓	✓	✓
Regularity stride		✓	✓	✓
Symmetry (regularity step/regularity stride)		✓	✓	✓
Peak		✓	✓	✓
RMS		✓	✓	✓
RMS CV		✓	✓	✓
Ratio (RMS/RMS resultant)		✓	✓	✓

AP, anterior-posterior axis; ML, medial-lateral axis; VT, vertical axis; CV, coefficient of variation; RMS, root mean squared. *Speed was determined from the GPS watch or treadmill setting, not the accelerometer signal.
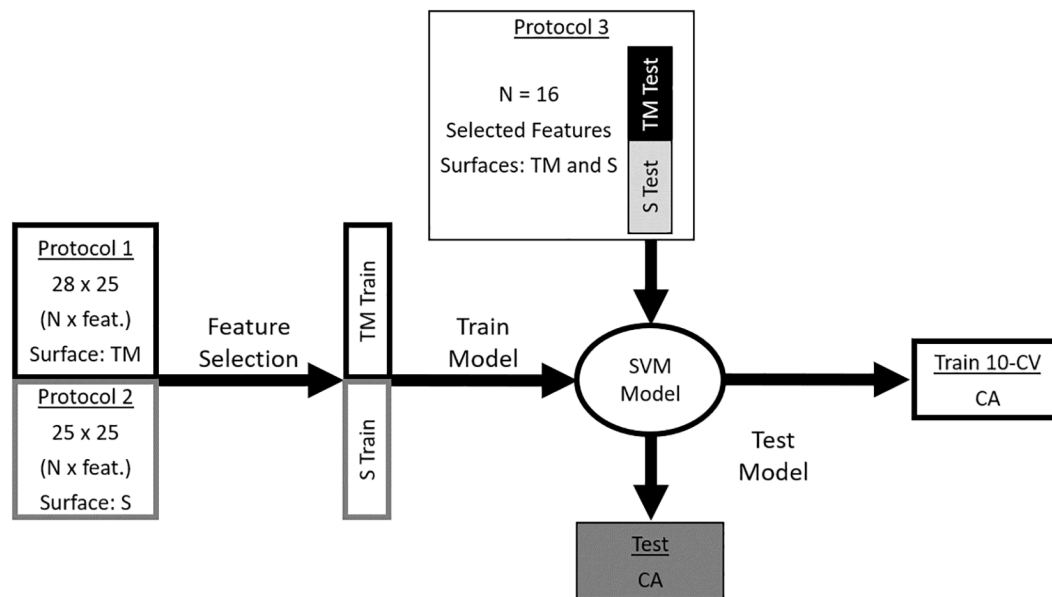
FIGURE 2 | The data from Protocol 1 and Protocol 2 were used to create a model to distinguish treadmill running from sidewalk running. Prior to building the model, the number of features in the training dataset was reduced following a feature selection task. The two environments from Protocol 3 were used as an independent testing dataset for the model. The features in the testing dataset matched the selected features in the training dataset. TM, treadmill; S, sidewalk; SVM, support vector machine; CA, classification accuracy; 10-CV, 10-fold cross-validation of the training dataset.

100 iterations were retained, and the selected features in Protocols 1 and 2 became the training dataset. All data processing and feature selection was done using custom MATLAB software (v9.1.0.441655, Mathworks, Inc., Natick, MA, United States).

Classification

The training dataset was used to train a binary support vector machine classifier (Shmilovici, 2010) for treadmill and sidewalk, with all hyper-parameters optimized with the MATLAB function *fitcsvm*. The model was tested two ways: (1) 10-fold cross-validation of the training dataset from Protocol 1 and 2, with each participant's data in only one fold at a time, and (2) the selected features from both runs in Protocol 3 were used as an independent testing dataset. The classification process was repeated for 100 iterations, and an average classification accuracy across all iterations was determined.

Statistical Analysis

Height, mass, age, and treadmill or sidewalk speed were checked for normality and compared across protocols in separate ANOVAs. A paired *t*-test was used to detect differences in speed between treadmill and sidewalk among participants within Protocol 3. Differences between treadmill and sidewalk for each of the selected features were determined with independent *t*-tests for the training dataset and paired *t*-tests for the testing dataset. For each statistical test, significance was determined at $p < 0.05$, with a Bonferroni adjustment based on number of comparisons. All statistical analyses were done using SPSS (v24.0.0.1, SPSS, Inc., Chicago, IL, United States).

RESULTS

There was no significant effect of protocol for height, mass, age, or treadmill or sidewalk speed ($p > 0.05$). Within Protocol 3, speed was significantly different ($p = 0.001$) between treadmill running ($M = 2.75$ m/s, SD = 0.39 m/s) and sidewalk running ($M = 3.10$ m/s, SD = 0.60 m/s).

Nine features were selected to discriminate treadmill and sidewalk running (**Table 3** and **Figure 3**). There was a greater ratio of vertical acceleration during treadmill running and a greater ratio of anterior-posterior acceleration during sidewalk running in both the training and testing dataset. Step and stride

TABLE 3 | Selected features used in the classification model.

Mean rank	Selected features
1.00	Ratio VT
1.05	Ratio AP
2.06	Regularity step ML
2.06	RMS CV ML
2.30	Regularity stride VT
2.39	RMS CV AP
2.67	RMS CV VT
2.86	Regularity stride ML
3.00	Regularity step VT

AP, anterior-posterior; ML, medial-lateral; VT, vertical; CV, coefficient of variation; RMS, root mean squared. Features were ranked according to the order in which they were selected during the 10-fold cross-validation of the feature selection algorithm, and the mean rank over 100 iterations of feature selection is reported for features selected at least 10% of the time.
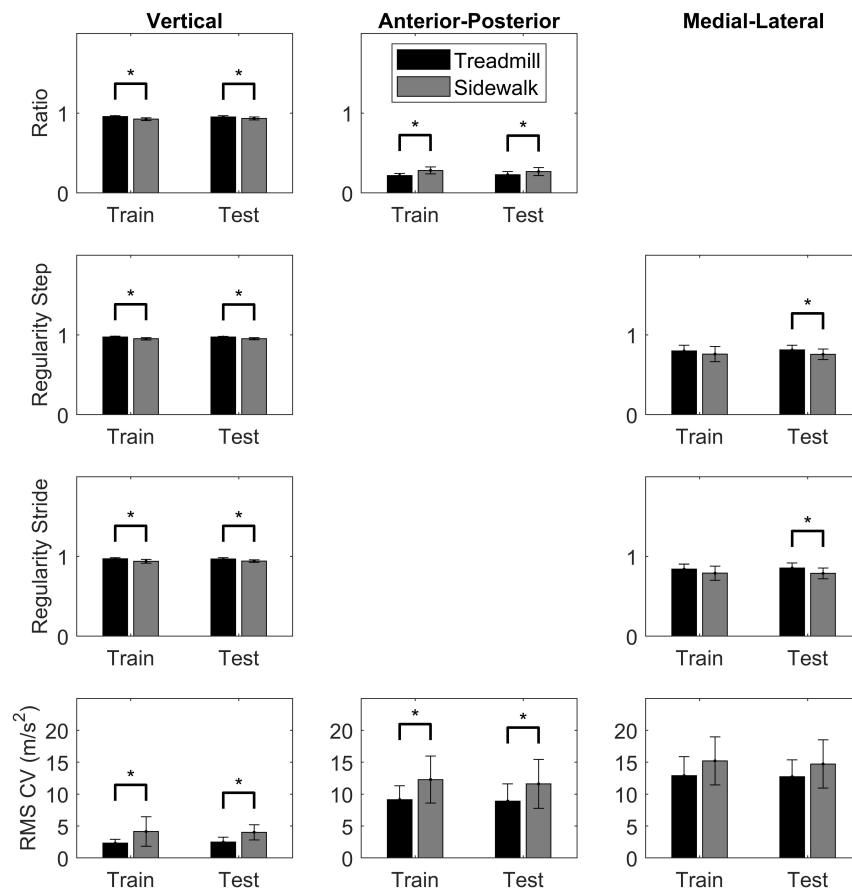
FIGURE 3 | Comparisons between treadmill (black) and sidewalk (gray) conditions for each of the nine selected features used in the model. Independent *t*-tests were used for the training dataset comparisons and paired *t*-tests were used for the testing dataset comparisons. Since a total of 18 comparisons were made, significance (*) was determined at $p < 0.003$.

regularity were significantly greater in the treadmill condition for the vertical axis in both the training and testing dataset, and in the medial-lateral axis for the testing dataset. During sidewalk running, there was significantly greater variability in the magnitude of the vertical and anterior-posterior accelerations for both datasets.

The initial classification accuracy based on 10-fold cross-validation of the training dataset ($M = 93.17\%$, SD = 2.43%) was greater than the classification accuracy of the independent testing dataset ($M = 83.81\%$, SD = 3.39%). Over 100 iterations, ten participants had both conditions correctly classified at least 82 times, and the remaining six had poor classification of one condition but perfect classification of the other condition (**Table 4**).

DISCUSSION

The purpose of this study was to classify running environments based on features extracted from a single accelerometer and identify features that would represent the difference between treadmill and sidewalk running. Sidewalk running

was characterized by lower regularity and greater variability than treadmill running and using these features, classification accuracy over 80% was achieved for both the training dataset and an independent dataset. These results are supported by Lindsay et al. (2014) who also reported that the treadmill running requires greater constraints and increased voluntary control during running gait. Thus, researchers must use caution when generalizing laboratory-based treadmill running results to real-world conditions for purposes such as rehabilitation of injuries, improved performance, and/or injury prevention (Benson et al., 2018a).

The observed changes in running patterns in different running environments are likely due to the consistency of the surfaces and/or speed in each environment. For example, a treadmill offers a smooth and consistent running surface and a constant speed for every step, whereas outdoor running presents more variable conditions with opportunities for changes in speed, surface, inclination, turns in the running path, other pedestrians/runners, and/or changes in weather or temperature (Ahamed et al., 2017, 2018; Benson et al., 2019). This lack of consistency likely contributed to the decrease in regularity in the vertical and medial-lateral dimensions, and changes in the ratios of the

Test participant	Number correct treadmill predictions	Number correct sidewalk predictions
1	100	1
2	100	21
3	100	25
4	100	95
5	100	96
6	100	100
7	100	100
8	100	100
9	99	100
10	92	98
11	87	99
12	84	97
13	82	100
14	53	100
15	40	100
16	13	100

magnitude of the acceleration. The decrease in regularity and observed shift to a greater ratio of horizontal accelerations than vertical accelerations when on sidewalk is consistent with previous research that has shown differences between stable and unstable surfaces based center of mass accelerations (Menz et al., 2003; Schütte et al., 2016) and stride time analyses (Lindsay et al., 2014). Sidewalk running was also characterized by greater variability in the magnitude of accelerations in all three dimensions. From a dynamical systems approach, a lack of coordinative variability in movement patterns may be associated with an unhealthy or pathological state (Hamill et al., 2012). However, the current study did not calculate coordinative variability in a manner similar to the methods proposed by Hamill et al. (2012), so future prospective studies should consider a link between the increased center of mass variability observed during sidewalk running and running-related injuries.

Due to the influence of speed on the magnitude of center of mass accelerations (Kobsar et al., 2014; Benson et al., 2018b), and the tendency to preferentially select a slightly slower speed during treadmill compared to overground running (Kong et al., 2012), speed was included as a potential feature in the classification model. However, speed was not one of the selected features used in the model. Therefore, differences in features related to the variability and consistency of the accelerometer signal had a greater role in discriminating between treadmill and sidewalk running.

The ability to generalize these results beyond the current study may be influenced by overfitting the classification model to the study participants (Ferber et al., 2016). Despite the use of 10-fold cross-validation of the training dataset to attempt to improve generalizability of classification, the model slightly overfit to the training dataset as there was lower classification accuracy for the independent testing dataset compared to the 10-fold cross-validation of the training dataset. Regarding real-world usability,

previous studies that have classified IMU-generated running and walking patterns have consistently reported classification accuracy greater than 80% (Kobsar et al., 2014, 2015; Phinyomark et al., 2014; Ahamed et al., 2018, 2019; Benson et al., 2018b; Clermont et al., 2018). Thus, the reported 93.17% accuracy for the training dataset and 83.81% accuracy for the independent testing dataset in the current study suggests that this classification mechanism has practical use.

The nearly 10% difference in classification accuracy between the training and testing datasets can be attributed to differences in running patterns between individuals in each dataset. In the cases where an individual in the testing dataset had a low classification rate for one environment, there was a perfect classification rate for the other environment. This result does not suggest that these misclassified participants have the same running pattern in both environments, but rather their running pattern on one environment is similar to the running patterns of other runners on the opposite. For example, the poor treadmill classification for test participant 16 (**Table 4**) was most likely driven by anterior-posterior variability in the treadmill condition that was greater than the sidewalk anterior-posterior variability for all participants in the training dataset. Yet, test participant 16 had perfect classification accuracy in the sidewalk condition as their anterior-posterior variability in the sidewalk condition was even greater than their treadmill value. Therefore, the misclassifications observed in this study highlight the potential strength of subject-specific models of running biomechanics to monitor changes in an individual's running biomechanics (Ahamed et al., 2018, 2019; Benson et al., 2019) and should be further investigated in future studies.

In addition to the previous limitations discussed, other limitations are acknowledged. First is the possibility that other unmeasured variables may also differ between running environments. The measured variables were previously used to quantify running patterns and were thus considered suitable for this study. However, *a priori* variable selection suggests a risk of investigator bias and may lead to the dismissal of potentially meaningful information that could be represented by other variables, such as metrics related to the accelerometer signal frequency content. Second, in addition to other accelerometer-based features, physiological metrics such as heart rate may differ between running environments. A further limitation is that although many of the features used in this study were on a scale of 0–1 (e.g., ratio of acceleration in a given axis, symmetry, regularity), other features were not on the same scale which may have influenced the contribution of each variable in the classification model. Nevertheless, six of the nine selected features, and four of the top-five features, were on the 0–1 scale, suggesting that features with values greater than 1 did not have an undue influence on the classification model.

In conclusion, we used a machine learning approach to successfully select features related to the consistency and variability of center of mass accelerations between treadmill and sidewalk running. Overall, step and stride regularity were significantly greater during treadmill running while sidewalk running resulted in significantly greater variability in the magnitude of the vertical and anterior-posterior accelerations.

Based on a 10-fold cross-validation of the training dataset we achieved a 93.17% classification accuracy, which was greater than the 83.81% classification accuracy of the independent testing dataset. The overall machine learning approach presented here could be utilized in future running biomechanical analyses to identify relevant differences in running patterns using IMUs.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Conjoint Human Research Ethics Board (CHREB) at the University of Calgary (REB16-1183). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RF, LB, and CC designed the study and were responsible for writing the manuscript. LB and CC were responsible for the data collection and data analysis.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Ahamed, N. U., Benson, L., Clermont, C., Osis, S. T., and Ferber, R. (2017). Fuzzy inference system-based recognition of slow, medium and fast running conditions using a triaxial accelerometer. *Procedia Comput. Sci.* 114, 401–407. doi: 10.1016/j.procs.2017.09.054

Ahamed, N. U., Kobsar, D., Benson, L., Clermont, C., Kohrs, R., and Osis, S. T. (2018). Using wearable sensors to classify subject-specific running biomechanical gait patterns based on changes in environmental weather conditions. *PloS One* 13:e0203839. doi: 10.1371/journal.pone.0203839

Ahamed, N. U., Kobsar, D., Benson, L. C., Clermont, C. A., Osis, S. T., and Ferber, R. (2019). Subject-specific and group-based running pattern classification using a single wearable sensor. *J. Biomech.* 84, 227–233. doi: 10.1016/j.jbiomech.2019.01.001

Avvenuti, M., Casella, A., and Cesarini, D. (2013). Using gait symmetry to virtually align a triaxial accelerometer during running and walking. *Electron. Lett.* 49, 120–121. doi: 10.1049/el.2012.3763

Barden, J. M., Clermont, C. A., Kobsar, D., and Beauchet, O. (2016). Accelerometer-based step regularity is lower in older adults with bilateral knee osteoarthritis. *Front. Hum Neurosci.* 10:625. doi: 10.3389/fnhum.2016.00625

Benson, L. C., Ahamed, N. U., Kobsar, D., and Ferber, R. (2019). New considerations for collecting biomechanical data using wearable sensors: number of level runs to define a stable running pattern with a single IMU. *J. Biomech.* 85, 187–192. doi: 10.1016/j.jbiomech.2019.01.004

Benson, L. C., Clermont, C. A., Bosnjak, E., and Ferber, R. (2018a). The use of wearable devices for walking and running gait analysis outside of the lab: a systematic review. *Gait Posture* 63, 124–138. doi: 10.1016/j.gaitpost.2018.04.047

Benson, L. C., Clermont, C. A., Osis, S. T., Kobsar, D., and Ferber, R. (2018b). Classifying running speed conditions using a single wearable sensor: optimal segmentation and feature extraction methods. *J. Biomech.* 71, 94–99. doi: 10.1016/j.jbiomech.2018.01.034

Caby, B., Kieffer, S., de Saint Hubert, M., Cremer, G., and Macq, B. (2011). Feature extraction and selection for objective gait analysis and fall risk assessment by accelerometry. *Biomed. Eng. Online* 10:1. doi: 10.1186/1475-925X-10-1

Chizi, B., and Maimon, O. (2010). "Dimension reduction and feature selection," in *Data Mining and Knowledge Discover Handbook*, eds O. Maimon, and L. Rokach (New York, NY: Springer), 83–100. doi: 10.1007/978-0-387-09823-4_5

Clermont, C. A., Benson, L. C., Osis, S. T., Kobsar, D., and Ferber, R. (2018). Running patterns for male and female competitive and recreational runners based on accelerometer data. *J. Sports Sci.* 37, 204–211. doi: 10.1080/02640414.2018.1488518

Dixon, P., Schütte, K., Vanwanseele, B., Jacobs, J. V., Dennerlein, J. T., Schiffman, J. M., et al. (2019). Machine learning algorithms can classify outdoor terrain types during running using accelerometry data. *Gait Posture* 74, 176–181. doi: 10.1016/j.gaitpost.2019.09.005

Ferber, R., Osis, S. T., Hicks, J. L., and Delp, S. L. (2016). Gait biomechanics in the era of data science. *J. Biomech.* 49, 3759–3761. doi: 10.1016/j.jbiomech.2016.10.033

Hamill, J., Palmer, C., and Van Emmerik, R. E. (2012). Coordinative variability and overuse injury. *Sports Med. Arthrosc. Rehabil. Ther. Technol.* 4:45. doi: 10.1186/1758-2555-4-45

Henriksen, M., Lund, H., Moe-Nilssen, R., Bliddal, H., and Danneskiod-Samsøe, B. (2004). Test–retest reliability of trunk accelerometric gait analysis. *Gait Posture* 19, 288–297. doi: 10.1016/s0966-6362(03)00069-9

Kobsar, D., Osis, S. T., Hettinga, B. A., and Ferber, R. (2014). Classification accuracy of a single tri-axial accelerometer for training background and experience level in runners. *J. Biomech.* 47, 2508–2511. doi: 10.1016/j.jbiomech.2014.04.017

Kobsar, D., Osis, S. T., Hettinga, B. A., and Ferber, R. (2015). Gait biomechanics and patient-reported function as predictors of response to a hip strengthening exercise intervention in patients with knee osteoarthritis. *PloS One* 10:e0139923. doi: 10.1371/journal.pone.0139923

Kong, P. W., Koh, T. M., Tan, W. C., and Wang, Y. S. (2012). Unmatched perception of speed when running overground and on a treadmill. *Gait Posture* 36, 46–48. doi: 10.1016/j.gaitpost.2012.01.001

Lee, J. B., Mellifont, R. B., and Burkett, B. J. (2010). The use of a single inertial sensor to identify stride, step, and stance durations of running gait. *J. Sci. Med. Sport* 13, 270–273. doi: 10.1016/j.jsams.2009.01.005

Lindsay, T. R., Noakes, T. D., and McGregor, S. J. (2014). Effect of treadmill versus overground running on the structure of variability of stride timing. *Percept. Mot. Skills* 118, 331–346. doi: 10.2466/30.26.pms.118k18w8

Menz, H. B., Lord, S. R., and Fitzpatrick, R. C. (2003). Acceleration patterns of the head and pelvis when walking on level and irregular surfaces. *Gait Posture* 18, 35–46. doi: 10.1016/s0966-6362(02)00159-5

Moe-Nilssen, R. (1998). A new method for evaluating motor control in gait under real-life environmental conditions. Part 1: the instrument. *Clin. Biomech.* 13, 320–327. doi: 10.1016/s0268-0033(98)00089-8

Moe-Nilssen, R., and Helbostad, J. L. (2004). Estimation of gait cycle characteristics by trunk accelerometry. *J. Biomech.* 37, 121–126. doi: 10.1016/s0021-9290(03)00233-1

Norris, M., Anderson, R., and Kenny, I. C. (2014). Method analysis of accelerometers and gyroscopes in running gait: a systematic review. *Proc.*

Inst. Mech. Eng. Part P J. Sports Engi. Technol. 228, 3–15. doi: 10.1177/1754337113502472

Phinyomark, A., Hettinga, B. A., Osis, S. T., and Ferber, R. (2014). Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *Plos One* 9:e105246. doi: 10.1371/journal.pone.0105246

Reenalda, J., Maartens, E., Homan, L., and Buurke, J. H. (2016). Continuous three dimensional analysis of running mechanics during a marathon by means of inertial magnetic measurement units to objectify changes in running mechanics. *J. Biomech.* 49, 3362–3367. doi: 10.1016/j.jbiomech.2016.08.032

Schütte, K. H., Aeles, J., De Beéck, T. O., van der Zwaard, B. C., Venter, R., Vanwanseele, B., et al. (2016). Surface effects on dynamic stability and loading during outdoor running using wireless trunk accelerometry. *Gait Posture* 48, 220–225. doi: 10.1016/j.gaitpost.2016.05.017

Shmilovici, A. (2010). "Support vector machines," in *Data Mining and Knowledge Discovery Handbook*, eds O. Maimon, and L. Rokach (New York, NY: Springer), 231–247.

Simon, S. R. (2004). Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems. *J. Biomech.* 37, 1869–1880. doi: 10.1016/j.jbiomech.2004.02.047

Taunton, J. E., Ryan, M. B., Clement, D. B., McKenzie, D. C., Lloyd-Smith, D. R., Zumbo, B. D., et al. (2003). A prospective study of running injuries: the Vancouver Sun Run "In Training" clinics. *Br. J. Sports Med.* 37, 239–244. doi: 10.1136/bjsm.37.3.239

Wundersitz, D. W., Gastin, P. B., Richter, C., Robertson, S. J., and Netto, K. J. (2015). Validity of a trunk-mounted accelerometer to assess peak accelerations during walking, jogging and running. *Eur. J. Sport Sci.* 15, 382–390. doi: 10.1080/17461391.2014.955131

Check for updates

Interpreting Deep Learning Features for Myoelectric Control: A Comparison With Handcrafted Features

Ulysse Côté-Allard[1†], Evan Campbell[2*†], Angkoon Phinyomark[2], François Laviolette[3], Benoit Gosselin[1‡] and Erik Scheme[2‡]*

[1] Department of Computer and Electrical Engineering, Université Laval, Quebec, QC, Canada, [2] Department of Electrical and Computer Engineering, Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, Canada, [3] Department of Computer Science and Software Engineering, Université Laval, Quebec, QC, Canada

Existing research on myoelectric control systems primarily focuses on extracting discriminative characteristics of the electromyographic (EMG) signal by designing handcrafted features. Recently, however, deep learning techniques have been applied to the challenging task of EMG-based gesture recognition. The adoption of these techniques slowly shifts the focus from feature engineering to feature learning. Nevertheless, the black-box nature of deep learning makes it hard to understand the type of information learned by the network and how it relates to handcrafted features. Additionally, due to the high variability in EMG recordings between participants, deep features tend to generalize poorly across subjects using standard training methods. Consequently, this work introduces a new multi-domain learning algorithm, named ADANN (Adaptive Domain Adversarial Neural Network), which significantly enhances ($p = 0.00004$) inter-subject classification accuracy by an average of 19.40% compared to standard training. Using ADANN-generated features, this work provides the first topological data analysis of EMG-based gesture recognition for the characterization of the information encoded within a deep network, using handcrafted features as landmarks. This analysis reveals that handcrafted features and the learned features (in the earlier layers) both try to discriminate between all gestures, but do not encode the same information to do so. In the later layers, the learned features are inclined to instead adopt a one-vs.-all strategy for a given class. Furthermore, by using convolutional network visualization techniques, it is revealed that learned features actually tend to ignore the most activated channel during contraction, which is in stark contrast with the prevalence of handcrafted features designed to capture amplitude information. Overall, this work paves the way for hybrid feature sets by providing a clear guideline of complementary information encoded within learned and handcrafted features.

Keywords: EMG, deep learning, MAPPER, feature extraction, gesture recognition, CNN, ConvNet, Grad-CAM

1. INTRODUCTION

Surface Electromyography (sEMG) is a technique employed in a vast array of applications from assistive technologies (Phinyomark et al., 2011c; Scheme and Englehart, 2011) to bio-mechanical analysis (Andersen et al., 2018), and more generally as a way to interface with computers and robots (Zhang et al., 2009; St-Onge et al., 2019). Traditionally, the sEMG-based gesture recognition literature primarily focuses on feature engineering as a way to increase the information density of the signal to improve gesture discrimination (Oskoei and Hu, 2007; Scheme and Englehart, 2011; Phinyomark et al., 2012a). In the last few years, however, researchers have started to leverage deep learning (Allard et al., 2016; Atzori et al., 2016; Phinyomark and Scheme, 2018a), shifting the paradigm from feature engineering to feature learning.

Deep learning is a multi-level representation learning method (i.e., methods that learn an embedding from an input to facilitate detection or classification), where each level generates a higher, more abstract representation of the input (LeCun et al., 2015). Conventionally, the output layer (i.e., classifier or regressor) only has direct access to the output of the highest representation level (LeCun et al., 2015; Alom et al., 2018). In contrast, several works have also fed the intermediary layers' output directly to the network's head (Sermanet et al., 2013; Long et al., 2015; Yang and Ramanan, 2015). Arguably, the most successful approach using this design philosophy is DenseNet (Huang et al., 2017), a type of convolutional network (ConvNet) where each layer receives the feature maps of all preceding layers as input. Features learned by ConvNets were also extracted to be employed in conjunction with (or replace) handcrafted features when training conventional machine learning algorithms (e.g., support vector machine, linear discriminant analysis, decision tree) (Poria et al., 2015; Nanni et al., 2017; Chen et al., 2019; Liu et al., 2019). Within the context of sEMG-based gesture recognition, deep learning was shown to be competitive with the current state of the art (Côté-Allard et al., 2019a) and when combined with handcrafted features, to outperform it (Chen et al., 2019). This last result seems to indicate that, for sEMG signals, deep-learned features provide useful information that may be complementary to those that have been engineered throughout the years. However, the black box nature of these deep networks means that understanding what type of information is encapsulated throughout the network, and how to leverage this information, is challenging.

The main contribution of this work is, therefore, to provide the first extensive analysis of the relationship between handcrafted and learned features within the context of sEMG-based gesture recognition. Understanding the feature space learned by the network could shed new insights on the type of information contained in sEMG signals. In turn, this improved understanding will allow the creation of better handcrafted features and facilitate the creation of new hybrid feature sets using this feature learning paradigm.

An important challenge arises when working with biosignals, as extensive variability exists between subjects (Guidetti et al., 1996; Batchvarov and Malik, 2002; Meltzer et al., 2007; Castellini et al., 2009; Halaki and Ginn, 2012). Especially within the

context of sEMG-based gesture recognition (Castellini et al., 2009; Halaki and Ginn, 2012). Consequently, features learned using traditional deep learning training methods can be highly participant-specific, which would hinder the goal of this work of learning a general feature representation of sEMG signals. By defining each participant as a different *domain*, however, this issue can be framed as a Multi-Domain Learning problem (MDL) (Yang and Hospedales, 2014), with the added restriction that the network's weights should be participant-agnostic. Multiple popular and effective MDL algorithms have been proposed over the years (Nam and Han, 2016; Rebuffi et al., 2018). For example, Nam and Han (2016) proposed to use a shared network across multiples domains with one predictive head per domain. In Yang and Hospedales (2014), a single head was shared across two parallel networks with one of them receiving the example's representation as input, while the other receives a vector representation of the associated domain of the example. These algorithms however are ill-suited for this work's context as they: do not explicitly impose domain-agnostic weight learning (Yang and Hospedales, 2014), can scale poorly with the number of domains (i.e., participants) (Nam and Han, 2016), or are restricted to encode a single domain within their learned features (and use adaptor blocks to bridge the gap between domains) (Rebuffi et al., 2018). Unsupervised domain-adversarial training algorithms (Ajakan et al., 2014; Ganin et al., 2016; Tzeng et al., 2017; Shu et al., 2018) predict an unlabeled dataset by learning a representation on a labeled dataset that makes it hard to distinguish between examples from either distribution. However, these algorithms are often not designed to learn a unique representation across more than two domains simultaneously (Ajakan et al., 2014; Ganin et al., 2016; Tzeng et al., 2017; Shu et al., 2018), can be destructive to the source domain representation (through iterative process) (Shu et al., 2018), and by nature of the problem they are trying to solve, do not leverage the labels of the target domains. As such, this work presents a new multi-domain adversarial training algorithm, named ADANN (Adaptive Domain Adversarial Neural Network). ADANN trains a network across multiple domains simultaneously while explicitly penalizing any domain-variant representations to study learned features that generalize well across participants.

In this work, the sEMG information encapsulated within the general deep learning features learned by ADANN, is characterized using handcrafted features as landmarks in a topological network. This network is generated via the Mapper algorithm (Singh et al., 2007), with *t*-Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008), a non-linear dimensionality reduction visualization method, as the filter function. Mapper is a Topological Data Analysis (TDA) tool that excels at determining the shape of high dimensional data, by providing a faithful representation of it through a topological network. This TDA tool has been applied as a solution to numerous challenging applications across a wide array of domains; for example, uncovering the dynamic organization of brain activity during various tasks (Saggar et al., 2018) or identifying a subgroup of breast cancer with 100% survival rate and no metastasis (Nicolau et al., 2011). Mapper has also been
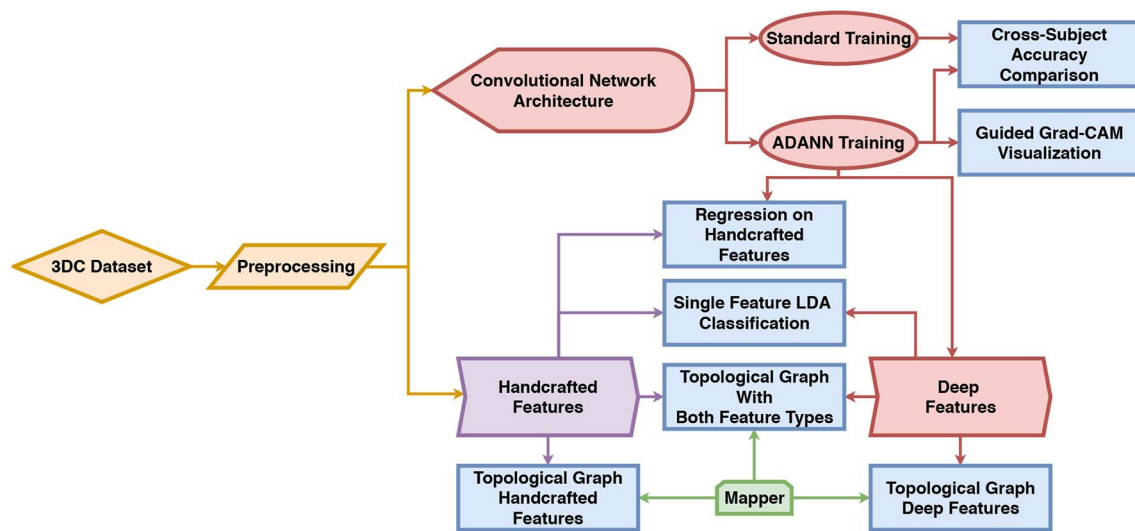
FIGURE 1 | Diagram of the workflow of this work. The *3DC Dataset* is first preprocessed before being used to train the network using standard training and the proposed ADANN training procedure. The handcrafted features are directly calculated from the preprocessed dataset, while the deep features are extracted from the ConvNet trained with ADANN. In the diagram, the blue rectangles represent experiments and the arrows show which methods/algorithms are required to perform them.

applied to determine relationships between feature space for physiological signal pain recognition (Campbell et al., 2019b), and EMG-based gesture recognition (Phinyomark et al., 2017). However, to the best of the authors' knowledge, the use of TDA to interpret information harnessed within deep-learned features using handcrafted features as landmarks has yet to be explored.

In this paper, convNet visualization techniques are also leveraged as a way to highlight how the network makes class-discriminant decisions. Several works (Simonyan et al., 2013; Springenberg et al., 2014; Zeiler and Fergus, 2014; Gan et al., 2015) have proposed to visualize network's predictions by emphasizing which input-pixels have the most impact on the network's output, consequently, fostering a better understanding of what the network has learned. For example, Simonyan et al. (2013) used partial derivatives to compute pixel-relevance for the network output. Another example is Guided Backpropagation (Springenberg et al., 2014), which modifies the computation of the gradient to only include paths within the network that positively contribute to the prediction of a given class. When compared with saliency maps (Simonyan et al., 2013), Guided Backpropagation results in qualitative visualization improvements (Selvaraju et al., 2017). While these methods produce resolutions at a pixel level, the images produced with respect to different classes are nearly identical (Selvaraju et al., 2017). Other types of algorithms provide highly class-discriminative visualizations, but at a lower resolution (Selvaraju et al., 2016; Zhou et al., 2016) and sometimes require a specific ConvNet architecture (Zhou et al., 2016) to use. Within this work, Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) (Selvaraju et al., 2017) is employed as it provides pixel-wise input resolution while being class-discriminative. Another advantage of this technique is that it can be implemented on any ConvNet-based architecture without

requiring re-training. To the best of the authors' knowledge, this is the first time that deep learning visualization techniques are applied to EMG signals.

2. MATERIALS AND METHODS

A flowchart of the material, methods and experiment is shown in **Figure 1**. This section is divided as follows: first, a description of the dataset and preprocessing used in this work is given in section 2.1. Then, the handcrafted features are presented in section 2.2. The ConvNet architecture and the new multi-domain adversarial training algorithm (ADANN) are presented in sections 2.3.1 and 2.3.2, respectively. A brief overview of Guided Grad-CAM is given in section 2.3.3, while sections 2.3.4 and 2.3.5 present single feature classification and handcrafted feature regression, respectively. Finally, the Mapper algorithm is detailed in section 2.4.

2.1. EMG Data

The dataset employed in this work is the *3DC Dataset* (Côté-Allard et al., 2019b), featuring 22 able-bodied participants performing ten hand/wrist gestures + neutral (see **Figure 2** for the list of gestures). This dataset was recorded with the 3DC Armband; a wireless, 10-channel, dry-electrode, 3D printed sEMG armband. The device samples data at 1,000 Hz per channel, allowing the feature extraction to take advantage of the full spectra of sEMG signals (Phinyomark and Scheme, 2018b). Informed consent was obtained from all participants, as approved by Laval University's Research Ethics Committee (Côté-Allard et al., 2019b).

The dataset was built as follows: Each participant was asked to perform and hold each gesture for a period of 5 s starting from the neutral position to produce a *cycle*. Three more cycles were
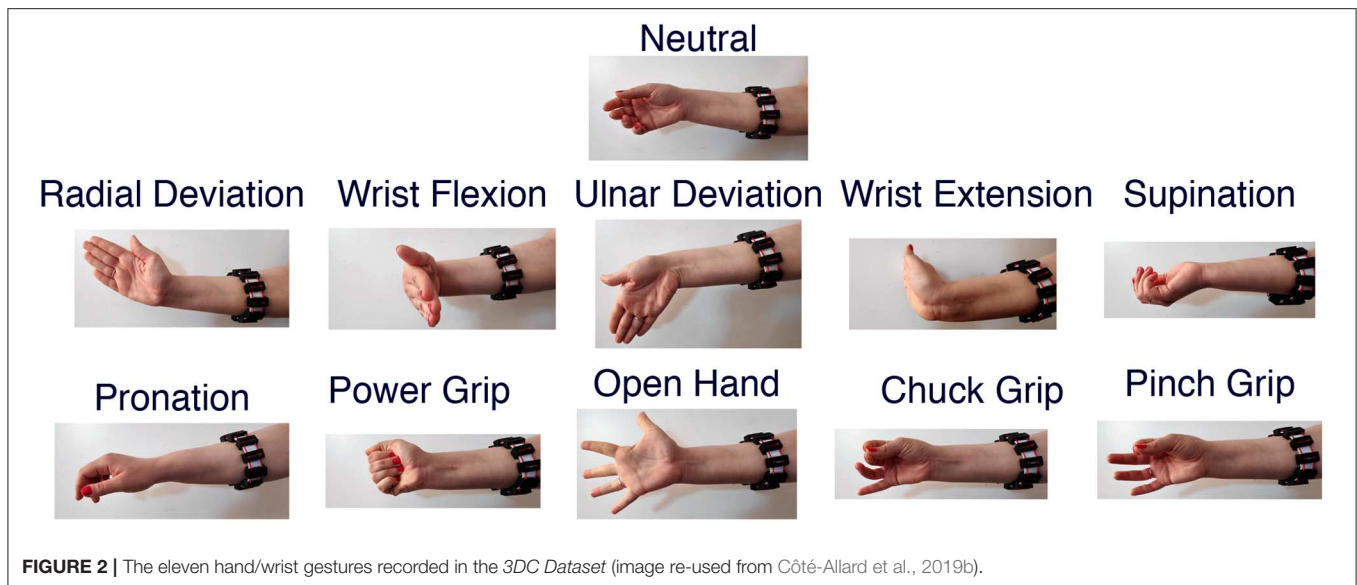
FIGURE 2 | The eleven hand/wrist gestures recorded in the *3DC Dataset* (image re-used from Côté-Allard et al., 2019b).

recorded to serve as the *training dataset*. After a 5 min break, four new cycles were recorded to serve as the *test dataset*. Note that the validation set and hyperparameter selection are made from the training dataset.

As this work aims to understand the type of features learned by deep network in the context of myoelectric control systems, a critical factor to consider is the input latency. Smith et al. (2010) showed that the optimal guidance latency was between 150 and 250 ms. As such, the data from each participant was segmented into 151 ms frames with an overlap of 100 ms. The raw data was then band-pass filtered between 20 and 495 Hz using a fourth-order Butterworth filter.

2.2. Handcrafted Features

Handcrafted features are characteristics extracted from windows of the EMG signal using established mathematical equations. The purpose of these feature extraction methods is to enhance the information density of the signal so as to improve discrimination between motion classes (Oskoei and Hu, 2007; Phinyomark et al., 2012a). Across the myoelectric control literature, hundreds of handcrafted feature extraction methods have been presented (Oskoei and Hu, 2007; Phinyomark et al., 2012a, 2013). As such, implementing the exhaustive set of features that has been proposed is impractical. Instead, within this study a comprehensive subset of 79 of the most commonly used features is employed. With a comprehensive set of features, past literature has identified five functional groups that summarize all sources of information current handcrafted feature extraction techniques describe: signal amplitude and power (SAP), non-linear complexity (NLC), frequency information (FI), time-series modeling (TSM), and unique (UNI) (Phinyomark et al., 2017; Campbell et al., 2019a). The SAP functional group includes time-domain energy or power features (e.g., Root Mean Squared, Mean Absolute Value). The FI functional group generally refers to features extracted from the frequency domain, or

features that describe spectral properties (e.g., Mean Frequency, Zero Crossings). The NLC functional group corresponds to features that describe entropy or similarity based information (e.g., Sample Entropy, Maximum Fractal Length). The TSM functional group represents features that attempt to reconstruct the data provided through stochastic or other algorithmic models (e.g., Autoregressive Coefficients, Cepstral Coefficients). Finally, the UNI functional group represents features that capture various other modalities of information, such as measures of signal quality or a combination of other functional groups (e.g., Signal to Motion Artifact Ratio, Time Domain Power Spectral Descriptors).

Table 1 presents the 56 handcrafted feature methods considered in this work. Note that some methods produce multiple features (e.g., Cepstral Coefficients, Histogram), resulting in a total of 79 features. The SAP, FI, NLC, TSM, and UNI feature groups are represented here by 25, 5, 6, 7, and 13 feature extraction methods, respectively. In the TDA of the deep learned features (see section 2.4), these handcrafted features serve as landmarks for well-understood properties of the EMG signal. In the regression model analysis (see section 2.3.5), the flow of information through the ConvNet is visualized by employing the handcrafted features methods as the target of the network.

2.3. Convolutional Network

The following subsections present the deep learning architecture, training methods and visualization techniques employed in this paper. The PyTorch (Paszke et al., 2017) implementation employed in this work is available at: https://github.com/UlysseCoteAllard/sEMG_handCraftedVsLearnedFeatures.

2.3.1. Architecture

Recent works on sEMG-based gesture recognition using deep learning have shown that ConvNets trained with the raw sEMG signal as input were able to achieve similar classification accuracy

TABLE 1 | Handcrafted features extracted for topological landmarks sorted by functional group.

References	Feature extraction method	Name	Group
Phinyomark et al. (2012a)	Amplitude of the first burst	AFB	SAP
Kim et al. (2011)	Difference absolute mean value	DAMV	SAP
Kim et al. (2011)	Difference absolute standard deviation value	DASDV	SAP
Zardoshti-Kermani et al. (1995)	Difference log detector	DLD	SAP
Phinyomark et al. (2012a)	Difference temporal moment	DTM	SAP
Zardoshti-Kermani et al. (1995)	Difference variance value	DVARV	SAP
Zardoshti-Kermani et al. (1995)	Difference v-order	DV	SAP
Park and Lee (1998)	Integral of electromyogram	IEMG	SAP
Zardoshti-Kermani et al. (1995)	Log detector	LD	SAP
Al-Timemy et al. (2015)	Second-order moment	M2	SAP
Oskoei and Hu (2008)	Modified mean absolute value 1	MMAV1	SAP
Oskoei and Hu (2008)	Modified mean absolute value 2	MMAV2	SAP
Saponas et al. (2008)	Mean absolute value	MAV	SAP
Phinyomark et al. (2012a)	Maximum	MAX	SAP
Du and Vuskovic (2004)	Multiple hamming windows	MHW	SAP
Du and Vuskovic (2004)	Mean power	MNP	SAP
Du and Vuskovic (2004)	Multiple trapezoidal windows	MTW	SAP
Saponas et al. (2008)	Root mean squared	RMS	SAP
Du and Vuskovic (2004)	Spectral moment	SM	SAP
Du and Vuskovic (2004)	Sum of squared integral	SSI	SAP
Phinyomark et al. (2012a)	Temporal moment	TM	SAP
Du and Vuskovic (2004)	Total power	TTP	SAP
Zardoshti-Kermani et al. (1995)	Variance	VAR	SAP
Zardoshti-Kermani et al. (1995)	v-Order	V	SAP
Phinyomark et al. (2012a)	Waveform length	WL	SAP
Oskoei and Hu (2006, 2008)	Frequency ratio	FR	FI
Thongpanja et al. (2013, 2015)	Median frequency	MDF	FI
Thongpanja et al. (2013, 2015)	Mean frequency	MNF	FI
Phinyomark et al. (2012a)	Slope sign change	SSC	FI
Zardoshti-Kermani et al. (1995)	Zero crossings	ZC	FI
Phinyomark et al. (2013)	Sample entropy	SAMPEN	NLC
Phinyomark et al. (2013)	Approximate entropy	APEN	NLC
Zardoshti-Kermani et al. (1995)	Willison's amplitude	WAMP	NLC
Gitter and Czerniecki (1995)	Box-counting fractal dimension	BC	NLC
Gupta et al. (1997)	Katz fractal dimension	KATZ	NLC
Arjunan and Kumar (2010)	Maximum fractal length	MFL	NLC
Park and Lee (1998)	Autoregressive coefficients	AR	TSM
Park and Lee (1998)	Cepstral coefficients	CC	TSM
Park and Lee (1998)	Difference autoregressive coefficient	DAR	TSM
Park and Lee (1998)	Difference cepstral coefficients	DCC	TSM
Phinyomark et al. (2011d, 2012b)	Detrend fluctuation analysis	DFA	TSM
Qingju and Zhizeng (2006)	Power spectrum ratio	PSR	TSM
Sinderby et al. (1995) and McCool et al. (2014)	Signal to noise ratio	SNR	TSM
Phinyomark et al. (2011a,b)	Critical exponent	CE	UNI
Sinderby et al. (1995) and McCool et al. (2014)	Maximum to minimum drop in power density ratio	DPR	UNI
Phinyomark et al. (2012a)	Histogram	HIST	UNI
Thongpanja et al. (2016) and Van Den Broek et al. (2006)	Kurtosis	KURT	UNI
Phinyomark et al. (2012a)	Mean absolute value slope	MAVS	UNI
Sinderby et al. (1995) and McCool et al. (2014)	Power spectrum deformation	OHM	UNI
Phinyomark et al. (2013)	Peak frequency	PKF	UNI
Talebinejad et al. (2009)	Power spectrum density fractal dimension	PSDFD	UNI
Thongpanja et al. (2016) and Van Den Broek et al. (2006)	Skewness	SKEW	UNI

(Continued)

TABLE 1 | Continued

References	Feature extraction method	Name	Group
Sinderby et al. (1995) and McCool et al. (2014)	Signal to motion artifact ratio	SMR	UNI
Al-Timemy et al. (2015)	Time domain power spectral descriptors	TSPSD	UNI
Phinyomark et al. (2012a)	Variance of central frequency	VCF	UNI
Phinyomark et al. (2013)	Variance fractal dimension	VFD	UNI

to the current state of the art (Zia ur Rehman et al., 2018; Côté-Allard et al., 2019a). Consequently, and to reduce bias, the preprocessed raw data (see section 2.1) is passed directly as an image of shape **10 × 151** (Channel × Sample) to the ConvNet.

The ConvNet's architecture, which is depicted in **Figure 3**, contains six *blocks* followed by a fully connected layer for gesture-classification. The network's topology was selected to obtain a deep network with a limited number of learnable parameters (to avoid overfitting) with simple layer connections to enable an easier, and thus more thorough analysis. All architecture choices and hyperparameter selection were performed using the training set of the *3DC Dataset* or inspired by previous works (Côté-Allard et al., 2019a,b). Each block encapsulates a convolutional layer (LeCun et al., 2015), followed by batch normalization (BN) (Ioffe and Szegedy, 2015), leaky ReLU (slope = 0.1) (Xu et al., 2015) and dropout (Gal and Ghahramani, 2016) (with a drop rate set at 0.35 following Côté-Allard et al., 2019a). The number of blocks within the network was selected to obtain a sufficiently deep network to study how the type of learned features evolve with respect to their layer. The depth of the network was limited by the number of examples available for training and more complex layer connections [e.g., residual network (He et al., 2016), dense network (Huang et al., 2017)] were avoided to not ambiguate the analysis performed in this work. The number of feature maps (64) was kept uniform for each layer, allowing for easier comparisons of learned features across the convolutional layers. The filter size was **1 × 26** so that, similarly to the handcrafted features, the learned features are channel independent. Due to the selected filter size, the dimensions of feature maps at the final layer is 10×1.

Adam (Kingma and Ba, 2014) was employed to optimize the ConvNet with an initial learning rate of 0.0404709 and batch size of 512 (as used in Côté-Allard et al., 2019b). The training dataset was divided into training and validation sets using the first three cycles and last cycle, respectively. Employing this validation set, learning rate annealing was applied with a factor of five and a patience of fifteen with early stopping applied when two consecutive annealings occurred without achieving a better validation loss.

For the purpose of the TDA, features maps were extracted after the non-linearity using per feature-map channel-wise average pooling. That is, the number of feature maps remained the same, but the feature map's value per channel was averaged to a single scalar (as is common with handcrafted features).

2.3.2. Multi-Domain Adversarial Training

To better understand what type of features are commonly learned at each layer of the network, it is desirable that the model generalizes well across participants. This feature generality principle also motivates the design of the handcrafted features (presented in section 2.2), as it would be impractical to create new features for each new participant. Learning a general feature representation across participants, however, cannot be achieved by simply aggregating the training data of all participants and then training a classifier normally. As, even when precisely controlling for electrode placement, cross-subject accuracy using standard learning methods is poor (Castellini et al., 2009). This problem is compounded by the fact that important differences exist between subjects of the *3DC Dataset* (i.e., position and rotation of the armband placed on the left or right arm).

Learning a participant-agnostic representation can be framed as a multi-domain learning problem (Nam and Han, 2016). In the context of sEMG-based gesture recognition, AdaBN, a domain adaptation algorithm presented in Li et al. (2016), was successfully employed as a way to learn a general representation across participants in Cote-Allard et al. (2017), Côté-Allard et al. (2019a). The hypothesis of AdaBN is that label-related information (i.e., hand gestures) will be contained within the network's weights, while the domain-related information (i.e., participants) are stored in their BN statistics. Training is thus performed by sharing the weights of the network across the subjects dataset while tracking the BN statistics independently for each participant.

To inhibit the shared network's weights from learning subject-specific representation, Domain-Adversarial Neural Networks (DANN) training (Ganin et al., 2016) is employed. DANN is designed to learn domain-invariant features across two domains from the point of view of the desired task. The approach used by DANN to achieve this objective consists of adding a second head (referred to as the *domain classification head*) to the network presented in section 2.3.1, which receives the output of block B6. The goal of this second head is to learn to discriminate between the domains. However, during backpropagation, the gradient computed from the domain loss is multiplied by a negative constant (set to -1 in this work) as it exits the domain classification head. This gradient reversal explicitly forces the feature distributions over the domains to be similar. Note that the backpropagation algorithm proceeds normally for the first head (gesture classification head). The loss function used for both heads is the cross-entropy loss. The two losses are combined as follows: $\mathcal{L}_y + \lambda \mathcal{L}_d$, where $\mathcal{L}_y$ and $\mathcal{L}_d$ are the prediction and domain loss, respectively (see **Figure 4**), while λ is a scalar that weights the domain loss (set to 0.1 in this work).

Using this approach, each participant of the 3DC Dataset represents a different domain (n=22). A direct application of DANN would thus initialize the domain classification head

FIGURE 3 | The ConvNet's architecture, employing 543,629 learnable parameters. In this figure, *Bi* refers to the ith feature extraction block (i∈{1,2,3,4,5,6}). Conv refers to Convolutional layer. As shown, the feature extraction is performed after the non-linearity (leaky ReLU).

FIGURE 4 | Overview of the training steps of ADANN (identical to DANN) for one labeled batch from the source ({x_s, y_s}, blue lines) and one unlabeled batch from the target ({x_t}, red dashed lines). The purple dotted lines correspond to the backpropagated gradient. The gradient reversal operation is represented by the purple diamond.

with 22 output neurons. This, however, could create a pitfall where the network is able to differentiate between the domains perfectly while simply predict one of the 21 other domains to maximize $\mathcal{L}_d$. Instead, the domain classification head is initialized with only two output neurons. At each epoch, a batch is created that contains examples from a single participant (this batch is referred to as the *source batch*, and is assigned the domain label 0). A second batch, referred to as the *target batch*, is also created that contains examples from one of the other participants selected at random, and is assigned domain label 1. As every participants data is used as the

source batch at each epoch, this ensures that the network is forced to learn a domain-independent feature representation. ADANN's goal is thus to force the network to be unable to accurately associate a participant with their examples while achieving a highly discriminative gesture representation across all participants. During training, the BN statistics are tracked individually for each subject. Therefore, when learning from a source or target batch, the network uses the BN statistics associated with the corresponding participant. Note that, by construction, the participant associated with the source is necessarily different from the participant associated with the

target. Consequently, the network is fed the source and target batch consecutively (i.e., not both batch simultaneously). Also note that the BN statistics are updated only in association with the source batch to ensure equal training updates across all participants. For a given iteration, once the source and target batch are constructed, the training step proceeds as described for DANN (see **Figure 4**).

To assess the performance of the proposed MDL algorithm, two identical ConvNet (as described in section 2.3.1) were created. One of the ConvNets was trained with ADANN, whereas the other used a standard training loop (i.e., aggregating the data from all participants), with both using the same hyperparameters. The networks trained with both methods were then tested on the test dataset with no participant-specific fine-tuning.

2.3.3. Learning Visualization

One of the main problems associated with deep learning is interpretability of how and why a model makes a prediction given a particular input. A first step in understanding a network prediction is through the visualization of the learned weights, feature maps and gradients resulting from a particular input. Consequently, several sophisticated visualization techniques have been developed, which are aimed at facilitating a better comprehension of the hierarchical learning that takes place within a network (Simonyan et al., 2013; Springenberg et al., 2014; Zhou et al., 2016). One popular such technique is Guided Grad-CAM, which combines high resolution pixel-space gradient visualization and class-discriminative visualization (Selvaraju et al., 2017). Guided Grad-CAM is thus employed to visualize how the ConvNet trained with ADANN makes its decisions, both on real examples from the *3DC Dataset* and on an artificially generated signals.

Given an image that was used to compute a forward pass in the network and a label y, the output of Guided Grad-CAM is calculated from four distinct steps (note that steps two and three are computed independently from each other using the output of step one):

1. Set all the gradients of the output neurons to zero, except for the gradient of the neuron associated with the label y (which is set to one) and name the gradient of the neuron of interest y^g.
2. Set all negative activations to zero. Then, perform backpropagation, but before propagating the gradient at each step, set all the negative gradients to zero again. Save the final gradients corresponding to the input image. This step corresponds to computing the guided backpropagation (Springenberg et al., 2014).
3. Let $F_{j,i}$ be the activation of the ith feature map of the jth layer with feature maps of the network. Select a layer F_j of interest (in this work F_j correspond to the rectified convolutional layer of B6). Backpropagate the signal from the output layer to $F_{j,i}$ (i.e., $\frac{\partial y^g}{\partial F_{j,i}}$). Then for each i compute the global average pooling of $\frac{\partial y^g}{\partial F_{j,i}}$ and name it $w_{j,i}$. Finally, compute: $ReLU\left(\sum_i w_{j,i} F_{j,i}\right)$. This third step corresponds to computing the Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2016).

4. Finally, fuse the output of the two previous steps using point-wise multiplication to obtain the output of Guided Grad-CAM (Selvaraju et al., 2017).

2.3.4. Learned Feature Classification

Similarly to Chen et al. (2019), the learned features were extracted to train a Linear Discriminant Analysis (LDA) classifier to show the discriminative ability of the learned features. LDA was selected as it was shown to provide robust classification within the context of sEMG-based gesture recognition (Campbell et al., 2019c), does not require hyperparameter tuning, and creates linear boundaries within the input feature space. LDA was trained in a cross-subject framework on the training dataset and tested on the test dataset. For comparison purposes, LDA was also trained on the handcrafted features described in section 2.2. Note that the implementation was from scikit-learn (Pedregosa et al., 2011).

2.3.5. Regression Model

One method of highlighting the information content encoded throughout a network is to see how well-known handcrafted features can be predicted from the network's feature maps at different stages. This can be achieved using an added output neuron (*regression head*) at the feature extraction stage [i.e., after the non-linearity, but before the average pooling (before the green trapezoid of **Figure 3**)] of each block. The goal of this output is to map from the learned features to the handcrafted features of interest. As all the features considered in section 2.2 are calculated channel-wise, only the information from the first sEMG channel (arbitrarily selected) of the feature maps will be fed to the regression head.

The training procedure to implement this is as follows: first, pre-train the network using ADANN (presented in section 2.3.2). Second, freeze all the weights of the network, except for the weights associated with the regression head of the block of interest. The Mean Square Error (MSE) is then employed as the loss function with the target being the value of the handcrafted feature of interest from the first sEMG channel. Due to the stochastic nature of the algorithm, the training was performed 20 times for each participant and the results were given as the average MSE computed on the test dataset across of all participants. Note that the targets derived from multi-output feature extraction methods (e.g., Autoregressive Coefficients) corresponded to the first principal component returned by Principal Component Analysis (PCA) (where singular value decomposition was performed on the training and test set for the training and test phase, respectively).

2.4. Topological Data Analysis—Mapper

Conventional TDA methods, such as Isomap (Balasubramanian and Schwartz, 2002) produce a low dimensional embedding by retaining geodesic distances between neighboring points. However, they often have limited topological stability (Choi and Choi, 2007) and lack the ability to produce a simplicial complex (a ball-and-stick simplification of the shape of the dataset) with size smaller than the original dataset (Singh et al., 2007). The Mapper algorithm (Singh et al., 2007) is a TDA method that creates interpretable simplifications of high-dimensional data sets that remain true to the shape of the data set. Mapper can thus produce

a stable representation of the topological shape of the dataset at a specified resolution, where the shape of the network has been simplified during a partial clustering stage. Further, the shape of the dataset is defined such that it is coordinate, deformation, and compression invariant. Consequently, this TDA algorithm can be employed to better understand how handcrafted and deep-learned features relate to one-another. In this work, Mapper is employed on three scenarios; (A), (B), and (C). In scenario (A), the algorithm only uses the handcrafted features as a way to validate the hyperparameters selected by cross-referencing the results with previous EMG works using Mapper (Phinyomark et al., 2017; Campbell et al., 2019a). For scenario (B), only the learned features are used to determine if features within the same block extract similar or dissimilar sources of information (i.e., the degree at which the features within the same block are dispersed across the topological network). Finally, in scenario (C), Mapper is applied to the combination of learned and handcrafted features to better understand their relationship and to provide new avenues of research for sEMG-based gesture recognition.

Sections 2.4.1–2.4.2, below, provide additional details about the approach, mathematical basis and implementation of Mapper in this work. Readers who are familiar with, or prefer to avoid these details, may jump directly to section 3.

2.4.1. Mapper Algorithm
The construction of the topological network created using the Mapper algorithm can be seen as a five stage pipeline:

1. *prepare*: organize the data set to produce a point cloud of features in high dimensional space.
2. *lens*: filter the high dimensional data into a lower dimensional representation using a lens.
3. *resolution*: divide the filtration into a set of regions.
4. *partial clustering*: for each region, cluster the contents in the original high dimensional space.
5. *combine*: combine the region isolated clusters into a single topological network using common points across regions (Geniesse et al., 2019).

2.4.2. Mathematical Definition of Mapper
A mathematical definition of the Mapper algorithm for feature extraction using a multi-channel recording device is as follows:

Let $\mathbf{x} \stackrel{\text{def}}{=} (\vec{x}_1, \ldots, \vec{x}_C)$ be a series of samples for each C channels, where $\vec{x}_c \in \mathbb{R}^S, \forall c \in \{1, \ldots, C\}$ and S is the length of a consecutive series of data. Define $\mathcal{X} \stackrel{\text{def}}{=} \{\mathbf{x}_n\}_{n=1}^N$ a set of N examples. Let also $\Phi \stackrel{\text{def}}{=} \{\phi_m\}_{m=1}^M$ be a set of M feature-generating functions of the form $\phi_m : \mathbb{R}^S \to \mathbb{R}$. Given $\mathbf{x}_{n,c}$ the c th element of $\mathbf{x}_n \in \mathcal{X}$, the resulting feature $f_{n,c}^m \in \mathbb{R}$ is obtained by applying ϕ_m such that $f_{n,c}^m \stackrel{\text{def}}{=} \phi_m(\mathbf{x}_{n,c})$. Consequently, the vector $\vec{f}_m \in \mathbb{R}^{N \times C}$ is obtained such that $\vec{f}_m \stackrel{\text{def}}{=} (f_{1,1}^m, f_{1,2}^m, \ldots, f_{1,C}^m, f_{2,1}^m, f_{2,2}^m, \ldots, f_{2,C}^m, \ldots, f_{N,C}^m)$.

The first step of the Mapper algorithm is to consider $\mathcal{F} \stackrel{\text{def}}{=} \{\vec{f}_m\}_{m=1}^M$, the transformed data points from $\mathcal{X}$. Then define $\psi : \mathbb{R}^{N \times C} \to \mathbb{R}^Z$, with $0 < Z \ll N \times C$ and consider the set $\mathcal{Z} \stackrel{\text{def}}{=} \{\psi(\vec{f}) | \vec{f} \in \mathcal{F}\}$. This dimensionality reduction ($N \times C \to Z$)

is employed to reduce the computational cost of the rest of the Mapper algorithm and can be considered as a hyperparameter of the Mapper algorithm.

In the second step of the algorithm, define $\sigma : \mathbb{R}^Z \to \mathbb{R}^W$, with $0 < W \ll Z$ and consider the set $\mathcal{W} \stackrel{\text{def}}{=} \{\sigma(\vec{z}) | \vec{z} \in \mathcal{Z}\}$. In the literature (Singh et al., 2007), the function σ is called filter function and $\mathcal{W}$ is the image or lens.

Third, let $\mathfrak{C}$ be the smallest hypercube of $\mathbb{R}^W$ which covers $\mathcal{W}$ entirely. As $\mathcal{X}$ is a finite set, each dimension of $\mathfrak{C}$ is a finite interval. Let $k \in \mathbb{N}^*$, be a hyperparameter that subdivides $\mathfrak{C}$ evenly into k^W smaller hypercubes. Note that the side lengths of these smaller hypercubes are $H = \frac{1}{k} \times$ the length size of $\mathfrak{C}$. Denotes $\mathcal{V}$ the set of all vertices of these smaller hypercubes. Next, fix $D > H$ as another hyperparameter. For each $\vec{v} \in \mathcal{V}$, consider the hypercube $c_{\vec{v}}$ of length D centered on $\vec{v}$. A visualization of step 3 is given in **Figure 5**.

Fourth, define $\mathcal{Z}_{\vec{v}} \stackrel{\text{def}}{=} \{\vec{z} \in \mathcal{Z} | \sigma(\vec{z}) \in c_{\vec{v}}\}$, the set of all elements of $\mathcal{Z}$ that is projected in the hypercube $c_{\vec{v}}$. Let ξ be a clustering algorithm and $\xi(\mathcal{Z}_{\vec{v}})$ be the resulting set of clusters. Define $\mathcal{B}$ as the set that consist of all so obtained clusters for all $\mathcal{Z}_{\vec{v}}$.

Fifth, compute the topological graph $\mathcal{G}$ using each element of $\mathcal{B}$ as a vertex and create an edge between vertices $\mathcal{G}_i$ and $\mathcal{G}_j$ ($i, j \in \{1, \ldots, |\mathcal{B}|\}, i \neq j$) if $\mathcal{G}_i \cap \mathcal{G}_j \neq \emptyset$.

2.4.3. Mapper Implementation Within This Work
In this work, as described in section 2.1 the dataset was recorded using the 3DC Armband which offers 10 channel-recording ($C = 10$) and an example is comprised of 151 data-points ($S = 151$) for each channel. The number of considered features in scenarios (A), (B), and (C), are 79, 384, and 465, respectively. Note that multi-output feature extraction techniques (e.g., AR, HIST), consider each component of that vector as a separate feature. Each element of $\mathcal{F}$ is obtained by computing the result of a feature from section 2.2 (corresponding to $\phi_m()$ in the mathematical definition given previously) over each channel of each example of the *Training Dataset*. The dataset undergoes the first dimensionality reduction ($\Psi()$) using PCA (Wold et al., 1987), where the number of principal components used corresponds to 99% of the total variance. For scenarios (A), (B), and (C), 99% of the variance resulted in 44, 77, and 119 components, respectively, extracted from 971,860 channel-wise examples.

A second dimensionality reduction is then performed ($\sigma()$), referred to as the filter function, with the goal of representing meaningful characteristics of the relationship between features (Singh et al., 2007). Within this study, t-Stochastic Neighborhood Embedding (t-SNE) (Maaten and Hinton, 2008) is used to encapsulate important local structure between features. The two-dimensional (2D) t-SNE lens was constructed with a perplexity of 30, as this configuration resulted in the most stable visualization over many repetitions [tested on scenario (A)]. Using t-SNE as part of the Mapper algorithm instead of on its own leverages its ability to represent local structure while avoiding the use of a low-dimensional manifold to encapsulate global structure. Instead, the global structure
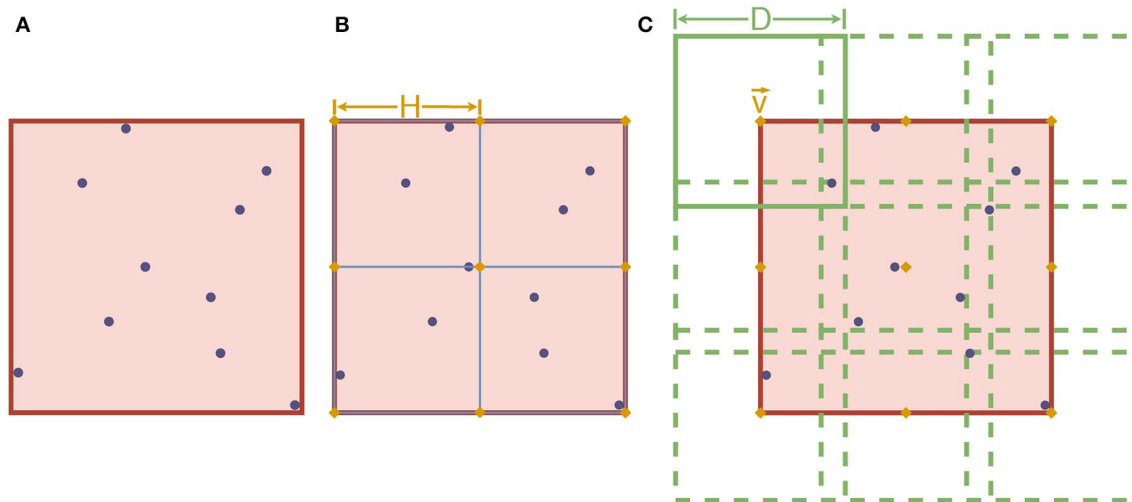
FIGURE 5 | An example of step 3 of the Mapper algorithm with $W = 2$. The purple dots represent the elements of $\mathcal{W}$. In **(A)**, the red square corresponds to $\mathfrak{C}$. In **(B)**, $\mathfrak{C}$ is subdivided using k^2 squares of length H (with $k = 2$ in this case). The orange diamonds, in both **(B,C)**, represent the elements of $\mathcal{V}$. Finally, the square $c_{\vec{v}}$ of length D is shown on the upper left corner of **(C)**, overlapping other squares centered on other elements of $\mathcal{V}$ (dotted lines).

is predominantly incorporated into the topological network produced by Mapper during the fifth stage.

The 2D lens was then segmented into a set of overlapped bins (the hypercubes centered on the elements of $\mathcal{V}$), called the cover. A stable topological network was obtained when each dimension was divided into 5 regions, forming a grid of 25 cubes that were overlapped by 65%. The number of regions correspond to the topological network's resolution, while the overlap has an influence on the amount of connection formed between nodes (Singh et al., 2007).

Data points in each region are then clustered in isolation to provide insight into the local structure of the feature space (the elements of $\mathcal{Z}_{\vec{v}}$ correspond to the data-point of a specific region). For each region, Ward's hierarchical clustering (ξ) was applied to construct a dendogram that grouped similar features together according to a reduction in cluster variance (Ward, 1963).

Finally, the dendograms produced using neighboring regions are combined to form the topological network ($\mathcal{G}$) using the features that lie in the overlapped area to construct the edges between the nodes.

The implementation of the Mapper algorithm was facilitated by a combination of the Kepler Mapper (van Veen and Saul, 2019) and the DyNeuSR (Dynamical Neuroimaging Spatiotemporal Representations) (Geniesse et al., 2019) Python modules. An extended coverage of processing pipelines for time-series TDA is given in Phinyomark et al. (2018).

3. RESULTS

3.1. Handcrafted Features

Figure 6 shows the topological network produced using only the handcrafted features. The Kullback-Leibler divergence of the t-SNE embedding of the handcrafted features plateaued at 0.50, indicating that the perplexity and number of iterations used was

appropriate for the dataset. The topological network consisted of 125 nodes and 524 edges.

The color of the nodes within the network indicates the percentage of members that belong to the feature group of interest [(A):SAP, (B): NLC, (C): FI, (D): TSM, and (E): UNI]. The presence of an edge symbolizes common features present in the connected nodes, which can be used at a global scale to verify that functional groups (similar information) cluster together. Due to the topological nature of the graph, information similarity between nodes is measured using the number of edges that separate two nodes and not the length of the edges. Detailed interpretation of the TDA networks are given in the discussion.

3.2. Deep Features

The average cross-subject accuracy on the test set when using the proposed ADANN framework was $84.43 \pm 0.05\%$. Using a Wilcoxon signed-rank test (Wilcoxon, 1992) with $n = 22$, and considering each participant as a separate dataset, this was found to significantly outperform ($p < 0.0001$) the average accuracy of $65.03 \pm 0.08\%$ obtained when training the ConvNet conventionally. Furthermore, based on Cohen's d, this difference in accuracy was considered to be huge (Sawilowsky, 2009). The accuracy obtained per participant for each training method is given in **Figure 7A**, and the confusion matrices calculated on the gestures are shown in **Figure 7B**.

Figure 8A provides visualizations of the ConvNet trained with ADANN using Guided Grad-CAM for several examples from the *3DC Dataset*, These visualizations highlight what the network considers "important" (i.e., which part of the signals had the most impact in predicting a given class) for the prediction of a particular gesture.

Instead of using Guided Grad-CAM to visualize how the network arrived at a decision for a known gesture, **Figure 8B** presents the results of the visualization algorithm when the

FIGURE 6 | Topological network generated exclusively for the handcrafted features, where nodes are colored to indicate percent composition of: **(A)** signal amplitude and power features (SAP), **(B)** non-linear complexity (NLC), **(C)** frequency information features (FI), **(D)** time series modeling features (TSM), and **(E)** unique features (UNI). Dashed boxes highlight dense groupings of the specified functional group in each of the networks.

network is told to find a gesture that is not present in the input. This is akin to using a picture of a cat as an input to the network and displaying the parts of the image that most resemble a giraffe. In **Figure 8B**, the input was randomly generated from a Gaussian distribution of mean 0 and standard deviation of 450 (chosen to have the same scale as the EMG signals of the *3DC Dataset*). For six of the eleven gestures (Radial Deviation, Wrist

Extension, Supination, Open Hand, Chuck Grip, and Pinch Grip) the network correctly identifies no relevant areas pertaining to these classes. While the network does highlight features in the input space associated with the other gestures, the magnitude of these contributions was substantially smaller (half or less) than when the requested gesture was actually present in the input signal.

FIGURE 7 | Classification results of deep learning architectures. **(A)** Per-participant test set accuracy comparison when training the network with and without ADANN, **(B)** Confusion matrices on the test set for cross-subject training with and without ADANN.
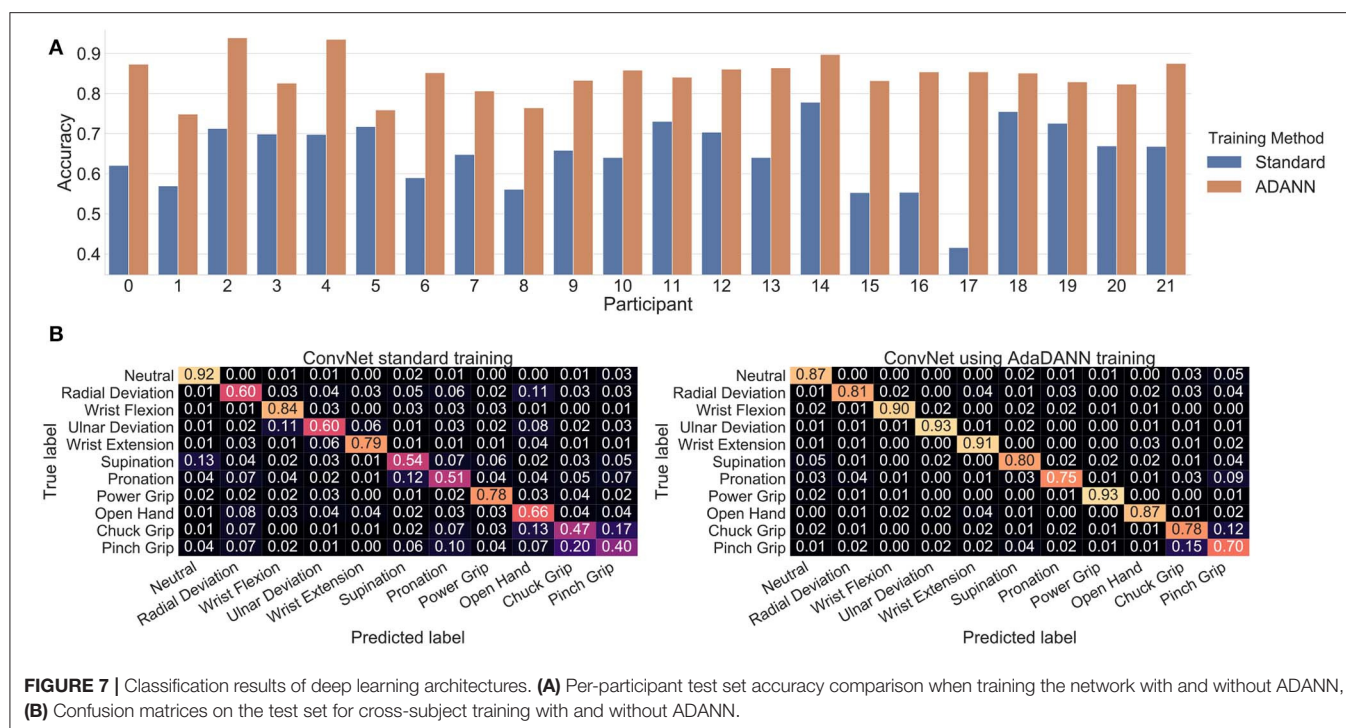
The topological network produced using only the learned features is given in **Figure 9**. The color of the nodes within the network indicates the percentage of members that belong to the feature group of interests [(A): B1, (B): B2, (C): B3, (D): B4, (E): B5, and (F): B6]. Interpretation of the TDA network follows the rational stated in section 3.1. The Kullback-Leibler divergence of the t-SNE embedding of the handcrafted features plateaued at 0.37, again indicating that the perplexity and number of iterations used was appropriate for the dataset. The topological network consisted of 115 nodes and 672 edges.

3.3. Hybrid Features
The topological network produced using both handcrafted and learned features is shown in **Figure 10**. The Kullback-Leibler divergence of the t-SNE embedding of all features plateaued at 0.53, again indicating that the perplexity and number of iterations used was appropriate for the dataset. The topological network consisted of 115 nodes and 770 edges. From this network, only a subset of nodes were occupied by both handcrafted and learned features. Those nodes were indicated in **Figure 10**.

The color of the nodes within the network indicates the percentage of members that belong to the feature group of interests (learned features). Information similarity was shown through a zoomed-in region of the network, where learned and handcrafted features clustered together. The feature members of the numbered nodes were listed in **Table 2**. Interpretation of the TDA network follows the rational stated in section 3.1.

Table 3 shows the average accuracy (grouped by block for the learned features and by group for the handcrafted features) obtained when training an LDA on each feature and when using all features within a category (i.e., within a block or

within a group of handcrafted feature). Note that for the learned features, PCA is applied to the feature map and the first component is employed to represent a given learned feature. **Figure 11** shows examples of confusion matrices computed from the LDA classifications of singular features (both handcrafted and learned). **Figure 11**, also shows some confusion matrices obtained from the LDA's classification result when using all features within a category.

Figure 12 shows the average mean square error computed when regressing from the ConvNet's learned features (see section 2.3.5) to fifteen handcrafted features (three per Functional Group). Note that the mean squared error is obtained by computing the regression using only the output of the block of interest.

4. DISCUSSION
4.1. Handcrafted Features
The result of the Mapper algorithm applied to handcrafted features (see **Figure 6**) showed that the handcrafted features agglomerated mostly with their respective groups, and that the topological graph is Y-shaped. This shows that the hyperparameters selected in this work are consistent with those found in previous EMG literature (Phinyomark et al., 2018; Campbell et al., 2019a).

4.2. ADANN and Deep Learning Visualization
Figure 7B shows that training the network with ADANN outperforms the standard training method in cross-subject classification. One advantage of ADANN in the context of this
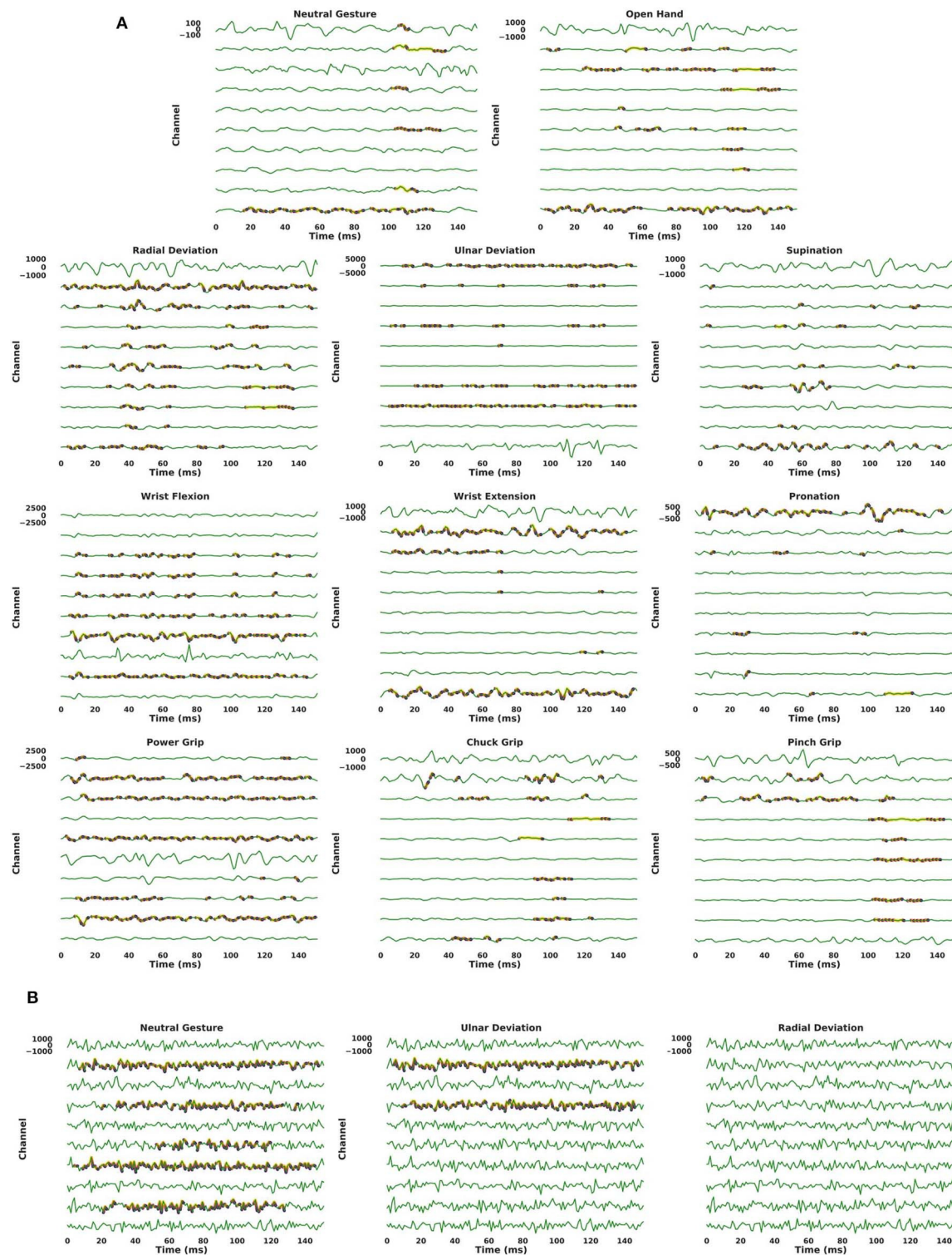
FIGURE 8 | Output of Guided Grad-CAM when asked to highlight specific gestures in an example. For all graphs, the y-axis of each channel are scaled to the same range of value (indicated on the first channel of each graph). Warmer colors indicate a higher "importance" of a feature in the input space for the requested gesture. The coloring use a logarithmic scale. For visualization purposes, only features that are within three order of magnitudes to the most contributing feature are colored. **(A)** The examples shown are real examples and correspond to the same gestures that Guided Grad-CAM is asked to highlight. **(B)** A single example, generated using Gaussian noise of mean 0 and standard deviation 450, is shown three times. While the visualization algorithm does highlight features in the input space (when the requested gesture is not truly present in the input), the magnitude of these contributions is substantially smaller (half or less) than when the requested gesture is present in the input.
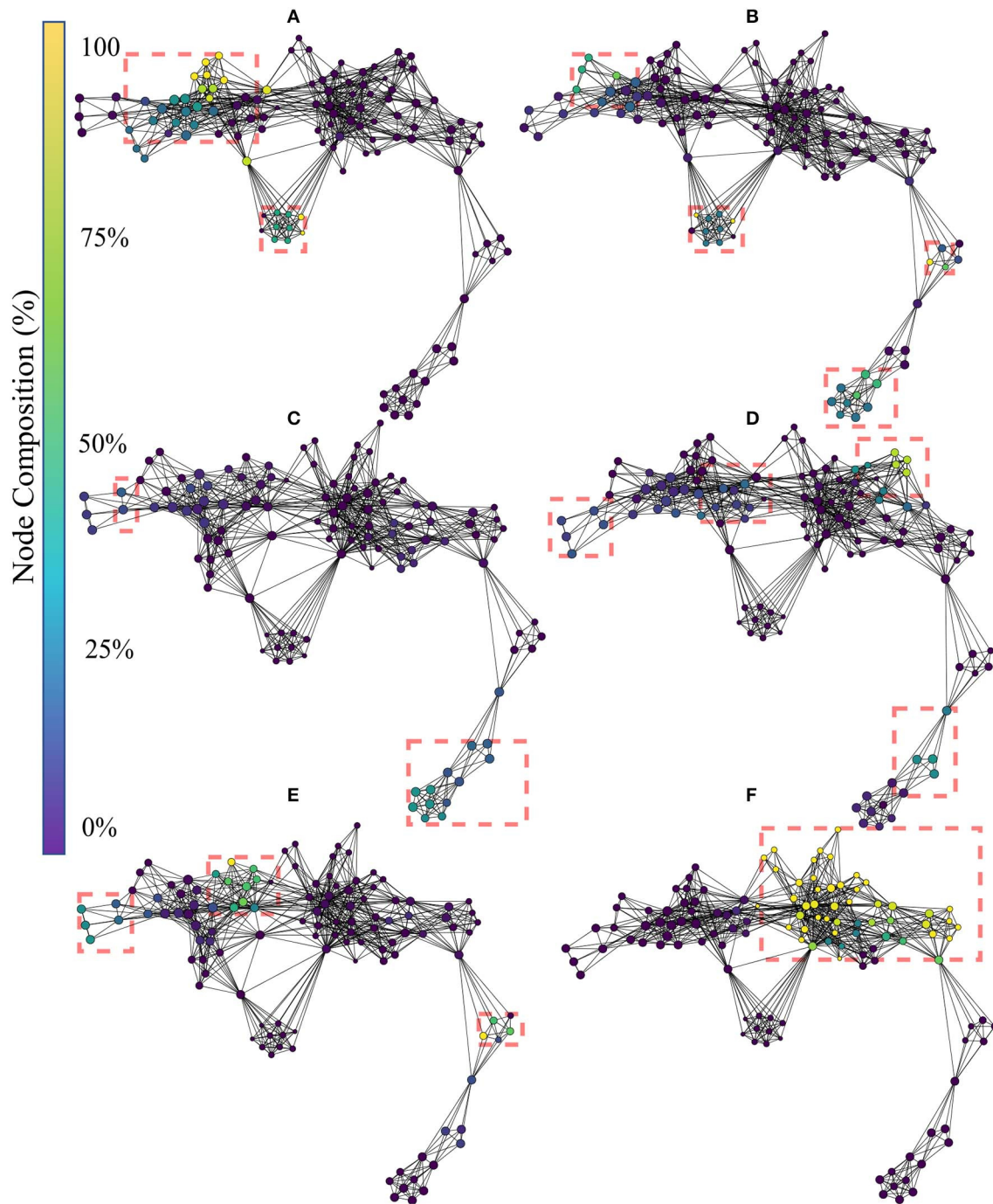
FIGURE 9 | Topological network generated for exclusively the learned features, where nodes are colored to indicate percent composition of: **(A)** Block 1's features, **(B)** Block 2's features, **(C)** Block 3's features, **(D)** Block 4's features, **(E)** Block 5's features, and **(F)** Block 6's features. Dashed boxes highlight dense groupings of the specified block features in each of the networks.

work is that the weights of the network have strong incentives to be subject-agnostic. As such, the learned features extracted from the network can be thought of as general features (and to a certain extent subject-independent) for the task of sEMG-based hand gesture recognition.

Applying Guided Grad-CAM, as in **Figure 8**, shows that the network mostly focuses on different channels for the detection of antagonist gestures. This suggests that the ConvNet was able to extract spatial features despite having access only to one dimensional convolutional kernels. Furthermore, it is notable
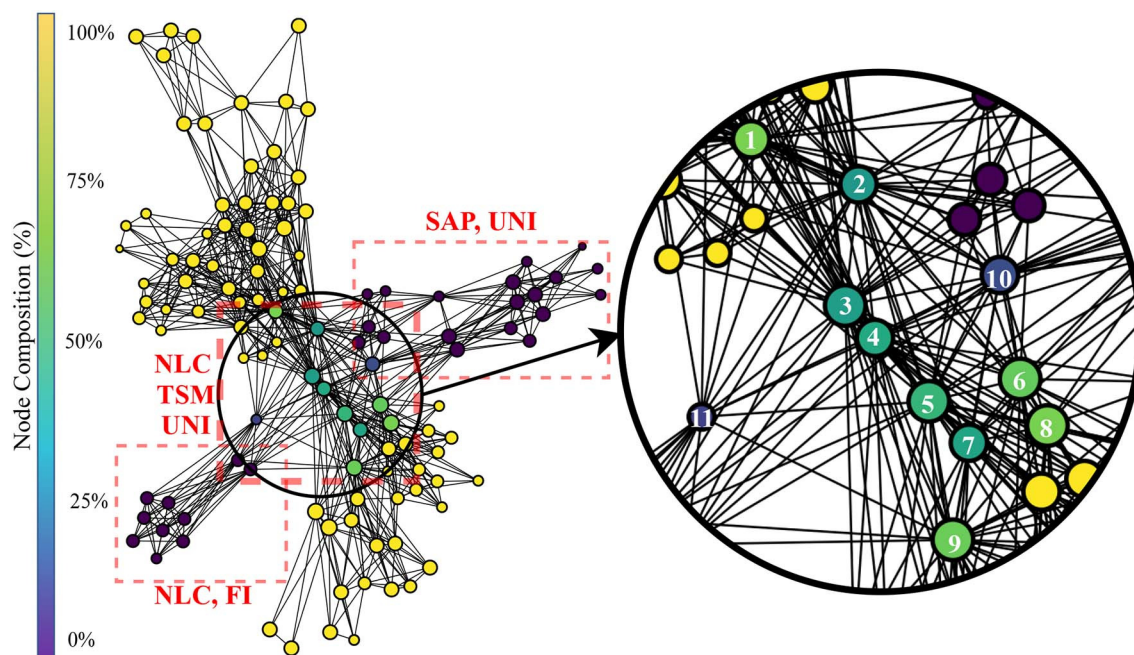
FIGURE 10 | Topological network generated for all features, where nodes were colored to indicate percent composition of learned features. The dashed boxes highlight dense grouping of handcrafted features with their associated type.

TABLE 2 | Members of nodes labeled in **Figure 6**. LeF**X** refers to a *Learned Feature* from block **X**.

#	Summary	Members
1	TSM+LeF5	AR2 AR4 DAR2 DAR4 CC1 CC4 DCC1
		DCC3 SNR 8xLeF1 1xLeF2 4xLeF4 10xLeF5 13xLe5
2	TSM+UNI+LeF6	APEN AR2 AR4 DAR2 DAR4 CC1 CC4 DCC1 DCC3 DCC4 CE DFA DPR HIST123
		SKEW MAVS OHM PSDFD PSR SMR SNR VCF VFD 1xLeF1 3xLeF2 3xLeF5 21xLeF6
3	TSM+UNI+LeF6	APEN AR2 AR4 DAR2 DAR4 CC1 CC4 DCC1 DCC3 DCC4 CE DFA DPR HIST12
		SKEW MAVS OHM PSDFD PSR SMR SNR VCF VFD 1xLeF1 1xLeF2 1xLeF5 27xLeF6
4	UNI+LeF6	APEN DCC4 CE DFA DPR HIST123
		SKEW MAVS OHM PSDFD PSR SMR VCF VFD 2xLeF2 2xLeF5 21xLeF6
2	TSM+UNI+LeF6	APEN CC1 CC4 DCC4 CE DFA DPR HIST123
		SKEW MAVS OHM PSDFD PSR SMR SNR VCF VFD 37xLeF6
6	TSM+UNI+LeF6	CC1 CC4 DCC4 CE DPR HIST123 SKEW MAVS PSDFD SMR
		SNR VCF VFD 5xLeF2 5xLeF4 1xLeF5 37xLeF6
7	UNI+LeF6	DCC4 CE DPR HIST123 SKEW MAVS
		PSDFD SMR VCF VFD 2xLeF2 15xLeF6
8	UNI+LeF6	DCC4 CE DPR HIST123 SKEW MAVS PSDFD SMR
		VCF VFD 5xLeF2 5xLeF4 1xLeF5 37xLeF6
9	UNI+LeF6	APEN DCC4 CE DFA DPR HIST2 SKEW MAVS
		OHM PSDFD PSR SMR VCF VFD 15xLeF2 36xLeF6
10	All Handcrafted+LeF6	APEN CC14 DCC4 CE DFA DPR HIST123 KURT SKEW M2 MAVS MAX MHW23
		MTW123 MNP TTP OHM PSDFD PSR SM SMR SNR SSI TM DTM VAR DVARV VCF VFD 11xLeF6
11	NLC+LeF6	APEN SAMPEN BC
		KATZ 1xLeF6

that for all the examples given in **Figure 8A**, the most active channel was not the primary channel used for the gesture prediction. In fact, for the vast majority of gestures, the channel with the highest amplitude did not contribute in a meaningful way to the network's prediction. This observation held true while looking at several other examples from the

TABLE 3 | Accuracy obtained on the test set using the handcrafted features and the learned features from their respective block.

	Single feature		All features
	Average accuracy (%)	STD (%)	Accuracy (%)
SAP	26.80	7.0	41.61
FI	19.95	2.87	34.80
NLC	22.32	7.15	31.49
TSM	22.24	3.33	37.18
UNI	15.32	5.11	48.37
Block 1	**28.49**	**3.84**	74.59
Block 2	28.28	4.66	78.26
Block 3	28.90	5.06	79.19
Block 4	29.21	5.15	78.77
Block 5	28.18	5.48	79.23
Block 6	26.62	6.19	**81.38**

The Single Feature accuracies are given as the average accuracy over all the features of their respective block/category.

3DC Dataset. This might indicate that the common practice of placing the recording channel directly on the most prominent muscle for a given gesture within the context of gesture recognition may not be optimal. One could thus use the type of information provided by algorithms, such as Guided Grad-CAM as another way of performing channel selection (instead of simply using classification accuracy). The absence of importance on amplitude characteristics is in contrast to conventional practices of handcrafted feature engineering—where the feature set typically relies heavily on amplitude characteristics. This perhaps explains the growing interest in handcrafted feature extraction techniques that do not capture amplitude information, such as TDPSD, that have been demonstrated to outperform conventional amplitude-reliant features in terms of accuracy and robustness to confounding factors (Khushaba et al., 2016).

When applying Guided Grad-CAM on a noise input (one where the target gesture is not present, as seen in **Figure 8B**), the reported activation level is substantially lower, and in some cases non-existent. When the standard deviation of the Gaussian noise was increased by 33%, the network did not find any features resembling any gesture. This is most likely due to the fact that increasing the spread of the noise leads to a potentially greater gap in value between two adjacent data-points (reduced smoothness) fostering the condition for a more unrealistic signal. One could thus imagine training a generative adversarial network with the discriminative function based on the activation level calculated by Guided Grad-CAM, and modulating the difficulty by augmenting the signal's amplitude. This could facilitate training a network to not only be able to generate realistic, synthetic EMG signal, but also have the signal resemble actual gestures.

In contrast to the topological networks based on handcrafted features, those based on the learned features appear as a long flair with a loop. From **Figure 9A**, the learned features from block 1 are concentrated in the left segment of the flare, and the lower segment of the loop. From **Figure 9B**, the learned features

from block 2 were located slightly more central to the network than the block 1 features. Additionally, a small subset of block 2 features appeared at the right segment of the flare, indicating a second distinct source of information was being harnessed. From **Figures 9C–E**, the features of block 3, 4, and 5 relocate their concentration of features to converge in the center of the network. Finally from **Figure 9F**, the concentration of all block 6 features lies in the center of the network. Thus, it can be seen that learned features from the same block tend to cluster together and remain close in the map to adjacent blocks in the network. The only exception to this is from the first block to the second, where substantially different features were generated by the latter. This suggests that the first layer may serve almost as a preprocessing layer which conditions the signal for the other layers.

4.3. Hybrid Features Visualization

The topological network generated from using both the handcrafted and learned features (see **Figure 10**) followed two orthogonal axes with the handcrafted features on one and the learned features on the other. The middle of the graph (where the two axis intercept) is where any nodes containing both handcrafted and learned features are found. The vast majority of these nodes are populated by features from block 6 and the NLC, TSM and UNI functional groupings. No nodes in the graph contained both handcrafted features and features from block 3, suggesting that block 3 extracted features not captured by current feature designs. Conversely, no learned features shared a node with features from the FI family, suggesting that these features may not have been extracted by the network.

While this topological network informs the type of information encoded within each individual feature, it is important to note that information can still be present but encoded in a more complex way within the weights of the deep network. This information flow can be visualized from the regression graphs of **Figure 12**. Features from the SAP family are more easily predicted within the early blocks whereas features from the TSM and NLC family require the latter blocks of the network to achieve the best predictions. Interestingly, while features from the FI family did not share any nodes learned features, one can see that the deep network is able to better extract this type of information within the intermediary blocks. This indicates (from **Figures 10**, **12**) that, while frequency information is not explicitly used by the ConvNet, this type of information is nonetheless indirectly used to compute the features from the latter blocks. An example of a feature for which the ConvNet was unable to leverage its topology is the HIST (see **Figure 12**).

4.4. Understanding Deep Features Predictions

The topological network of **Figure 10** showed that the type of information encoded within the lower blocks of the ConvNet tended to be highly dissimilar to what the handcrafted features encoded. Interestingly, however, **Figure 11** shows that the role fulfilled by these features is similar. That is, both the handcrafted and learned features (from the lower blocks) try to encode general properties that can distinguish between all classes.
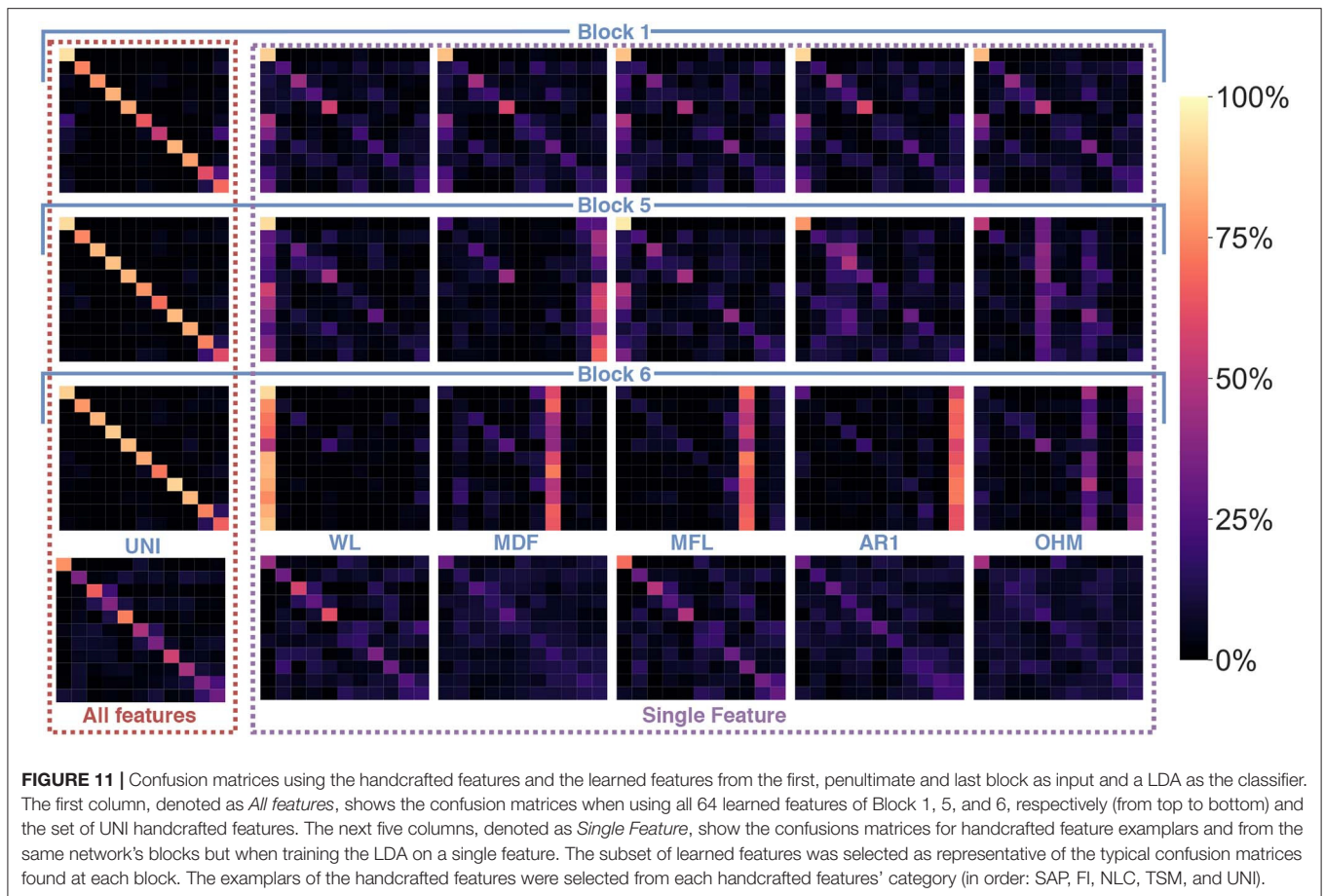
FIGURE 11 | Confusion matrices using the handcrafted features and the learned features from the first, penultimate and last block as input and a LDA as the classifier. The first column, denoted as *All features*, shows the confusion matrices when using all 64 learned features of Block 1, 5, and 6, respectively (from top to bottom) and the set of UNI handcrafted features. The next five columns, denoted as *Single Feature*, show the confusions matrices for handcrafted feature examplars and from the same network's blocks but when training the LDA on a single feature. The subset of learned features was selected as representative of the typical confusion matrices found at each block. The examplars of the handcrafted features were selected from each handcrafted features' category (in order: SAP, FI, NLC, TSM, and UNI).

The confusion matrices obtained from training an LDA on a single feature highlight this behavior (see **Figure 11** for some examples) as both the handcrafted features and the learned features (before the last block) are able to distinguish between gestures relatively equally. In contrast, the features extracted from the last block (and to a lesser extent from the penultimate block) have been optimized to be a gesture detector instead of a feature detector. A clear visual of this behavior is illustrated in **Figure 11**, where the main line highlighted in the confusion matrices from block 6 was a single column (corresponding to the prediction of a single gesture), instead of the typical diagonal. In other words, during training, the neurons of the final block are encoded to have maximum activation when a particular class was provided in the input window and minimum activation when other classes were provided; effectively creating a one-vs.-all (OVA) classifier. This behavior is consistent with the feature visualization literature found in image classification and natural language processing, where semantic dictionaries or saliency maps have depicted neuron representations becoming more abstract at later layers (Simonyan et al., 2013; LeCun et al., 2015). This also explains why the features from the last block obtained the worst average accuracy when taken individually while achieving the highest accuracy as a group (see **Table 3**). That is, as each feature map of the last layer tries to detect a particular gesture, its activation for the other gestures should

be minimal, making the distinction between the other gestures significantly harder. The final decision layer of the network can then be thought of as a weighted average of these OVA classifiers to maximize the performance of the learned feature maps. Note that in **Table 3**, the lower accuracies obtained from the handcrafted features as a group were expected as each feature within the same family provides similar type of information, even more so than the learned features of the network (as seen in **Figures 6**, **9**, **10**). Overall, the best performing handcrafted feature set as a group was the features from the UNI family despite the fact that they were the worst on average when alone. This is most likely due to the fact that by definitions, features within this family are more heterogeneous.

5. CONCLUSION

This paper presents the first in-depth analysis of features learned using deep learning for EMG-based hand gesture recognition. The type of information encoded within learned features and their relationship to handcrafted features were characterized employing a mixture of topological data analysis (Mapper), network interpretability visualization (Guided Grad-CAM), machine learning (feature classification prediction), and by visualizing the information flow using feature regression.
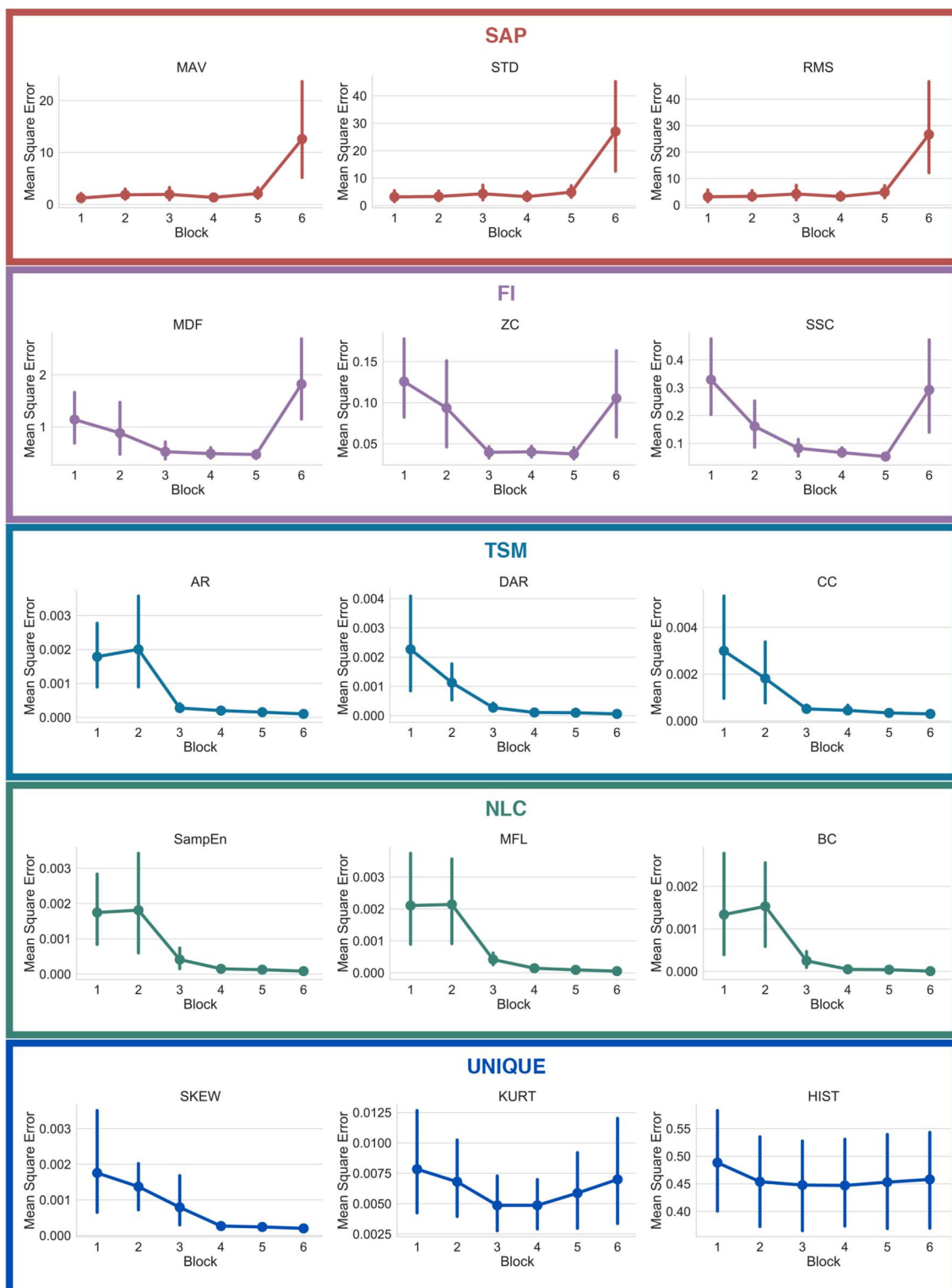
FIGURE 12 | Mean squared error of the regressions from learned features to handcrafted features, with respect to the number of blocks employed for the regression. The features are grouped with their respective functional groups.

As a secondary, but significant contribution, this work presented ADANN, a novel multi-domain training algorithm particularly suited for EMG-based gesture recognition shown to significantly outperform traditional training on cross-subject classification accuracy.

This manuscript paves the way for hybrid classifiers that contain both learned and handcrafted features. An ideal application for the findings of this work would rely on a mix of handcrafted features and learned features taken from all four extremities of the hybrid topological network, and at the center to provide complementary, and general features to the classifier. A network could then be trained to augment its sensitivity to similar classes. For example, to alleviate ambiguity between pinch grip and chuck grip, a learned feature that encodes the one-vs.-all information of pinch grip could be included into the original feature set or into an otherwise handcrafted only feature set. Alternatively, handcrafted feature extraction stages may be installed within the deep learning architecture by means of neuroevolution of augmenting topologies (Chen and Alahakoon, 2006), a genetic algorithm that optimizes the weights and connections of deep learning architectures.

The main limitation of this study was the use of a single architecture to generate the learned features. Though this architecture was chosen to be representative of current practices in myoelectric control and be extensible to other applications, the current work did study the impact of varying the number of blocks and the composition of these block on the different experiments. Additionally, although the set of handcrafted features was selected to be comprehensive over the sources of information available from the EMG signal, explicit time-frequency features, such as those based on spectrograms and wavelet were not included in the current work, as they were ill-adapted to the framework employed in this study. Furthermore, an analysis including a larger amount of gestures should also be conducted. Importantly, these results are presented for a single 1D electrode array, and may not be representative of larger 2D arrays, such as those used in high density EMG applications. Similarly, explicit spatio-temporal features, such as coherence between electrodes, were not explored, and the convolutional kernels were restricted to 1D (although as seen in **Figure 8A** the network was still able to learn spatial information to a certain extent). Omitting these type of complex features was a design choice as this work represents a first step in understanding and characterizing learned features within the context of EMG signal. As such, using this manuscript as a basis, future works should study the impact of diverse architectures on the type of learned features and will incorporate spatio-temporal features (both handcrafted and from 2D convolutional kernels). Additionally, formal feature set generation and hybrid classifiers should be investigated using the tools presented in this work.

DATA AVAILABILITY STATEMENT

The datasets analyzed and the source code for this study can be found at the following link: https://github.com/UlysseCoteAllard/sEMG_handCraftedVs LearnedFeatures.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comités d'Éthique de la Recherche avec desêtres humains de l'Université Laval. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

UC-A, EC, AP, FL, BG, and ES: conceptualization. UC-A, EC, AP, FL, and ES: methodology. UC-A and EC: software and validation. UC-A, EC, AP, FL, and ES: formal analysis and investigation. FL, BG, and ES: resources. UC-A and EC: data curation and writing original draft preparation. UC-A, EC, AP, FL, BG, and ES: writing review and editing. UC-A and EC: visualization. FL, BG, and ES: supervision. UC-A, EC, AP, FL, BG, and ES: project administration. UC-A, AP, FL, BG, and ES: funding acquisition.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. (2014). Domain-adversarial neural networks. *arXiv [Preprint] arxiv*:1412.4446.

Allard, U. C., Nougarou, F., Fall, C. L., Giguère, P., Gosselin, C., Laviolette, F., et al. (2016). "A convolutional neural network for robotic arm guidance using SEMG based frequency-features," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Seoul: IEEE), 2464–2470.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., et al. (2018). The history began from alexnet: a comprehensive survey on deep learning approaches. *arXiv [Preprint] arxiv*:1803.01164.

Al-Timemy, A. H., Khushaba, R. N., Bugmann, G., and Escudero, J. (2015). Improving the performance against force variation of EMG controlled multifunctional upper-limb prostheses for transradial amputees. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 650–661. doi: 10.1109/TNSRE.2015. 2445634

Andersen, V., Fimland, M. S., Mo, D.-A., Iversen, V. M., Vederhus, T., Hellebø, L. R. R., et al. (2018). Electromyographic comparison of barbell deadlift, hex bar deadlift, and hip thrust exercises: a cross-over study. *J. Strength Condit. Res.* 32, 587–593. doi: 10.1519/JSC.0000000000001826

Arjunan, S. P., and Kumar, D. K. (2010). Decoding subtle forearm flexions using fractal features of surface electromyogram from single and multiple sensors. *J. Neuroeng. Rehabil.* 7:53. doi: 10.1186/1743-0003-7-53

Atzori, M., Cognolato, M., and Müller, H. (2016). Deep learning with convolutional neural networks applied to electromyography data: a resource for the classification of movements for prosthetic hands. *Front. Neurorobot.* 10:9. doi: 10.3389/fnbot.2016.00009

Balasubramanian, M., and Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science* 295:7. doi: 10.1126/science.295.5552.7a

Batchvarov, V., and Malik, M. (2002). Individual patterns of QT/RR relationship. *Cardiac Electrophysiol. Rev.* 6, 282–288. doi: 10.1023/A:1016393328485

Côté-Allard, U., Fall, C. L., Drouin, A., Campeau-Lecours, A., Gosselin, C., Glette, K., et al. (2019a). Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 760–771. doi: 10.1109/TNSRE.2019.2896269

Côté-Allard, U., Gagnon-Turcotte, G., Laviolette, F., and Gosselin, B. (2019b). A low-cost, wireless, 3-D-printed custom armband for semg hand gesture recognition. *Sensors* 19:2811. doi: 10.3390/s19122811

Campbell, E., Phinyomark, A., Al-Timemy, A. H., Khushaba, R. N., Petri, G., and Scheme, E. (2019a). "Differences in EMG feature space between able-bodied and amputee subjects for myoelectric control," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)* (San Fransisco, CA), 33–36.

Campbell, E., Phinyomark, A., and Scheme, E. (2019b). Feature extraction and selection for pain recognition using peripheral physiological signals. *Front. Neurosci.* 13:437. doi: 10.3389/fnins.2019.00437

Campbell, E. D., Phinyomark, A., and Scheme, E. (2019c). "Linear discriminant analysis with bayesian risk parameters for myoelectric control," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (GlobalSIP 2019)* (Ottawa, ON).

Castellini, C., Fiorilla, A. E., and Sandini, G. (2009). Multi-subject/daily-life activity EMG-based control of mechanical hands. *J. Neuroeng. Rehabil.* 6:41. doi: 10.1186/1743-0003-6-41

Chen, H., Zhang, Y., Li, G., Fang, Y., and Liu, H. (2019). Surface electromyography feature extraction via convolutional neural network. *Int. J. Mach. Learn. Cybernet.* 11, 185–196. doi: 10.1007/s13042-019-00966-x

Chen, L., and Alahakoon, D. (2006). "Neuroevolution of augmenting topologies with learning for data classification," in *2006 International Conference on Information and Automation* (Shandong), 367–371.

Choi, H., and Choi, S. (2007). Robust kernel isomap. *Pattern Recogn.* 40, 853–862. doi: 10.1016/j.patcog.2006.04.025

Cote-Allard, U., Fall, C. L., Campeau-Lecours, A., Gosselin, C., Laviolette, F., and Gosselin, B. (2017). "Transfer learning for semg hand gestures recognition using convolutional neural networks," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (IEEE), 1663–1668.

Du, S., and Vuskovic, M. (2004). "Temporal vs. spectral approach to feature extraction from prehensile emg signals," in *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004* (Las Vegas, NV: IEEE), 344–350.

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Gan, C., Wang, N., Yang, Y., Yeung, D.-Y., and Hauptmann, A. G. (2015). "Devnet: a deep event network for multimedia event detection and evidence recounting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2568–2577.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. arxiv:1505.07818.

Geniesse, C., Sporns, O., Petri, G., and Saggar, M. (2019). Generating dynamical neuroimaging spatiotemporal representations (dyneusr) using topological data analysis. *Netw. Neurosci.* 3, 763–778. doi: 10.1162/netn_a_00093

Gitter, J. A., and Czerniecki, M. J. (1995). Fractal analysis of the electromyographic interference pattern. *J. Neurosci. Methods* 58, 103–108. doi: 10.1016/0165-0270(94)00164-C

Guidetti, L., Rivellini, G., and Figura, F. (1996). EMG patterns during running: intra-and inter-individual variability. *J. Electromyogr. Kinesiol.* 6, 37–48. doi: 10.1016/1050-6411(95)00015-1

Gupta, V., Suryanarayanan, S., and Reddy, N. P. (1997). Fractal analysis of surface EMG signals from the biceps. *Int. J. Med. Inform.* 45, 185–192. doi: 10.1016/S1386-5056(97)00029-4

Halaki, M., and Ginn, K. (2012). "Normalization of EMG signals: to normalize or not to normalize and what to normalize to?" in *Computational Intelligence in Electromyography Analysis–A Perspective on Current Applications and Future Challenges* ed G. R. Naik (London: IntechOpen), 175–194.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 4700–4708.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv [Preprint] arxiv:*1502.03167.

Khushaba, R. N., Al-Timemy, A., Kodagoda, S., and Nazarpour, K. (2016). Combined influence of forearm orientation and muscular contraction on emg pattern recognition. *Expert Syst. Appl.* 61, 154–161. doi: 10.1016/j.eswa.2016.05.031

Kim, K. S., Choi, H. H., Moon, C. S., and Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Curr. Appl. Phys.* 11, 740–745. doi: 10.1016/j.cap.2010.11.051

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arxiv [Preprint] arxiv:*1412.6980.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. (2016). Revisiting batch normalization for practical domain adaptation. *arxiv [Preprint] arxiv:*1603.04779.

Liu, X., Zhang, R., Meng, Z., Hong, R., and Liu, G. (2019). On fusing the latent deep CNN feature for image classification. *World Wide Web* 22, 423–436. doi: 10.1007/s11280-018-0600-3

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

McCool, P., Fraser, G. D., Chan, A. D., Petropoulakis, L., and Soraghan, J. J. (2014). Identification of contaminant type in surface electromyography (EMG) signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* 22, 774–783. doi: 10.1109/TNSRE.2014.2299573

Meltzer, J. A., Negishi, M., Mayes, L. C., and Constable, R. T. (2007). Individual differences in EEG theta and alpha dynamics during working memory correlate with fMRI responses across subjects. *Clin. Neurophysiol.* 118, 2419–2436. doi: 10.1016/j.clinph.2007.07.023

Nam, H., and Han, B. (2016). "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 4293–4302.

Nanni, L., Ghidoni, S., and Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recogn.* 71, 158–172. doi: 10.1016/j.patcog.2017.05.025

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7265–7270. doi: 10.1073/pnas.1102826108

Oskoei, M. A., and Hu, H. (2006). "Ga-based feature subset selection for myoelectric classification," in *2006 IEEE International Conference on Robotics and Biomimetics* (Kunming: IEEE), 1465–1470.

Oskoei, M. A., and Hu, H. (2007). Myoelectric control systems–a survey. *Biomed. Signal Process. Control* 2, 275–294. doi: 10.1016/j.bspc.2007.07.009

Oskoei, M. A., and Hu, H. (2008). Support vector machine-based classification scheme for myoelectric control applied to upper limb. *IEEE Trans. Biomed. Eng.* 55, 1956–1965. doi: 10.1109/TBME.2008.919734

Park, S.-H., and Lee, S.-P. (1998). EMG pattern recognition based on artificial intelligence techniques. *IEEE Trans. Rehabil. Eng.* 6, 400–405. doi: 10.1109/86.736154

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in pytorch," in *NIPS-W* (Long Beach, CA).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Phinyomark, A., Ibanez-Marcelo, E., and Petri, G. (2018). *Topological Data Analysis of Biomedical Big Data. 1st Edn.*, Volume 1 of 1, Chapter 11. Boca Raton: CRC Press, 209–233.

Phinyomark, A., Khushaba, R. N., Ibáñez-Marcelo, E., Patania, A., Scheme, E., and Petri, G. (2017). Navigating features: a topologically informed chart of electromyographic features space. *J. R. Soc. Interface* 14:20170734. doi: 10.1098/rsif.2017.0734

Phinyomark, A., Phothisonothai, M., Phukpattaranont, P., and Limsakul, C. (2011a). Critical exponent analysis applied to surface EMG signals for gesture recognition. *Metrol. Meas. Syst.* 18, 645–658. doi: 10.2478/v10178-011-0061-9

Phinyomark, A., Phothisonothai, M., Phukpattaranont, P., and Limsakul, C. (2011b). Evaluation of movement types and electrode positions for EMG pattern classification based on linear and non-linear features. *Eur. J. Sci. Res* 62, 24–34.

Phinyomark, A., Phukpattaranont, P., and Limsakul, C. (2011c). A review of control methods for electric power wheelchairs based on electromyography signals with special emphasis on pattern recognition. *IETE Tech. Rev.* 28, 316–326. doi: 10.4103/0256-4602.83552

Phinyomark, A., Phukpattaranont, P., and Limsakul, C. (2012a). Feature reduction and selection for EMG signal classification. *Expert Syst. Appl.* 39, 7420–7431. doi: 10.1016/j.eswa.2012.01.102

Phinyomark, A., Phukpattaranont, P., and Limsakul, C. (2012b). Fractal analysis features for weak and single-channel upper-limb EMG signals. *Expert Syst. Appl.* 39, 11156–11163. doi: 10.1016/j.eswa.2012.03.039

Phinyomark, A., Phukpattaranont, P., Limsakul, C., and Phothisonothai, M. (2011d). Electromyography (EMG) signal classification based on detrended fluctuation analysis. *Fluct. Noise Lett.* 10, 281–301. doi: 10.1142/S0219477511000570

Phinyomark, A., Quaine, F., Charbonnier, S., Serviere, C., Tarpin-Bernard, F., and Laurillau, Y. (2013). EMG feature evaluation for improving myoelectric pattern recognition robustness. *Expert Syst. Appl.* 40, 4832–4840. doi: 10.1016/j.eswa.2013.02.023

Phinyomark, A., and Scheme, E. (2018a). Emg pattern recognition in the era of big data and deep learning. *Big Data Cogn. Comput.* (Seoul) 2:21. doi: 10.3390/bdcc2030021

Phinyomark, A., and Scheme, E. (2018b). "A feature extraction issue for myoelectric control based on wearable EMG sensors," in *2018 IEEE Sensors Applications Symposium (SAS)* (IEEE), 1–6.

Poria, S., Cambria, E., and Gelbukh, A. (2015). "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon), 2539–2544.

Qingju, Z., and Zhizeng, L. (2006). "Wavelet de-noising of electromyography," in *2006 International Conference on Mechatronics and Automation* (Luoyang: IEEE), 1553–1558.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2018). "Efficient parametrization of multi-domain deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 8119–8127.

Saggar, M., Sporns, O., Gonzalez-Castillo, J., Bandettini, P., Carlsson, G., Glover, G., et al. (2018). Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nat. Commun.* 9:1399. doi: 10.1038/s41467-018-03664-4

Saponas, T. S., Tan, D. S., Morris, D., and Balakrishnan, R. (2008). "Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence: ACM), 515–524.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* 8:26. doi: 10.22237/jmasm/1257035100

Scheme, E., and Englehart, K. (2011). Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use. *J. Rehabil. Res. Dev.* 48, 643–659. doi: 10.1682/JRRD.2010.09.0177

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 618–626.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). Grad-CAM: why did you say that? *arxiv [Preprint] arxiv*:1611.07450.

Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 3626–3633.

Shu, R., Bui, H. H., Narui, H., and Ermon, S. (2018). A dirt-t approach to unsupervised domain adaptation. *arXiv [Preprint] arxiv*:1802.08735.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv [Preprint] arxiv*:1312.6034.

Sinderby, C., Lindstrom, L., and Grassino, A. (1995). Automatic assessment of electromyogram quality. *J. Appl. Physiol.* 79, 1803–1815. doi: 10.1152/jappl.1995.79.5.1803

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). "Topological methods for the analysis of high dimensional data sets and 3D object recognition," in *SPBG* (Prague), 91–100.

Smith, L. H., Hargrove, L. J., Lock, B. A., and Kuiken, T. A. (2010). Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay. *IEEE Trans. Neural Syst. Rehabil. Eng.* 19, 186–192. doi: 10.1109/TNSRE.2010.2100828

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: the all convolutional net. *arxiv [Preprint] arxiv*:1412.6806.

St-Onge, D., Côté-Allard, U., Glette, K., Gosselin, B., and Beltrame, G. (2019). Engaging with robotic swarms: commands from expressive motion. *ACM Trans. Hum. Robot Interact.* 8:11. doi: 10.1145/3323213

Talebinejad, M., Chan, A. D., Miri, A., and Dansereau, R. M. (2009). Fractal analysis of surface electromyography signals: a novel power spectrum-based method. *J. Electromyogr. Kinesiol.* 19, 840–850. doi: 10.1016/j.jelekin.2008.05.004

Thongpanja, S., Phinyomark, A., Hu, H., Limsakul, C., and Phukpattaranont, P. (2015). The effects of the force of contraction and elbow joint angle on mean and median frequency analysis for muscle fatigue evaluation. *ScienceAsia* 41, 263–263. doi: 10.2306/scienceasia1513-1874.2015.41.263

Thongpanja, S., Phinyomark, A., Phukpattaranont, P., and Limsakul, C. (2013). Mean and median frequency of EMG signal to determine muscle force based on time-dependent power spectrum. *Elektron. Elektrotech.* 19, 51–56. doi: 10.5755/j01.eee.19.3.3697

Thongpanja, S., Phinyomark, A., Quaine, F., Laurillau, Y., Limsakul, C., and Phukpattaranont, P. (2016). Probability density functions of stationary surface EMG signals in noisy environments. *IEEE Trans. Instr. Meas.* 65, 1547–1557. doi: 10.1109/TIM.2016.2534378

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7167–7176.

Van Den Broek, E. L., Schut, M. H., Westerink, J. H., van Herk, J., and Tuinenbreijer, K. (2006). "Computing emotion awareness through facial electromyography," in *European Conference on Computer Vision* (Graz: Springer), 52–63.

van Veen, H. J., and Saul, N. (2019). *Kepler Mapper*.

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Wilcoxon, F. (1992). "Individual comparisons by ranking methods," in *Breakthroughs in Statistics* ed M. Davidian (Washington, DC: Springer), 196–202.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2:37–52.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv [Preprint] arxiv*:1505.00853.

Yang, S., and Ramanan, D. (2015). "Multi-scale recognition with DAG-CNNs," in *Proceedings of the IEEE International Conference on Computer Vision* (Nice), 1215–1223.

Yang, Y., and Hospedales, T. M. (2014). A unified perspective on multi-domain and multi-task learning. *arXiv [Preprint] arxiv*:1412.7489.

Zardoshti-Kermani, M., Wheeler, B. C., Badie, K., and Hashemi, R. M. (1995). EMG feature evaluation for movement control of upper extremity prostheses. *IEEE Trans. Rehabil. Eng.* 3, 324–333. doi: 10.1109/86.481972

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Zurich: Springer), 818–833.

Zhang, X., Chen, X., Wang, W.-H., Yang, J.-H., Lantz, V., and Wang, K.-Q. (2009). "Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors," in *Proceedings of the 14th International Conference on Intelligent User Interfaces* (Sanibel Island, FL: ACM), 401–406.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2921–2929.

Zia ur Rehman, M., Waris, A., Gilani, S., Jochumsen, M., Niazi, I., Jamil, M., et al. (2018). Multiday EMG-based classification of hand motions with deep learning techniques. *Sensors* 18:2497. doi: 10.3390/s18082497

3D Tracking of Human Motion Using Visual Skeletonization and Stereoscopic Vision

Matteo Zago¹*, Matteo Luzzago², Tommaso Marangoni², Mariolino De Cecco³, Marco Tarabini² and Manuela Galli¹

¹ Department of Electronics, Information and Bioengineering, Polytechnic of Milan, Milan, Italy, ² Department of Mechanical Engineering, Polytechnic of Milan, Milan, Italy, ³ Department of Industrial Engineering, University of Trento, Trento, Italy

The design of markerless systems to reconstruct human motion in a timely, unobtrusive and externally valid manner is still an open challenge. Artificial intelligence algorithms based on automatic landmarks identification on video images opened to a new approach, potentially e-viable with low-cost hardware. OpenPose is a library that t using a two-branch convolutional neural network allows for the recognition of skeletons in the scene. Although OpenPose-based solutions are spreading, their metrological performances relative to video setup are still largely unexplored. This paper aimed at validating a two-cameras OpenPose-based markerless system for gait analysis, considering its accuracy relative to three factors: cameras' relative distance, gait direction and video resolution. Two volunteers performed a walking test within a gait analysis laboratory. A marker-based optical motion capture system was taken as a reference. Procedures involved: calibration of the stereoscopic system; acquisition of video recordings, simultaneously with the reference marker-based system; video processing within OpenPose to extract the subject's skeleton; videos synchronization; triangulation of the skeletons in the two videos to obtain the 3D coordinates of the joints. Two set of parameters were considered for the accuracy assessment: errors in trajectory reconstruction and error in selected gait space-temporal parameters (step length, swing and stance time). The lowest error in trajectories ($\sim$20 mm) was obtained with cameras 1.8 m apart, highest resolution and straight gait, and the highest ($\sim$60 mm) with the 1.0 m, low resolution and diagonal gait configuration. The OpenPose-based system tended to underestimate step length of about 1.5 cm, while no systematic biases were found for swing/stance time. Step length significantly changed according to gait direction ($p = 0.008$), camera distance ($p = 0.020$), and resolution ($p < 0.001$). Among stance and swing times, the lowest errors (0.02 and 0.05 s for stance and swing, respectively) were obtained with the 1 m, highest resolution and straight gait configuration. These findings confirm the feasibility of tracking kinematics and gait parameters of a single subject in a 3D space using two low-cost webcams and the OpenPose engine. In particular, the maximization of cameras distance and video resolution enabled to achieve the highest metrological performances.

Keywords: movement measurement, gait analysis, computer vision, artificial intelligence, markerless motion capture

INTRODUCTION

The measurement of human motion represents one of the most interesting and challenging topics of metrology. Optical motion tracking solutions can be broadly categorized into marker-based and markerless systems (Winter, 1990; Zhou and Hu, 2008). Mostly represented by the first group, the modern technological standards ground on established measurement principles and techniques: the position of joints and the orientation body segments are obtained through the three-dimensional localization of passive (or less often, active) markers, fixed on subjects' body and captured by a calibrated multi-camera stereophotogrammetric video system (Cappozzo et al., 2005). The human body is a complex, self-occluding and only partially rigid entity (Mündermann et al., 2006). Thus, instead of directly tracking human body pose, these systems work by identifying common object features in consecutive images (fiducial points or landmarks), which are used to track the motion of a series of rigid bodes connected by rotational joints (Winter, 1990). This solution provides the best metrological performances, in terms of accuracy in the markers' localization (usually in the order of 10ths of millimeters), repeatability and frequency of measurements (Ma'touq et al., 2018). Owing to their cost, complexity and required personnel to run the recording and place the markers on specific anatomical landmarks, marker-based systems are mainly used in specialized laboratories for clinical/rehabilitation applications or entertainment and digital animation (Winter, 1990; Cappozzo et al., 2005).

With the aim of limiting these drawbacks, in the last decades the interest toward markerless solution has grown rapidly, trying either to reduce the cost of technology or to simplify the process (Abbondanza et al., 2016; Ronchi and Perona, 2017; Colyer et al., 2018; Mizumoto et al., 2018; Tanaka et al., 2018; Tarabini et al., 2018a; Clark et al., 2019). Markerless systems are based on four main components, namely a camera system, a body model, the image features used and the algorithms that determine shape, pose and location of the model itself (Colyer et al., 2018). Two families of camera systems can be used, differing by whether or not they produce a so-called "depth map," i.e., an image where each pixel describes the distance of a point in the scene from the camera. Probably the best-known depth-sensing camera systems (often referred to as RGB-D cameras as they capture both color and depth) are Microsoft Kinect, Intel Realsense, and StereoLabs Zed. These solutions are particularly effective for real-time full body pose estimation in interactive systems and videogames (Shotton et al., 2011; Ye et al., 2013), but they also have limitations that hinder their wide application in clinical or biomechanical setting: short-range, inoperability in bright sun light, and potential interference between multiple sensors (Colyer et al., 2018). In addition, the accuracy in motion tracking is still lower than marker-based systems, which actually remain the gold standard.

Recently, novel artificial intelligence algorithms based on automatic landmarks identification on video images (computer vision) opened to a new approach for markerless motion capture, which became potentially feasible with low-cost hardware (Cao et al., 2016; Colyer et al., 2018; Clark et al., 2019). In that,

machine learning techniques were exploited to identify the nodes of a skeletal structure describing the posture of a human subject within a given image frame. As the associated computational burden made this method practicably unviable, the process was optimized by a research group from the Carnegie Mellon University, who released a processing framework called *OpenPose* (Cao et al., 2016). OpenPose takes as input color images from simple web-cameras and using a two-branch convolutional neural network (CNN) produces as output confidence maps of keypoints, and affinity for each keypoint pair (that is, belonging to the same skeleton). This way, OpenPose allows for the recognition of skeletons of multiple persons in the same scene. Some Authors adopted these OpenPose-based solutions as a functional block of their research: an example is Huang et al. (2017), in which OpenPose was used as an initialization step for the reconstruction of 3D human shape; a different approach is presented in Mehta et al. (2017), in which a 3D skeletal model was obtained starting from a single planar image.

Although promising results were obtained, the design of markerless systems able to reliably reconstruct human motion in a timely, unobtrusive and externally valid manner is still an open challenge (Colyer et al., 2018). Among the fast-growing studies on the application to various case studies, only a few focused on the accuracy of subjects' three-dimensional reconstruction: the performance of OpenPose in the computation of the lower limbs angles were analyzed with a single camera (Gu et al., 2018), and compared to a multi-camera marker based system. However, to the best of our knowledge, a targeted metrological characterization of data processing with multiple viewpoints is still missing in the case of automated walking analysis. At present, example of OpenPose applications for the extraction of gait parameters are scant. We hypothesize that the cameras resolution and positioning, as well as the walking direction (i.e., angle with respect to cameras) could affect the accuracy and thus feasibility of such systems in the clinical setting. Thus, this paper aims at describing and validating an OpenPose-based markerless motion tracking system for gait analysis against a gold-standard commercial marker-based motion capture system, discussing the extent to which the aforementioned factors affect the tracking quality.

METHODS

Experimental Design and Participants

This observational case-series study was designed to determine the metrological performance of the stereoscopic system featured by OpenPose. The study involved two healthy volunteers who performed a walking test at comfortable walking speed within an instrumented gait analysis laboratory. The two participants were both 24-years-old male adults, with the following heights and body masses: 1.73 m and 61 kg, 1.82 m and 75 kg. They wore minimal, close-fitting clothes. Participants were instructed about the aims and benefits of the study, and they both signed a written informed consent prior to laboratory sessions. As this study did not involve any clinical intervention or functional/physical evaluation, the approval from the Ethics Committee was not

FIGURE 1 | Laboratory setup, schematic **(left)** and pictorial **(right)** view.
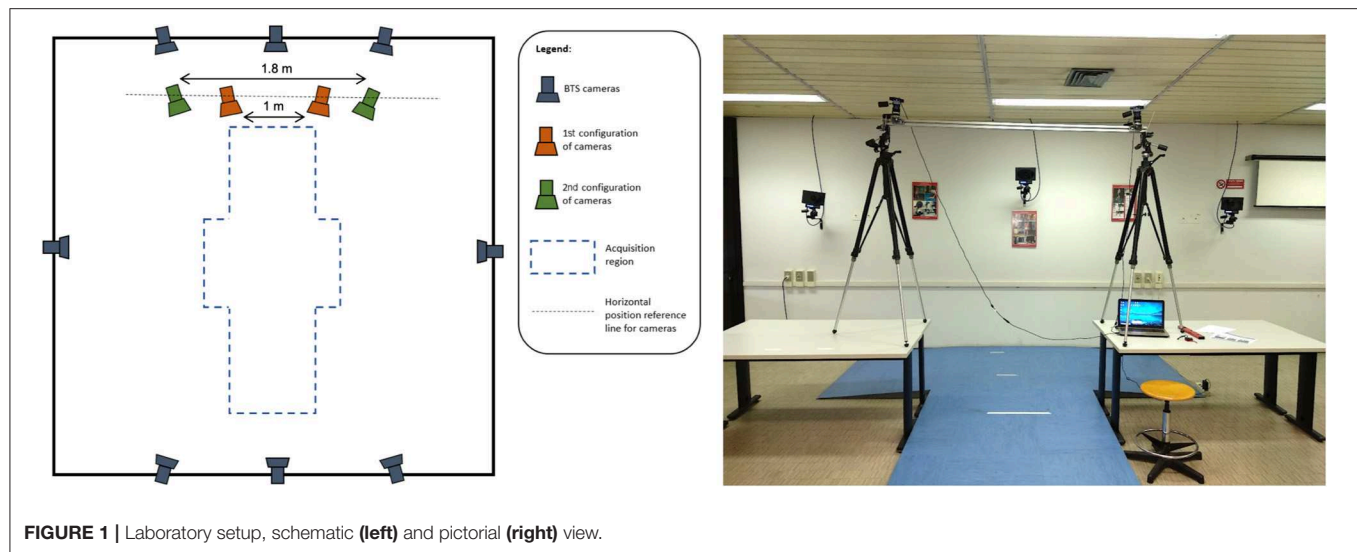
required. The study was carried out in accordance with the 1964 Helsinki declaration and its later amendments.

The effect of three factors potentially influencing the accuracy of the proposed system were considered:

1. Cameras' relative distance: cameras were positioned 1 m and then 1.8 m apart;
2. Gait direction, straight or diagonal, defined by means of visual references positioned along the path (the same for all the tests repetitions). In the second, additional sources of error may arise from the occlusions between body parts; subjects walked on a footboard and the walking direction was perpendicular to the cameras' connecting axis.
3. Video resolution: high (1,312 × 736 pixel), and standard (640 × 480 pixel). Both resolutions were obtained by scaling the camera native resolution with a cubic interpolation, this way we avoided the repetition of recording sessions.

Given that each factor assumed two levels, 2^3 (8) configurations were possible. Each test configuration was replicated 3 times per each volunteer; 48 tests were therefore performed.

Measurement System and Equipment

Two full-HD webcams (PC-W1, Aukey, Shenzhen, China) with a native image resolution of 1920 × 1080 pixels and a 1/2.7" CMOS sensor were used. Cameras acquired images at 30 Hz, with contrast and brightness automatically selected by the software provided by the manufacturer. Cameras were fastened on an aluminum bar perpendicular to the strait gait direction at a height of 2.3 m, framing the subject frontally.

A stereophotogrammetric motion analyser (Smart-D, BTS Bioengineering, Milano, Italy) equipped with eight infrared cameras sampling at 100 Hz was used as reference measurement system. The system was calibrated according to the manufacturer's specification, and the error in markers' location reconstruction was 0.2 mm on a working volume of 3 × 2 × 2 m. **Figure 1** shows the implemented measurement infrastructure.

Procedures

The measurement process can be summarized as follows:

1. Calibration of the stereoscopic system using planar patterns (Zhang, 2000; Hartley and Zisserman, 2003). Cameras calibration was performed within Matlab (v2018b, The Mathworks Inc., Natwick, USA) by means of the Camera Calibration Toolbox. A black and white checkerboard whose geometry is known (70 × 50 cm) is framed by the two cameras while spanning the checkerboard into the working volume. The Toolbox returns an estimate of the cameras internal and external parameters (i.e., lens distortion, camera relative orientation and position). To get a calibration metric, the reprojection error is computed by projecting the checkerboard points from world coordinates into image coordinates. Mean reprojection error was 0.18 pixels in the 1-m distance configuration, and 0.12 pixels in the 1.8 m configuration.
2. Acquisition of two video recordings, *a* and *b* (one per each webcam), using the cameras of the stereoscopic system. Each recording allowed to collect between four and five steps, according to the laboratory dimension, and lasted 4.5–6.5 s.
3. Simultaneous recording using the reference, marker-based optical system. Twenty-four reflective markers were placed on the subject in the following anatomical landmarks (see **Figure 2**): sternum and sacrum; right and left acromion, medial and lateral humeral epicondyles, radius and ulnar styloid process, antero-superior iliac spines, greater trochanter, medial and lateral femoral epicondyles, medial and lateral malleoli. This marker set was adapted from standard protocols used in clinical gait analysis (Davis et al., 1991; Zago et al., 2017), and was designed to match the skeletal configuration of OpenPose (**Figure 2**). To do so, wrists, elbows, knees and ankles joint centers were located at the midpoint (average) of medial and lateral markers. Hip joint centers were computed using regression equations as prompted by the International Society of Biomechanics standards (Wu and Cavanagh, 1995).

FIGURE 2 | Stick diagrams as returned by the marker-based optical system (top, left) and OpenPose model (top, right); corresponding 3D reconstruction of the skeletal structures during walking (bottom).

keypoints are located in relevant body landmarks (such as left hand, right hand, face, etc.) and were used to determine the 3D Cartesian coordinates, positioning the skeletal model of the subject in the space with respect to reference system of one camera. This operation was performed using the Matlab Computer Vision System Toolbox (v2018b, The Mathworks Inc., Natwick, USA), obtaining the 3D stereoscopic triangulation of the camera pixel coordinates, which included:

- the intrinsic calibration parameters of each camera, for the assessment of focal lengths, camera centers and distortion parameters;
- the extrinsic calibration parameters, accounting for the relative position of cameras;
- the undistortion of pixel coordinates;
- the application of a functional triangulation for each of the 2D keypoints for the identification of the corresponding 3D coordinates in the epipolar plane.

The resulting output was the 3D skeletal model of the subject, as shown in the bottom-right panel of **Figure 2**.

Prior to further processing, coordinates returned by both the OpenPose and the marker-based reference system were filtered using a zero-lag, 2nd order Butterworth filter with a cut-off frequency of 10 Hz.

Data Synchronization and Spatial Alignment

Since a physical trigger for the synchronization of the cameras with the motion capture system was not available, we asked the subjects to perform a sequence of repeated actions (to beat the right hand on the right hip). The synchronization procedure was repeated before each single test and it was achieved by overlapping the time series of the distance between the right wrist and right hip markers returned by the two systems. Prior to do so, the signals were both downsampled (cubic splines interpolation) to 30 Hz. Drift errors due to different sampling rates (100 Hz for the marker-based system, 30 Hz for the webcams) were negligible given the test duration of a few seconds.

The spatial alignment of the reference systems completed the measurement systems calibration: the 3D coordinates provided by the triangulation of OpenPose data were originally expressed in a reference system located in the optical center of one of the cameras, oriented as the camera itself. The marker-based system returns 3D coordinates resolved in global (laboratory) reference system fixed on the ground at the center of the working volume. These two coordinate systems were moved to a new, coincident, reference frame, positioned midway between the two cameras, with the origin at the ground level and with the axes directed as those of the original marker-based system. The alignment procedure was taken from Kabsch (1976) and involved the initial rotation of the OpenPose reference system, followed by the translation toward the desired origin.

Target Parameters Computation

Within the OpenPose-based system, the definition of the gait phases relies on the recognition of the foot condition—stance or swing (Saggin et al., 2013). The distance between two successive stance statuses represents the target measurement. In our case,

4. Video processing within OpenPose to extract the skeleton S of a (single) subject in each video recording (S_a and S_b).
5. Synchronization of the two videos (see paragraph Data Synchronization and Spatial Alignment).
6. Triangulation of the skeletons S_a and S_b using the calibration outcome (step 1) to obtain the three-dimensional coordinates of the joints and alignment between coordinate system of step 4.
7. Computation of gait parameters (see paragraph Target Parameters Computation) based on the three-dimensional coordinates obtained.
8. Evaluation of the OpenPose accuracy for each single test according to the metrics defined in the following paragraph.
9. Evaluation of the dependence of accuracy from the factors' levels using a 3×2 Analysis of Variance.

Data Processing

A set of 18 2D keypoints coordinates for body pose estimation (in pixels) are returned by OpenPose from video images; 2D

FIGURE 3 | Extraction of gait phases from the trajectories of ankle nodes' velocity, explanatory example taken from a straight gait test. OP, OpenPose-based system; MB, marker-based optical system.

the processing structure was taken from Tarabini et al. (2018b) and involved the analysis on the velocities of the nodes located at the ankle level (**Figure 3**). Given a window of n elements, the magnitude of the velocity (v) of the two ankle nodes was computed as:

$$v = f \cdot \sqrt{\prod_{i=1}^{n} (x_i - x_{i-1})^2 + \prod_{i=1}^{n} (y_i - y)^2 + \prod_{i=1}^{n} (z_i - z_{i-1})^2}$$

where f is the sample frequency (30 Hz). To minimize the influence of noise and ease foot status detection, a moving average lowpass filter was then applied on v, with a period of 12 samples (with a 30 Hz sampling frequency, the first zero of the filter transfer function is at 1.25 Hz). Two thresholds on the filtered velocity signal of the ankle node were set for the identification of the foot status: *HystLowSpeed* and *HystHighSpeed*. These were automatically obtained for each subject from a complete gait test. After an initial sorting of all velocities assessed from the test and reorganized in the form of a histogram, the values were computed as:

- *HystHighSpeed*: upper threshold limit, as the value corresponding to 65% of the sorted velocities: when the joint's filtered speed was higher than this value, then the foot was considered in the swing state (1).
- *HystLowSpeed*: lower threshold limit, equal to 80% of *HystLowSpeed*, to avoid erroneous swing's end caused by small variations induced by residual noise components. In this case, the foot was considered in the stance state (0).

To get correct gait parameters' values, it is essential to consider complete steps only. For such a reason, four cases were analyzed:

1. Foot enters the considered acquisition window in swing state (1) and exits still in swing state (1) (if the acquisition

contained at least a complete step, first and last step were not considered).
2. Foot enters the considered acquisition window in stance state (0) and exits in swing (1) state (the last step was not considered).
3. Foot enters the considered acquisition window in swing state (1) and exits in stance state (0) (the first step was not considered).
4. Foot enters in the considered acquisition window in stance state (0) and exits still in stance state (0) (if the acquisition contained any number of steps, they were all considered).

Evaluation of Accuracy

In each test, the accuracy of the proposed system was evaluated in terms of two sets of parameters, retrieved from the same recorded dataset:

- **Error in the reconstruction of the trajectories**, computed as the Root Mean Square (RMS) distance between the trajectories of selected, corresponding skeletal nodes. In doing so, the most similar physical fiducial points were considered: wrists, elbows, knees and ankles. Indeed, the reference and the proposed skeletal structures do not correspond perfectly. Thus, a minimization procedure was used to align the thirteen landmarks of the skeleton, and a roto-translation of the trajectories obtained with the OpenPose-based system was performed to align them to the correspondent reference (marker-based coordinates). The complete procedure is described in Abbondanza et al. (2016) and Tarabini et al. (2018b) and it is based on the calculation of the Euclidean distance in each frame between correspondent keypoints of the two systems. This method was already used to synchronize trajectories acquired with different measurement systems, and it proved to be
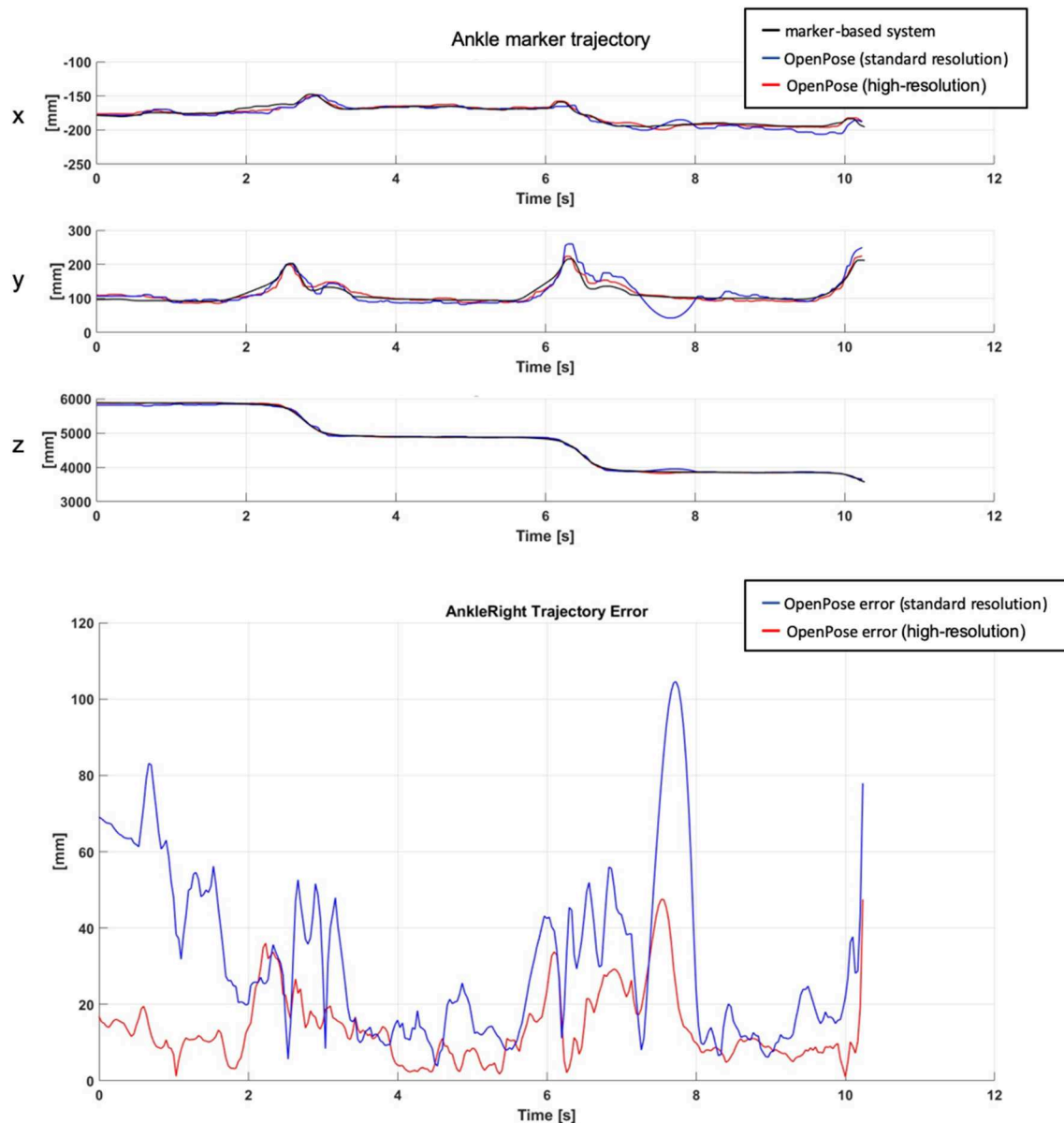
FIGURE 4 | Sample trajectories of a landmark (position of the right ankle) obtained from the reference marker-based (black) and markerless, OpenPose-based (blue and red) systems **(top)**; corresponding RMS distance **(bottom)**.

unbiased also in presence of offset between the skeletal markers (Abbondanza et al., 2016; Tarabini et al., 2018b).

- **Error in gait space-temporal parameters**: step length (distance between consecutive heel-strikes position), stance and swing time were extracted (Perry and Burnfield, 2010). The RMS error with the correspondent parameters computed with the reference marker-based system was computed.

Statistical Analysis

The effect of the three factors (cameras' distance, gait direction, and resolution) on the measurement error was assessed using the following analysis of variance (ANOVA) model design (Moschioni et al., 2013):

$$\xi = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{(1,2)} x_1 x_2 + \cdots + \epsilon$$

where ξ is the dependent variable, namely the skeletal node position error (RMS) or the error of one of the gait analysis parameters (step length, stance and swing time), and x_i are the independent variables (previously referred to as factors). β_0 is the global tests average, β_i and $\beta_{(1,2)}$ are used to describe the effect of the independent variables and their interactions (in particular, gait direction × camera distance interaction was assessed); ϵ is the residual, namely the difference between the actual data

TABLE 1 | Root Mean Square errors measured at different skeleton nodes as a function of gait direction, camera distance and video resolution.

#	Gait type	Straight gait				Diagonal gait			
	Distance (m)	1.8	1.8	1.0	1.0	1.8	1.8	1.0	1.0
	Resolution	LR	HR	LR	HR	LR	HR	LR	HR
1	Sternum	25.2	16.8	41.1	22.1	65.8	54.0	69.5	60.1
2	Shoulder, left	32.3	20.0	42.2	26.5	46.2	45.5	52.8	51.2
3	Shoulder, right	27.5	17.7	37.2	22.1	42.1	39.2	48.3	41.2
4	Elbow, left	28.2	18.3	49.8	24.6	69.9	61.6	69.1	58.3
5	Elbow, right	27.2	18.8	50.3	27.9	62.7	74.4	48.5	44.6
6	Wrist, left	23.3	17.1	45.1	22.9	51.6	51.8	79.1	54.0
7	Wrist, right	25.4	16.1	48.6	22.3	65.1	66.9	56.1	48.2
8	Hip, left	33.0	21.9	48.5	29.0	79.9	81.1	67.5	75.7
9	Hip, right	34.6	23.4	50.1	41.4	79.5	82.5	71.6	73.5
10	Knee, left	30.9	23.4	60.0	29.2	53.7	53.0	58.7	55.4
11	Knee, right	33.9	24.8	39.6	19.9	58.4	57.6	63.7	48.3
12	Ankle, left	39.5	26.1	62.2	24.0	69.1	68.7	87.4	103.3
13	Ankle, right	35.8	26.2	34.5	21.4	63.5	62.7	64.9	61.3
-	Mean	30.5	20.8	46.9	25.6	62.1	61.5	64.4	59.6
-	SD	4.6	3.5	7.9	5.4	11.1	12.6	11.1	16.0

Values in mm.

TABLE 2 | Root Mean Square errors measured between the reference (marker-based) and OpenPose-based systems on selected spatio-temporal gait parameters.

Gait type	Straight gait				Diagonal gait			
Distance (m)	1.8	1.8	1.0	1.0	1.8	1.8	1.0	1.0
Resolution	LR	HR	LR	HR	LR	HR	LR	HR
Step length (cm)	3.26	1.53	7.42	1.93	2.45	1.66	3.25	1.23
Swing time (s)	0.04	0.05	0.05	0.02	0.03	0.03	0.03	0.06
Stance time (s)	0.05	0.05	0.05	0.05	0.07	0.07	0.05	0.08

behavior and the model prediction. A significance alpha level of 5% was implemented throughout.

In addition, Bland-Altman plots were used to graphically compare gait analysis parameters obtained with the reference and OpenPose-based systems.

RESULTS

Figure 4 shows an explanatory plot of the original joints coordinates and of the corresponding measurement error over time. Overall measurement errors (RMS) are reported in **Tables 1**, **2**, and graphically summarized in the boxplots of **Figure 5**; **Table 3** displays the relevant statistic: all factors ($p < 0.01$) and interaction ($p < 0.001$, see **Figure 6**) resulted to be statistically significant relative to trajectories reconstruction error. The lowest error (about 20 mm) was obtained with the 1.8 m, highest resolution and straight gait configuration, and the highest (>60 mm) with the 1.0 m, low resolution and diagonal gait configuration.

Bland-Altman plots displaying gait parameters comparison are shown in **Figure 7**: the proposed system tended to underestimate step length of about 1.5 cm, while no systematic biases were found for swing/stance time. Step length significantly changed according to gait direction ($p = 0.008$), camera distance ($p = 0.020$) and resolution ($p < 0.001$, see **Table 3**). Consistently with trajectories' RMS, the lowest error in step length (1.53 cm) was obtained with the 1.8 m, high resolution and straight gait configuration. Among stance and swing times, only for the first emerged a significant factor, i.e., camera distance ($p = 0.038$), and the lowest errors (0.02 s and 0.05 s for stance and swing, respectively) were obtained with the 1 m, high resolution and straight gait configuration.

DISCUSSION

The findings of this work confirm to the feasibility of tracking kinematics and gait parameters of a single subject in a 3D space using two low-cost cameras and the OpenPose engine. The accuracy of markerless motion tracking depends on three factors: the occlusions between body parts, cameras position/orientation and video resolution; considering the best combination of the considered factors (cameras distance 1.8 m, maximum resolution, and no occlusions due to straight walking) the lowest error in 3D trajectories reconstruction was about 20 mm, the lowest error in swing/stance time was 0.03 s and 1.23 cm in step length. Values are comparable with intra-subject variability in clinical gait analysis investigations (Ciprandi et al., 2017; Temporiti et al., 2019), thus encourage a preliminary adoption of OpenPose-based markerless solutions in this setting. However, it should be noticed that a different configuration (smaller camera distance, lower resolution or diagonal gait direction) can negatively affect the results.

Accuracy

In our optimal configuration, average markers RMS was about 20 mm. This can be considered a notable result, as it is only slightly higher than the error reported in a previous study (about 15 mm), where however eight cameras (fs = 120 Hz) and a subject-specific, way more complex anatomical model were used (Corazza et al., 2010). Dunne et al. reported an error of ~50 mm in reconstructing foot contact position with a single camera system (Dunn et al., 2014).

While several studies compared the outcome of an OpenPose markerless system to a traditional marker-based one (Clark et al., 2019), the majority focused on joint angles (Colyer et al., 2018) and to the best of our knowledge, none of them provided gait analysis parameters. Thus, a direct evaluation of the performances of our system is not straightforward. As a reference, Kinect-based markerless systems returned a lower accuracy of 2.5–5.5 cm in step length and a slightly lower accuracy of 60–90 ms in stance/swing time (Latorre et al., 2018). Previously, Barone et al. obtained comparable or slightly better results (accuracy of 3.7 cm for step length and 0.02 s for step duration) but they combined a markerless system with the signal coming from the accelerometer embedded in a smartphone (Barone et al., 2016).
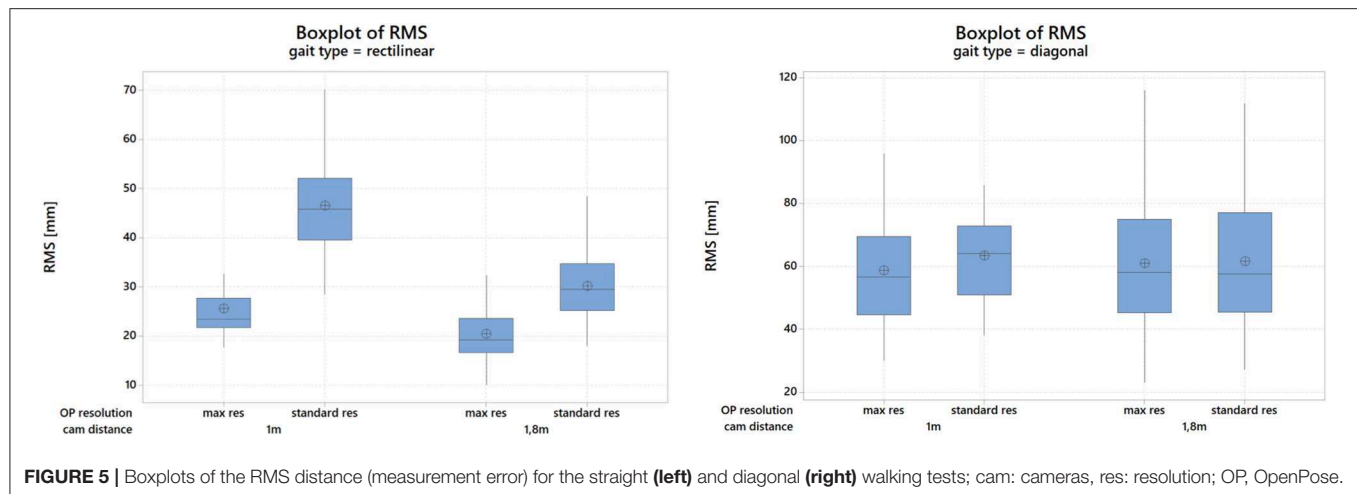
FIGURE 5 | Boxplots of the RMS distance (measurement error) for the straight **(left)** and diagonal **(right)** walking tests; cam: cameras, res: resolution; OP, OpenPose.

TABLE 3 | Statistical outcomes from the ANOVA computed on trajectories' RMS and on gait spatio-temporal parameters Root Mean Square errors.

Variable	Gait direction		Camera distance		Resolution		Direction * distance	
	F	p	F	p	F	p	F	p
RMS	392.39	**<0.001**	8.11	**0.005**	44.5	**<0.001**	19.6	**<0.001**
Step length	7.84	**0.008**	5.84	**0.020**	28.46	**<0.001**	6.09	**0.018**
Stance time	1.45	0.235	3.24	0.079	4.57	**0.038**	0.29	0.591
Swing time	10.55	0.002	0.06	0.811	1.22	0.275	0.05	0.817

Statistically significant values are in bold.

The resolution of the stereoscopic system is not constant, being dependent on the physical distance between the subject and cameras. The method performances worsen as the subject distance from the sensor increases: errors presented in these work are average values, both summarizing the ideal situation in which the subject is filling the two image planes and the situation in which the subject is far from the camera with a less favorable optical sensor resolution.

Effect of Camera Setup

Occlusions represented the most detrimental factor emerging from the comparison between gait types. The accuracy of results obtained in the diagonal gait tests was always lower than those obtained with a straight gait. In the 3D reconstruction all the other factors are almost negligible when occlusions are present. When body parts are occluded, OpenPose provides an estimation of the hidden landmarks, introducing an error that propagates in the 3D reconstruction. The problem is common with all the vision-based measurement systems and can be solved using more than 2 cameras simultaneously (most optoelectronic systems use from 6 to 12 cameras) so that each marker or joint is seen from more than 2 sensors.

Increasing the camera distance (from 1.0 to 1.8 m) in straight gait tests improved the accuracy of the reconstruction by 22.5%. Cameras relative distance and orientation influences the uncertainty of the triangulation, affecting the dimensions of the volume where the triangulated point can be placed. By positioning the two cameras further apart, the framed person is seen from a different perspective and the cameras are more convergent. This leads to a decrease of the capture volume where the triangulated point can be placed, but also a lower uncertainty in the triangulation process. In short, the higher the cameras distance, the narrower the working volume—but characterized by a higher accuracy.

When increasing video resolution, the error decreased of about 46% (1.0 m camera distance) and 32% (1.8 m camera distance). By increasing the video resolution, the uncertainty in the identification of the landmarks coordinates on the 2D images decreases, and the 3D reconstruction results more accurate. This comes at a cost: the main drawbacks are either higher processing time, to a first approximation linearly dependent to the number of pixels in the image, and more expensive hardware required to data processing. The spatial resolution of the system is not constant in the observed volume: the pixel to distance conversion factor depends on the position of the subject with respect to the cameras; consequently, the optical resolution worsen when the subject is far and occupies a small portion of the image. The problem can be solved by putting more cameras surrounding the subject and observing the motion from different directions, as in common optoelectronic systems. Since in our test the subject distance from the cameras varied approximately between 2 and 6 m, errors' numerical values are the average between optimal conditions (in which the subject fills the image) and worst ones. Consequently, in static applications when the subject is not moving, we can expect better performances with respect to values reported here.

Limitations and Perspectives

This pilot study was limited to two healthy subjects; a larger population could be considered in further research

FIGURE 6 | Effects plot for the RMS distance, according to the gait type (walking direction), cameras distance and resolution **(left)**. Gait type×distance interaction plot **(right)**. Cam: cameras, res: resolution; OP, OpenPose.

FIGURE 7 | Bland-Altman plot of the pooled (selected) gait parameters, comparison between the OpenPose- and the marker-based systems.
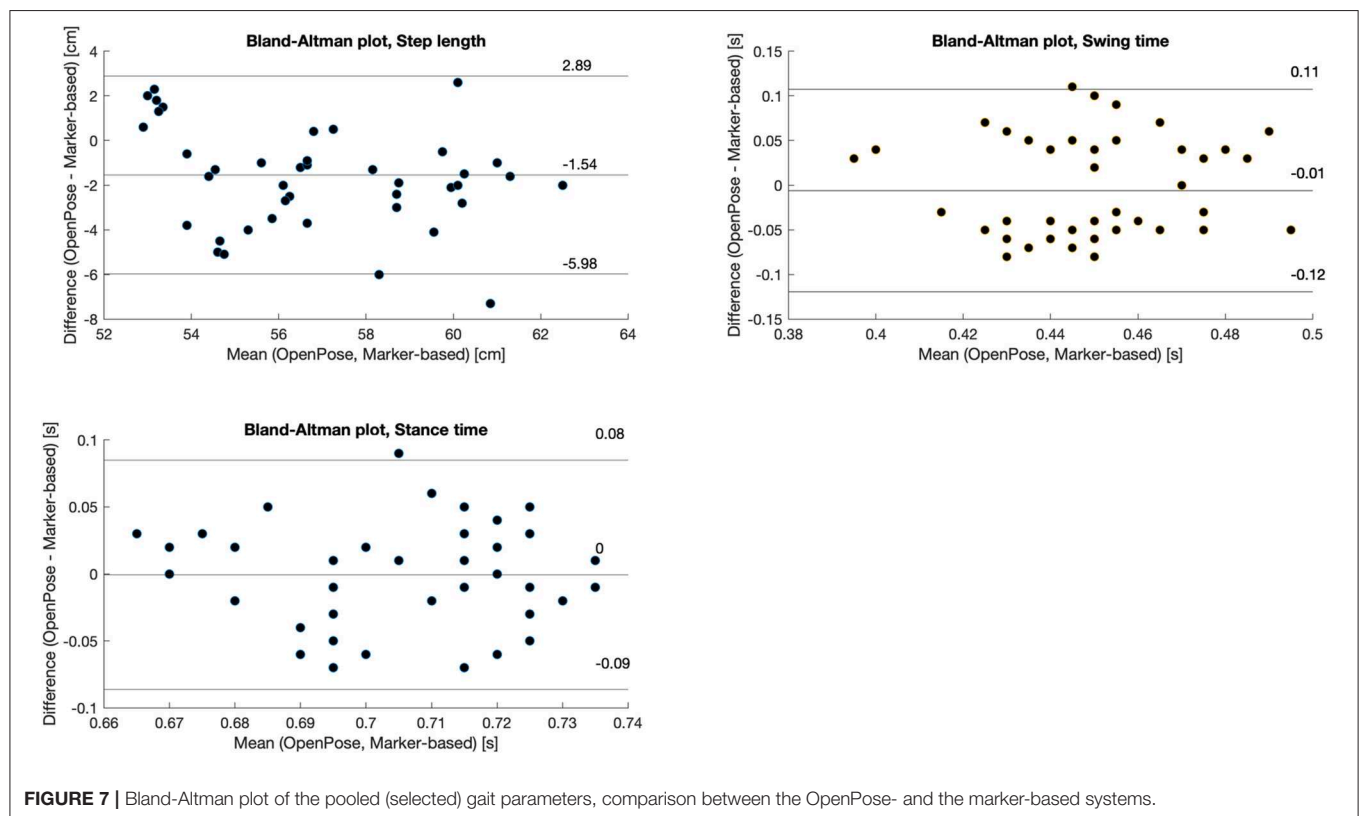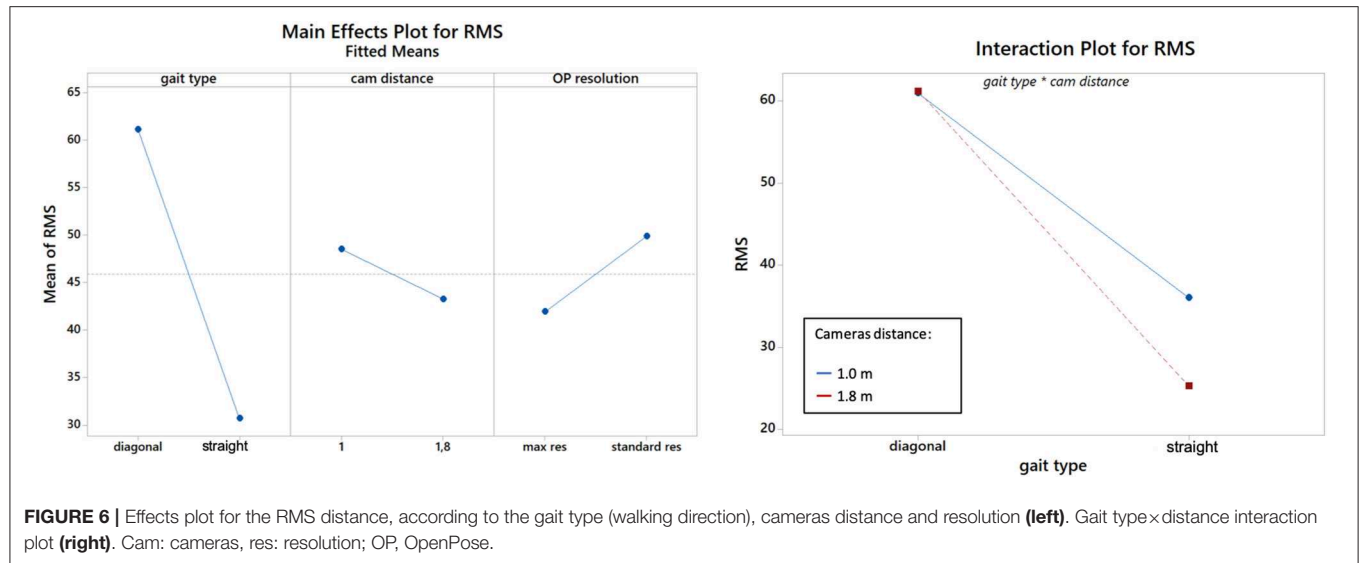
to address, for instance, the effect of body size on the tracking accuracy, as well as potential effects of clothes. It is advisable that future research lines address the metrological characterization of multi-camera systems, which will enable a complete 360 degrees view of the subject. In this, the occurrence of occlusions will be minimized, and a more accurate reconstruction is expected, at the expenses of a more complex hardware infrastructure.

CONCLUSION

In this work, a metrological characterization of OpenPose processing in the context of gait analysis by mean of low-cost stereoscopy was presented. Intentionally, no changes were made to the original software interface, working instead on the test configuration and on the influence factors in the metrological setup. Thus, all the insights concern the actual processing

algorithms, not considering improvements deriving from the optimization or tuning of the code for a specific task.

Although future improvements in OpenPose performance are expected, both in terms of accuracy in landmarks identifications and processing speed, the proposed analysis considered general, "external" factors that will remain practically valid. In particular, we showed that the maximization of cameras distance and video resolution enabled to achieve the highest metrological performances. Therefore, system accuracy could further be improved by reducing the presence of occlusions not only through a better joint location prediction in the source images, but also multiplying the number of cameras, thus obtaining a perspective closer to the straight walking condition.

This work points the way to further applications in environments where a video-based acquisition would be particularly useful, i.e., those where a quick and economical evaluation by non-expert operators is required.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

ML, TM, MZ, and MD carried out the experiment and undertook the data analysis. MZ and MT wrote the manuscript. MG and MT supervised the multidisciplinary project and revised the manuscript.

ACKNOWLEDGMENTS

REFERENCES

Abbondanza, P., Giancola, S., Sala, R., and Tarabini, M. (2016). "Accuracy of the microsoft kinect system in the identification of the body posture," in *International Conference on Wireless Mobile Communication and Healthcare* (Milan), 289–296. doi: 10.1007/978-3-319-58877-3_37

Barone, V., Verdini, F., Burattini, L., Di Nardo, F., and Fioretti, S. (2016). A markerless system based on smartphones and webcam for the measure of step length, width and duration on treadmill. *Comput. Methods Programs Biomed.* 125, 37–45. doi: 10.1016/j.cmpb.2015.12.003

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. *arXiv Prepr.* 1:arXiv1611.08050. doi: 10.1109/CVPR.2017.143

Cappozzo, A., Della Croce, U., Leardini, A., and Chiari, L. (2005). Human movement analysis using stereophotogrammetry. Part 1: theoretical background. *Gait Posture* 21, 186–196. doi: 10.1016/S0966-6362(04)00025-6

Ciprandi, D., Bertozzi, F., Zago, M., Ferreira, C. L. P., Boari, G., Sforza, C., et al. (2017). Study of the association between gait variability and physical activity. *Eur. Rev. Aging Phys. Act.* 14, 14:19. doi: 10.1186/s11556-017-0188-0

Clark, R. A., Mentiplay, B. F., Hough, E., and Pua, Y. H. (2019). Three-dimensional cameras and skeleton pose tracking for physical function assessment: a review of uses, validity, current developments and Kinect alternatives. *Gait Posture* 68, 193–200. doi: 10.1016/j.gaitpost.2018.11.029

Colyer, S. L., Evans, M., Cosker, D. P., and Salo, A. I. T. (2018). A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sport. Med. Open* 4:24. doi: 10.1186/s40798-018-0139-y

Corazza, S., Mündermann, L., Gambaretto, E., Ferrigno, G., and Andriacchi, T. P. (2010). Markerless motion capture through visual hull, articulated ICP and subject specific model generation. *Int. J. Comput. Vis.* 87, 156–169. doi: 10.1007/s11263-009-0284-3

Davis, R. B., Õunpuu, S., Tyburski, D., and Gage, J. R. (1991). A gait analysis data collection and reduction technique. *Hum. Mov. Sci.* 10, 575–587. doi: 10.1016/0167-9457(91)90046-Z

Dunn, M., Haake, S., Wheat, J., and Goodwill, S. (2014). Validation of a single camera, spatio-temporal gait analysis system. *Procedia Eng.* 72, 243–248. doi: 10.1016/j.proeng.2014.06.043

Gu, X., Deligianni, F., Lo, B., Chen, W., and Yang, G.-Z. (2018). "Markerless gait analysis based on a single RGB camera," in *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (Las Vegas, NV), 42–45. doi: 10.1109/BSN.2018.8329654

Hartley, R., and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511811685

Huang, M., Liao, L., and Pang, M. Y. C. (2017). Effects of whole body vibration on muscle spasticity for people with central nervous system disorders : a systematic review. *Clin Rehabil.* 31, 23–33. doi: 10.1177/0269215515621117

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A Cryst. Physics, Diffraction Theor. Gen. Crystallogr.* 32, 922–923. doi: 10.1107/S0567739476001873

Latorre, J., Llorens, R., Colomer, C., and Alcañiz, M. (2018). Reliability and comparison of Kinect-based methods for estimating spatiotemporal gait parameters of healthy and post-stroke individuals. *J. Biomech.* 72, 268–273. doi: 10.1016/j.jbiomech.2018.03.008

Ma'touq, J., Hu, T., and Haddadin, S. (2018). Sub-millimetre accurate human hand kinematics: from surface to skeleton. *Comput. Methods Biomech. Biomed. Engin.* 21, 113–128. doi: 10.1080/10255842.2018.1425996

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., et al. (2017). Vnect: real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.* 36:44. doi: 10.1145/3072959.3073596

Mizumoto, T., Fornaser, A., Suwa, H., Yasumoto, K., and De Cecco, M. (2018). "Kinect-based micro-behavior sensing system for learning the smart assistance with human subjects inside their homes," in *2018 Workshop on Metrology for Industry 4.0 and IoT* (Brescia), 1–6. doi: 10.1109/METROI4.2018.8428345

Moschioni, G., Saggin, B., Tarabini, M., Hald, J., and Morkholt, J. (2013). Use of design of experiments and Monte Carlo method for instruments optimal design. *Measurement* 46, 976–984. doi: 10.1016/j.measurement.2012.10.024

Mündermann, L., Corazza, S., and Andriacchi, T. P. (2006). The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *J. Neuroeng. Rehabil.* 3:6. doi: 10.1186/1743-0003-3-6

Perry, J., and Burnfield, J. N. (2010). *Gait Analysis: Normal and Pathological Function. 2nd Edn.* Thorofare, NJ: SLACK Incorporated.

Ronchi, M. R., and Perona, P. (2017). "Benchmarking and error diagnosis in multi-instance pose estimation," in *Computer Vision (ICCV), 2017 IEEE International Conference on* (Venice), 369–378. doi: 10.1109/ICCV.2017.48

Saggin, B., Scaccabarozzi, D., and Tarabini, M. (2013). Metrological performances of a plantar pressure measurement system. *IEEE Trans. Instrum. Meas.* 62, 766–776. doi: 10.1109/TIM.2013.2245185

Shotton, J., Sharp, T., Fitzgibbon, A., Blake, A., Cook, M., Kipman, A., et al. (2011). "Real-Time human pose recognition in parts from single depth images," in *CVPR 2011* (Providence, RI: IEEE), 1297–1304. doi: 10.1109/CVPR.2011.5995316

Tanaka, R., Takimoto, H., Yamasaki, T., and Higashi, A. (2018). Validity of time series kinematical data as measured by a markerless motion capture system on a flatland for gait assessment. *J. Biomech.* 71, 281–285. doi: 10.1016/j.jbiomech.2018.01.035

Tarabini, M., Marinoni, M., Mascetti, M., Marzaroli, P., Corti, F., Giberti, H., et al. (2018a). "Monitoring the human posture in industrial environment: a feasibility study," in *2018 IEEE Sensors Applications Symposium (SAS)* (Seoul), 1–6. doi: 10.1109/SAS.2018.8336710

Tarabini, M., Marinoni, M., Mascetti, M., Marzaroli, P., Corti, F., Giberti, H., et al. (2018b). "Real-time monitoring of the posture at the workplace using low cost sensors," in *Congress of the International Ergonomics Association* (Florence), 678–688. doi: 10.1007/978-3-319-96083-8_85

Temporiti, F., Zanotti, G., Furone, R., Molinari, S., Zago, M., Loppini, M., et al. (2019). Gait analysis in patients after bilateral versus unilateral total hip arthroplasty. *Gait Posture* 72, 46–50. doi: 10.1016/j.gaitpost.2019.05.026

Winter, D. A. (1990). *Biomechanics and Motor Control of Human Movement. 2nd Edn.* New York, NY: Wiley.

Wu, G., and Cavanagh, P. R. (1995). ISB recommendations in the reporting for standardization of kinematic data. *J. Biomech.* 28, 1257–1261. doi: 10.1016/0021-9290(95)00017-C

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., and Gall, J. (2013). "A survey on human motion analysis from depth data," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Berlin). doi: 10.1007/978-3-642-44964-2_8

Zago, M., Camerota, T. C. T. C., Pisu, S., Ciprandi, D., and Sforza, C. (2017). Gait analysis of young male patients diagnosed with primary bladder neck obstruction. *J. Electromyogr. Kinesiol.* 35, 69–75. doi: 10.1016/j.jelekin.2017.05.005

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1330–1334. doi: 10.1109/34.888718

Zhou, H., and Hu, H. (2008). Human motion tracking for rehabilitation—a survey. *Biomed. Signal Process. Control* 3, 1–18. doi: 10.1016/j.bspc.2007.09.001

Can We Predict the Motor Performance of Patients With Parkinson's Disease Based on Their Symptomatology?

Karina Lebel [1,2]*, Christian Duval [3,4], Etienne Goubault [3,4], Sarah Bogard [3,4] and Pierre. J. Blanchet [5,6]

[1] Département de Génie électrique et de Génie Informatique, Faculté de Génie, Université de Sherbrooke, Sherbrooke, QC, Canada, [2] Centre de Recherche sur le Vieillissement, Sherbrooke, QC, Canada, [3] Laboratoire de Simulation et Modélisation du Mouvement, École de Kinésiologie et des Sciences de l'activité physique, Université de Montréal, Montreal, QC, Canada, [4] Centre de Recherche Institut Universitaire de Gériatrie de Montréal, Montreal, QC, Canada, [5] Faculté de Médecine Dentaire, Université de Montréal, Montréal, QC, Canada, [6] Centre Hospitalier de l'Université de Montréal (C.H.U. Montreal), Montreal, QC, Canada

Introduction: Parkinson's disease hinders the ability of a person to perform daily activities. However, the varying impact of specific symptoms and their interactions on a person's motor repertoire is not understood. The current study investigates the possibility to predict global motor disabilities based on the patient symptomatology and medication.

Methods: A cohort of 115 patients diagnosed with Parkinson's disease (mean age = 67.0 ± 8.7 years old) participated in the study. Participants performed different tasks, including the Timed-Up & Go, eating soup and the Purdue Pegboard test. Performance on these tasks was judged using timing, number of errors committed, and count achieved. K-means method was used to cluster the overall performance and create different motor performance groups. Symptomatology was objectively assessed for each participant from a combination of wearable inertial sensors (bradykinesia, tremor, dyskinesia) and clinical assessment (rigidity, postural instability). A multinomial regression model was derived to predict the performance cluster membership based on the patients' symptomatology, socio-demographics information and medication.

Results: Clustering exposed four distinct performance groups: *normal* behavior, slightly affected in fine motor tasks, affected only in TUG, and affected in all areas. The statistical model revealed that low to moderate level of dyskinesia increased the likelihood of being in the *normal* group. A rise in postural instability and rest tremor increase the chance to be affected in TUG. Finally, LEDD did not help distinguishing between groups, but the presence of Amantadine as part of the medication regimen appears to decrease the likelihood of being part of the groups affected in TUG.

Conclusion: The approach allowed to demonstrate the potential of using clinical symptoms to predict the impact of Parkinson's disease on a person's mobility performance.

Keywords: mobility, motor impact, Parkinson, clustering, K-means

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease characterized by both motor and non-motor symptoms, including tremor, postural instability, muscle rigidity, and bradykinesia or akinesia (Sveinbjornsdottir, 2016). These symptoms affect the ability of patients to perform activities of daily living (ADL) to a varying extent. There is currently no cure for PD, and symptoms are chiefly managed with medication. While the treatment goal is to maximize the person's ability to perform everyday tasks, the impact of each symptom on ADL, and most importantly, of the combination of symptoms, is not well-understood. Past studies have tried to identify different phenotypes in PD to help with this issue, and to guide diagnosis, prognosis, and treatment (Eisinger et al., 2019). These studies identified a tremor dominant subtype and a postural instability

FIGURE 1 | Study inclusion flowchart.
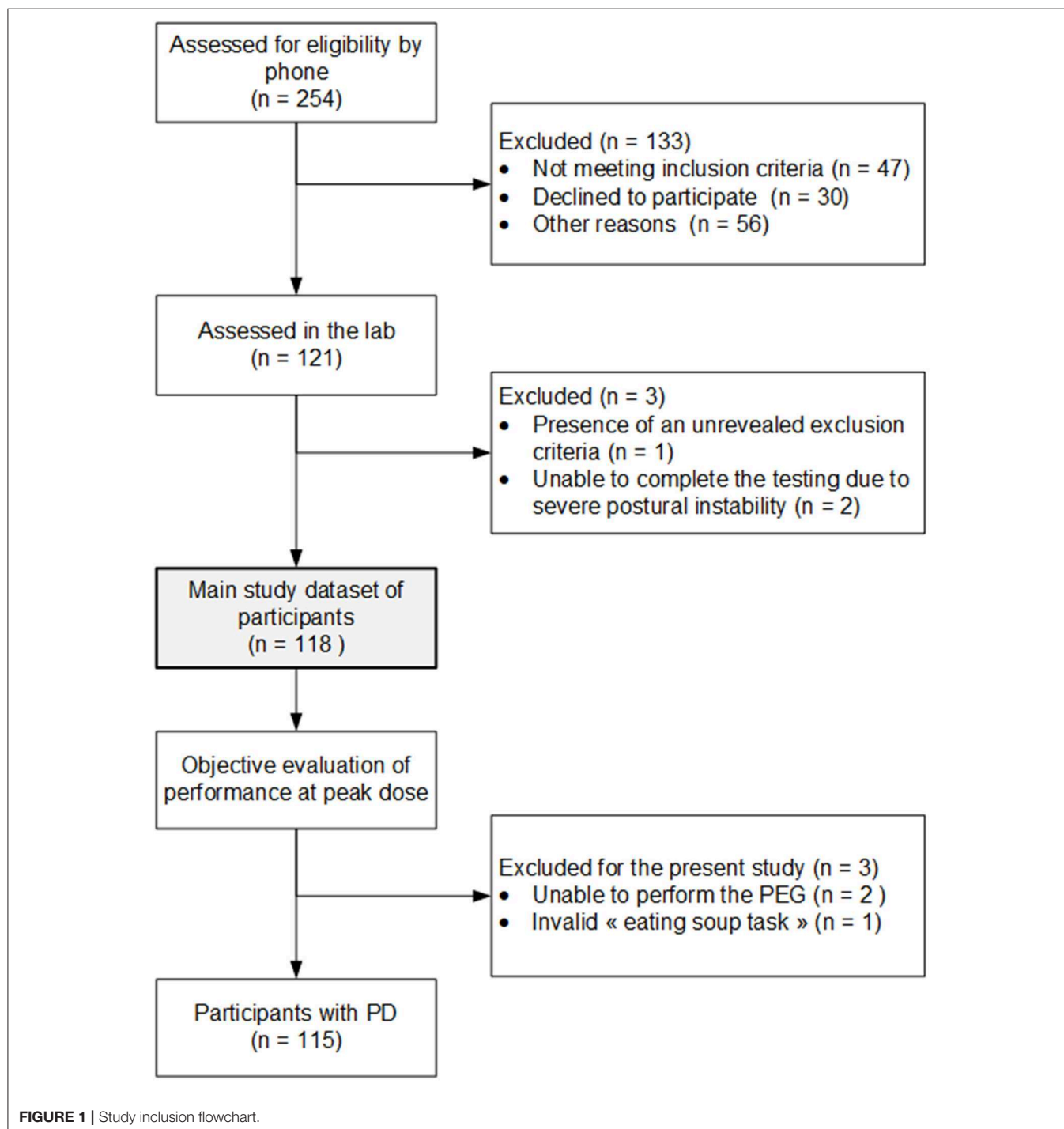
gait disorder group (Foltynie et al., 2002). Some studies also recognize an indeterminate subtype to PD, while others propose further sub-groups such as axial dominant, appendicular dominant, and rigidity dominant (Eisinger et al., 2017). In other words, classical approaches for PD phenotyping is mainly based on an *a priori* hypothesis of the importance of the dominant motor symptom on the patient's ability to perform ADL. Yet, patients with PD are often affected by more than one symptom. Combination of symptoms may exacerbate mobility issues or limit the efficacy of compensatory strategies. Furthermore, recent studies have outlined the impact of non-motor symptoms on the patients' ability to perform various ADL (Berganzo et al., 2016). The heterogeneity of the clinical profiles associated with PD therefore result in an unclear relationship between the traditional PD subgroups and the patients' proficiency in ADL. Thus, it appears desirable to revise this classification to allow a better correspondence with the treatment goal. One way to do so is to redirect the sub-typing approach toward an understanding of the functional impact of a patient's symptomatology on its global motor repertoire. Functional impact of a disease is traditionally assessed using questionnaires (Shulman et al., 2016). However, self-reported questionnaire are inclined to over or under-estimation of the patient's ability to perform activities and may suffer from flooring effect, as recently demonstrated by Regnault et al. (2019) for early Parkinson's disease. In an attempt to shed light on this type of issue, our lab has been working on developing methodologies to assess and objectively quantify symptoms and motor skills performance to better understand the relationship between PD symptoms and motor skills performance. We herein set to explore the capacity to assess the impact of the different symptoms on the motor skills repertoire in a global fashion. Specifically, this study aimed at: (1) exploring motor skill performance profiles in patients with PD; (2) identifying the factors (in our case symptoms) influencing the affiliation with a specific motor performance profile; and (3) verifying the possibility to create a model allowing to predict the motor performance profile based on the symptomatology.

METHODS

Participants

Data were extracted from a cohort of 121 patients diagnosed with PD. These participants were recruited in collaboration with the Quebec Parkinson Network and the Movement Disorders Clinic of the University of Calgary. Inclusion criteria consisted of a valid PD diagnosis given by a neurologist based on the UK Parkinson's Disease Society Brain Bank clinical diagnosis criteria (Hughes et al., 1992). Patients requiring assistance to walk, having an orthopedic condition that could hinder the performance of the tasks, as well as patients with a psychosis, were all excluded from the study. Previous publications using the data bank focused on the concomitant presence of cardinal symptoms of PD with dyskinesia (Goubault et al., 2018), as well as the influence of dyskinesia on motor performance (Goubault et al., 2019). For the present study, six additional participants were excluded as detailed in **Figure 1**. As a result, a sample of 115 patients, described in **Table 1**, was considered for the present study.

TABLE 1 | Study participants description.

	Healthy elderly (*n* = 69)	Parkinson's disease patients (*n* = 115)	*p*
GENERAL			
Gender (% male)	56.5%	58.3%	$P = 0.8779$
Age (yr)	68.1 ± 7.7	67.0 ± 8.7	$P = 0.4246$
Height (m)	1.67 ± 0.09	1.69 ± 0.10	$P = 0.4804$
Weight (Kg)	71.3 ± 14.5	69.9 ± 13.1	$P = 0.6188$
BMI	25.3 ± 3.9	24.5 ± 4.0	$P = 0.1928$
MMSE (/30)	28.6 ± 1.5	27.3 ± 2.5	$P < 0.001$
DISEASE INFO			
H&Y	–	1: 22.6% 3: 15.7% 2: 53.9% 4: 7.8%	–
UPDRS gait	–	1.1 ± 0.9	–
UPDRS freezing of gait	–	0.3 ± 0.8	–
UPDRS postural stability	–	1.1 ± 1.0	–
UPDRS posture	–	0.8 ± 1.0	–
UPDRS global spontaneity of movement	–	1.1 ± 1.0	–
UPDRS Postural tremor	–	0.5 ± 0.9	–
UPDRS rest tremor	–	0.2 ± 0.6	–
UPDRS rigidity	–	*Arms*: 0.7 ± 0.7	–
		Legs: 1.1 ± 0.7	–
Years since diagnosis	–	10.5 ± 5.8	–
MEDICATION			
LEDD	–	1029.1 ± 509.2**	–
Levodopa (%)	–	100*	–
Agonist (%)	–	32.7*	–
Amantadine (%)	–	39.8*	–
COMT or MAOB (%)	–	49.6*	–

BMI, Body Mass Index; MMSE, Mini-Mental State Exam; H&Y, Hoehn and Yahr scale; UPDRS, Unified Parkinson's Disease Rating Scale; LEDD, Levodopa Equivalent Daily Dose.
*Missing medication profile for 2 participants; **Missing info for 6 participants.

A second group of participants composed of 69 age and gender-matched community-dwelling elderly (43.5% female, age = 68.1 ± 7.7 years old, BMI = 25.3 ± 3.9) was also recruited through the Center de Recherche de l'Institut universitaire de gériatrie de Montréal (CRIUGM) to provide control data. The study protocol was approved by both the CRIUGM and the Conjoint Health Research ethics boards, and all participants provided written informed consent.

Experimental Protocol

The experimental protocol has been described in detail previously (Goubault et al., 2018, 2019). In brief, participants were tested on their regular medication and equipped with an inertial suit containing 17 sensors (IGS-180, Synertial Ltd, UK), allowing the capture of the entire body kinematics. Each sensor is composed of a 3-axis accelerometer, measuring linear acceleration, a 3-axis gyroscope, assessing angular velocity, and a 3-axis magnetometer. Upon arrival to the lab, participants

took their medication and were asked to fill-up a socio-demographic questionnaire as well as cognitive and quality of life questionnaires. The study's objective data acquisition process was then divided into two blocks of ADL, nested between blocks of symptoms evaluation. The chosen ADL included a variety of activities corresponding to a wide range in velocity and amplitude of motion (upper and/or lower limbs), in order to represent the complete motor repertoire. Tasks were selected from a variety of ADL and instrumented ADL scales (Klein and Bell, 1982; Fahn et al., 1987; Lozano et al., 1995; Boraud et al., 2001; Health Canada Interdepartmental Committee on Aging Seniors Issues, 2001; Krystkowiak and Defebvre, 2002; Guttman et al., 2003; Health Canada/Parkinson Society Canada, 2003; Goetz et al., 2008; Colosimo et al., 2010; Carignan et al., 2011). Chosen tasks included eating soup, cutting and eating a piece of apple, taking pills, drinking a glass of water, counting money, reading a book out loud, reaching to grab an object on the ground, rising from a chair, walking, turning, sitting down (Timed-Up and Go, TUG), and the Purdue Pegboard task. Participants were cued to initiate the task when a light, positioned in front of them, turned on. For this specific study, a subset of tasks was considered in order to limit the degrees of freedom in the analysis. The selected tasks included the TUG, eating soup (ES), and the Purdue Pegboard test (PB). While ES involves short range, slow speed movements, PB requires short range and fast motion, while the TUG relates to long range, medium speed global motion. For ES, participants sat down on a bench with both hands flat on the table. Once cued by the light, participants were instructed to take the spoon positioned on the table using their dominant hand, take four spoons of water at their preferred pace to reproduce true living conditions, return the spoon to the table, and position their hands back on the table. The time required to performed the task corresponds to the time elapsed between the light stimuli and the time the hands are placed back on the table. For PB, a board with two parallel rows composed of 25 holes each was placed in front of the participant. Upon signal, participants were instructed to insert as many pins in the holes as possible in 30 s, using both hands alternately. The TUG was initiated with the participant sitting on a bench. Upon signal (i.e., light), the participant was asked to rise from the bench without any help if possible (i.e., no hands), walk for 3 m at their preferred pace, and return to their initial sitting position. Performance was assessed using the time required to perform the task (ES, TUG), the count achieved (PB), and the number of errors committed (ES: dropping water, dropping the spoon; TUG: needing assistance, using hands to rise/sit; PB: dropping pins).

The symptoms assessment blocks consisted of a mixture of clinical evaluation and objective assessment of the symptomatology: postural instability was assessed using the pull-back test (Unified Parkinson's Disease Rating Scale item 3.12), rigidity was evaluated manually for each limb (item #3.3), bradykinesia was appraised objectively using a rapid alternating task, while tremor, drug-induced dyskinesia (DID) and freezing of gait (FoG) were all assessed objectively during appropriate tasks using inertial data (Goubault et al., 2018). Briefly, tremor was assessed using the signal captured by the gyroscopes positioned on the hands. These signals were band-pass filtered between 3 and 7.5 Hz to isolate the tremor frequency range. A power density spectrum was then used to identify the signal dominant frequency, as well as its dispersion. Tremor was detected when dispersion was below 2 Hz, in which case the corresponding tremor value was fixed to the dispersion bandwidth. DID was assessed during the tasks, using signals from the sensors not directly involved in the specific task. Signals from the gyroscopes were again band-pass filtered, this time between 0.5 and 4 Hz. The energy of the resulting signal was then computed, per segment. The average energy among the different segments considered corresponds to the DID value attributed for the task. Freezing of gait was assessed during the walking portion of the TUG. The process uses the ratio of the power of the signal within the walking bandwidth to the power located within the freezing bandwidth to identify freezing events.

Performance Clusters Identification

A clustering approach was used to explore the presence of motor skills performance profiles within a group of patients medicated for PD. This method allows the groups to emerge directly from the data without bias (Rui and Wunsch, 2005). In this specific case, performance clusters were based on five metrics extracted from three selected tasks: TUG time, TUG errors, Eating soup time, Eating soup errors, and Pegboard number of pins. To ensure all metrics have a similar influence during the clustering process, timing features as well as the Pegboard pins count were first normalized based on the control group performance data. Extreme values, defined as values outside the ± 4 Z-score, were also set to the closest valid limit.

Clustering was performed using the K-means method. In brief, this approach uses an iterative process to minimize the sum of the distances between each point and its cluster's centroid, while maximizing the difference between the clusters (Rui and Wunsch, 2005). This method, however, requires the user to specify the desired number of clusters. We defined the ideal number of clusters as a trade-off between the sum of the Euclidean distance between each point and its cluster's centroid and the resulting number of very small clusters, herein defined as groups composed of fewer than 10 participants. In other words, the clustering process was performed using a varying number of clusters, from 1 to 115 (the number of participants), and the quality of the resulting clusters was evaluated based on both the distance cost and the resulting number of small clusters, to identify the optimal number of clusters. The ability of the clusters to differentiate performance was then evaluated using a Kruskall-Wallis ANOVA test. The clustering and validation processes were performed in Matlab Release 2018a (The MathWorks, Massachusetts, United States).

Performance Profiles Features Identification & Membership Prediction

The second objective of this study consists in analyzing which features, amongst the motor and the non-motor symptoms as well as the participants' characteristics, explain the affiliation to a specific motor performance profile or cluster. To do so, symptomatology was first normalized based on the control group

FIGURE 2 | Performance clusters identification. **(A)** Visual representation of the clusters based on a subset of 3 factors. **(B)** Boxplots of the distribution of the different performance factors within the clusters. The *p*-values correspond to the result of the Kruskall-Wallis ANOVA test.
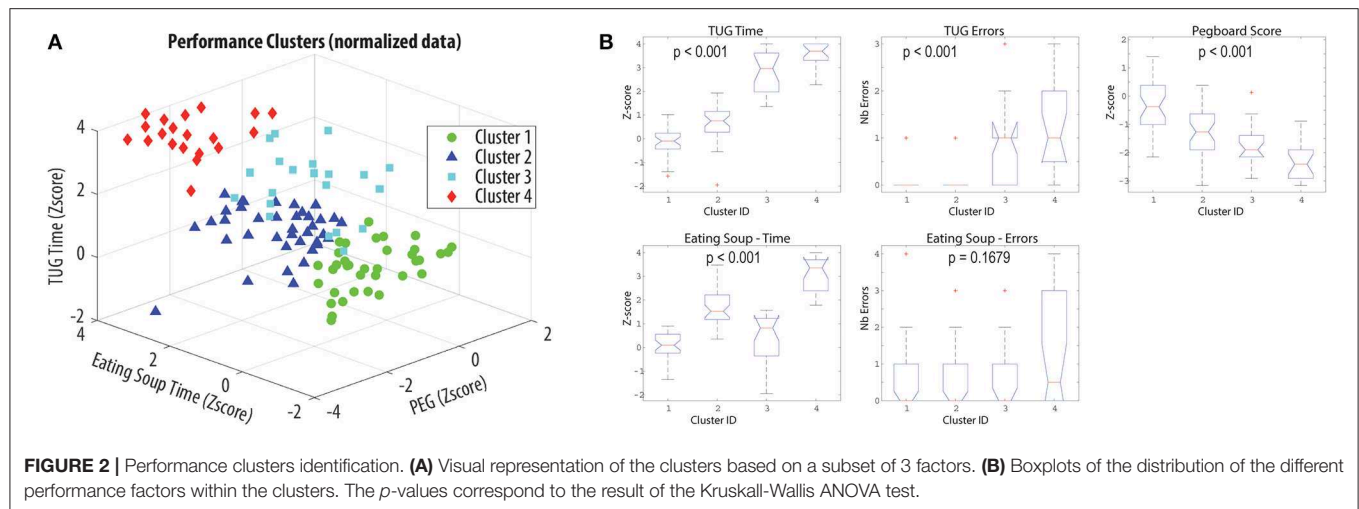
data acquired. Then, the sample was divided into a training and a validation datasets (80–20%). Using the training dataset, univariate multinomial regression analysis was performed on each variable, that is age, gender, BMI, years living with PD, Mini-Mental State Exam (MMSE), symptomatology (dyskinesia, bradykinesia, rest tremor, postural tremor, kinetic tremor, rigidity, postural instability, freezing of gait), and medication regimen [Levodopa Equivalent Daily Dose (LEDD), Levodopa, Agonist, Amantadine, COMT or MAOB]. All variables with a marginal significance (i.e., *p*-value) smaller or equal to 0.2 were identified as potentially explicative variables (PEV) for a specific cluster membership. A multivariate multinomial regression analysis was then performed using these PEV. The model was designed using 80% of the sample; and verified with the remaining 20%. The accuracy of the proposed model was then evaluated based on a contingency table. All statistical analyses were performed using SPSS v23 (IBM Corp., Armonk, NY).

RESULTS

Clusters Identification Results

Four clusters of performance were identified (**Figure 2A**) and confirmed by statistical analyses. Only the number of errors made while eating soup was not shown to be a discriminative factor (**Figure 2B**). As detailed in **Table 2**, Cluster 1 is composed of participants who performed within normal for all tasks and parameters. Cluster 2 corresponds to participants slightly affected in fine motor tasks. Cluster 3 is made of participants mainly affected during the TUG, while the last cluster is composed of participants affected in all activities.

Performance Features Identification Results

Cluster membership was attributed to each participant, following the process described in section Performance Profiles Features Identification & Membership Prediction. The resulting portrait of the patients' symptomatology profiles, per cluster, is reported in **Table 3**.

TABLE 2 | Clusters performance details.

	Cluster ID				*p*
	1	**2**	**3**	**4**	
TUG time (s) *Median [Q1,Q3]*	13.0 [12.3, 13.6] **Not affected**	14.6 [13.7, 15.5] **Not Affected**	20.2 [17.5, 22.2] **Affected**	22.4 [21.2, 25.0] **Affected**	*p* < 0.001
TUG err *Median [Q1,Q3]*	0 [0, 0] **Not affected**	0 [0, 0] **Not affected**	1.0 [0, 1.0] **Affected**	1.0 [0.5, 2.0] **Affected**	*p* < 0.001
Pegboard #pins *Median [Q1,Q3]*	15.0 [12.5, 18.0] **Not affected**	11.5 [9.0, 14.0] **Slightly affected**	9.0 [8.0, 11.0] **Slightly affected**	7.0 [5.0, 9.0] **Affected**	*p* < 0.001
Eating Soup time (s) *Median [Q1,Q3]*	18.9 [17.9, 20.4] **Not affected**	23.9 [22.6, 26.8] **Slightly affected**	21.3 [17.5, 22.8] **Not affected**	32.3 [27.6, 34.3] **Affected**	*p* < 0.001
Eating Soup Errors *Median [Q1,Q3]*	0 [0, 1]	0 [0, 1]	0 [0, 1]	0.5 [0, 3]	*p* = 0.1679

Univariate multinomial analysis performed on this set of data allowed to identify 10 potentially explanatory variables: age (*p* = 0.134), MMSE (*p* = 0.200), dyskinesia (*p* < 0.001), bradykinesia (*p* < 0.001), rest tremor (*p* = 0.024), kinetic tremor (*p* = 0.010), rigidity (*p* = 0.010), postural instability (*p* < 0.001), freezing of gait (*p* < 0.001), and the presence of Amantadine in the medication regimen (*p* = 0.180). Including all these potentially explanatory variables into a single multinomial regression allowed to derive a significant model (χ^2 = 140.628, *p* < 0.001) with a good representativeness (Nagelkerke pseudo R^2 = 0.839). This global model identified postural instability (*p* < 0.001), dyskinesia (*p* = 0.024), bradykinesia (*p* = 0.022), rigidity (*p* = 0.026), freezing of gait (*p* = 0.040), as well as Amantadine (*p* = 0.003) as the main significant variables, while cognitive impairment (*p* = 0.064) and rest tremor (*p* = 0.086) significantly discriminates between sub-groups 3 and 1 despite being globally significant.

TABLE 3 | Patients symptomatology portrait per performance cluster.

	Cluster ID			
	1 (n = 36)	**2** (n = 38)	**3** (n = 21)	**4** (n = 20)
GENERAL				
Gender (% male)	52.8%	68.4%	61.9%	45.0%
Age (yr)	64.0 ± 8.7	66.5 ± 8.4	70.0 ± 8.7	70.2 ± 7.5
BMI	23.9 ± 3.3	24.0 ± 3.8	24.9 ± 3.4	26.0 ± 5.5
MMSE (/30)	28.2 ± 1.5	27.9 ± 2.0	26.5 ± 2.6	25.3 ± 3.4
DISEASE INFO				
H&Y	1: 41.7%	1: 23.7%	1: 9.5%	1: 0.0%
	2: 55.6%	2: 68.4%	2: 33.3%	2: 45.0%
	3: 2.8%	3: 7.9%	3: 47.6%	3: 20.0%
	4: 0.0%	4: 0.0%	4: 9.5%	4: 35.0%
Years since diagnosis	8.8 ± 3.7	10.6 ± 5.9	11.8 ± 7.7	12.2 ± 6.4
MEDICATION				
LEDD	994.9 ± 450.8	866.0 ± 448.1	1378.8 ± 567.4	1051.5 ± 523.7
Levodopa (%)	100%	100%	100%	100%
Agonist (%)	38.9%	35.1%	35.0%	15.0%
Amantadine (%)	50.0%	35.1%	15.0%	55.0%
COMT or MAOB(%)	47.2%	59.5%	40.0%	45.0%
SYMPTOMS				
Dyskinesia (normalized value)	1.1 ± 1.6 [−1.7, 4.2]	−0.5 ± 1.4 [−2.7, 2.5]	0.6 ± 2.1 [−3.2, 3.2]	−0.8 ± 2.2 [−4.1, 4.1]
Bradykinesia (normalized value)	−1.1 ± 1.1 [−3.9, 0.7]	−1.9 ± 1.8 [−6.0, 1.1]	−1.7 ± 1.6 [−4.6, 1.1]	−3.6 ± 1.5 [−6.0, −0.9]
Rest tremor (normalized value)	0.3 ± 4.1 [−6.0, 6.0]	0.7 ± 3.9 [−6.0, 6.0]	1.1 ± 4.3 [−6.0, 6.0]	2.8 ± 3.2 [−2.3, 6.0]
Postural tremor (normalized value)	[1.6, 6.0]	[1.6, 6.0]	[1.6, 6.0]	[1.6, 6.0]
Kinetic tremor (normalized value)	0.5 ± 1.8 [−2.8, 6.4]	−0.1 ± 1.8 [−3.0, 4.9]	1.6 ± 2.1 [−2.1, 5.5]	1.3 ± 2.3 [−3.5, 6.0]
Postural instability	0.7 ± 0.7 [0.0, 2.5]	0.7 ± 0.7 [0.0, 1.0]	2.1 ± 1.2 [0.0, 4.0]	1.9 ± 1.0 [1.0, 4.0]
Freezing (%)	0.0 ± 0.0 [0.0, 0.0]	0.2 ± 1.2 [0.0, 2.0]	5.0 ± 15.7 [0.0, 71.6]	3.8 ± 9.0 [0.0, 36.6]
Rigidity	0.8 ± 0.6 [0, 2.5]	0.9 ± 0.6 [0, 2.0]	0.8 ± 0.8 [0, 2.5]	1.3 ± 0.6 [0.2, 2.5]

Detailed analysis of the model revealed that:

- an increase in postural instability increases the chance to be part of cluster 3 or 4, relative to cluster 1 or 2 ($p_{3rel1} = 0.001$, $OR_{3rel1} = 9.323$ [2.430, 35.773]; $p_{3rel2} = 0.001$, $OR_{3rel2} = 6.785$ [2.107, 21.851]; $p_{4rel1} = 0.009$, $OR_{4rel1} = 6.268$ [1.574, 24.957]; $p_{4rel2} = 0.012$, $OR_{4rel2} = 4.561$ [1.399, 14.868]);
- an increase of one standard deviation in dyskinesia level increases the chance to be in cluster 1 compared to cluster 2 or 3 ($p_{1rel2} = 0.014$, OR_{1rel2}: 2.12 [1.17, 3.86]; $p_{1rel3} = 0.023$, $OR_{1rel3} = 2.92$ [1.16, 7.30]);
- an increase of one standard deviation in bradykinesia level increases the likelihood of being in cluster 4 relative to cluster 3 ($p = 0.025$, OR = 6.06 [1.26, 29.41]);
- an increase in rigidity increases the chance to be in cluster 4 relative to cluster 1 ($p = 0.025$; OR = 34.17 [1.54, 757.05]) and cluster 2 ($p = 0.036$, OR = 24.97 [1.23, 506.53]);
- the presence of Amantadine in the medication regimen appears to decrease the risk of being in cluster 3, when compared to cluster 1 or 2 ($p_{3rel1} = 0.025$, $OR_{3rel1} = 3.22\text{E-}4$

TABLE 4 | Contingency table.

		Predicted cluster				
		1	**2**	**3**	**4**	**% Correct**
Observed cluster	1	27	7	1	1	75.0%
	2	9	26	0	2	70.3%
	3	1	2	16	1	80.0%
	4	1	2	0	17	85.0%
Overall percentage		33.6%	32.7%	15.0%	18.6%	76.1%

[2.92E-7, 0.356]; $p_{3rel2} = 0.044$, $OR_{3rel2} = 0.001$ [7.30E-7, 0.825]).

This model allowed to classify the participants within their respective cluster of performance with an accuracy of 76%, as illustrated in the contingency table (**Table 4**).

DISCUSSION

This study first aimed at investigating the presence of motor skills performance profiles in patients medicated for PD. Using a clustering approach, four different profiles emerged from the data. Analyzing the variation in metrics within each cluster revealed that cluster 1 is composed of participants who are not affected in the motor tasks assessed under medication. Cluster 2 participants are affected only slightly in fine motor tasks. Cluster 3 participants are mainly affected in mobility tasks, while cluster 4 involves participants affected in all areas. These clusters were shown to be statistically different for four performance metrics out of five, demonstrating the potential of the method. This approach offers an innovative view for PD classification, focussed on the global impact of the disease on the patient's motor repertoire as opposed to a more classical dominant symptom classification (Foltynie et al., 2002; Eisinger et al., 2017, 2019; Erro et al., 2019). To our knowledge, this study is the first to address the phenotype problematic from this point of view. Direct comparison between the two classification approaches would be worth investigating. Nevertheless, it is clear from the description of the symptomatology profile per cluster that symptoms coexist within the clusters. This observation supports a global approach of symptomatology characterization for motor performance prediction.

Although the clusters identified are statistically significant and appear to hold a clinical meaning, it shall be noted that the clustering method could be further refined. Indeed, the K-mean method requires the user to determine in advance the number of desired clusters. In order to remain as objective as possible, we first investigated different potential avenues for clusters quantity identification, such as the use of the silhouette validity index and the Calinski-Harabasz index (Arbelaitz et al., 2013). However, Hennig (2015) exposed an interesting way of looking at true clusters based on the direct aim pursued by the clustering process. Indeed, the idea for true or ideal clusters may vary depending on the situation. In the current study, we know

that the optimal number of clusters represents different mobility profiles, however somehow subtle these differences may be. As such, the dissimilarity between clusters criterion may not be obvious, and thus, the classic validity indexes may not be optimal. As such, we identified the ideal number of clusters as a trade-off between the within clusters similarity and the number of small clusters created. The pragmatic approach was appropriate for the current study, but may need to be revisited in other cases.

The second part of the study aimed at exploring potential factors influencing the affiliation of a participant to a specific cluster of performance in the "ON" medicated state. It was shown that postural instability, dyskinesia, bradykinesia, rigidity, freezing of gait, and Amantadine all play a significant role in the classification process. Consistent with the literature, postural instability and freezing of gait discriminated patients with disabled mobility (Muslimovic et al., 2008; Goubault et al., 2019). Unsurprisingly, an increase in bradykinesia raised the risk to be affected in fine motor tasks, but the model suggests that this is true only for the subset of the sample also affected in mobility tasks. Indeed, bradykinesia by itself did not come out as a significant factor to discriminate participants with normal performance and participants slightly affected in fine motor tasks (i.e., clusters 1 and 2). However, an increase in residual (i.e., on medication) bradykinesia increased the likelihood of being affected in all domains as opposed to being affected only in mobility tasks, suggesting that this factor is more relevant to appendicular rather than axial motor control. Only dyskinesia came out as a significant factor in the differentiation between patients with normal performance and patients slightly affected in fine motor tasks. Indeed, the present way of analyzing this data confirms what has been described in previous studies (Goubault et al., 2019) using the same patients that dyskinesia increases, to a certain extent, the likelihood of being in the *normal* performance group when compared to the group slightly affected in fine motor tasks or the group affected in mobility. We acknowledge the fact that few patients displayed severe dyskinesia in the current sample, but low to moderate levels of dyskinesia certainly did not interfere with the patients' performance. Results have also demonstrated that when all other symptoms are equivalent, the addition of Amantadine in the medication regimen decreases the risk of being part of the cluster affected in mobility task, when compared to the normal performance group. These results are concordant with the effect of Amantadine on gait in PD patients under deep-brain stimulation [16]. Yet, the impact of Amantadine on gait is still unclear [17], as well as the fraction of benefit that may derive from the reduction in levodopa daily dose afforded by this drug. Cognitive impairment did not come out as a global significant factor, but it did have a significant impact in differencing people with disabled mobility.

It is worth mentioning that the reported results could have been different if patients were tested in their OFF state. Indeed, all patients were tested at peak dose, assuming medication was optimal. The reported impact of the different symptoms on the performance cluster affiliation therefore refers to the effect of the residual symptoms. Further studies should consider running similar analyses ON and OFF states to assess not only the direct impact of the symptoms, but also to bring one step further the

analysis of the medication's impact instead of only considering the number of years since diagnosis in the analyses. Another limit to the current study regards the subset of tasks used for the analysis. Future work will focus on applying a similar protocol on the entire set of tasks collected.

The statistical model developed using the global patient symptomatology allowed to predict the impact of the disease on the patients' motor repertoire with an accuracy of 76%. The model was specifically good at recognizing patients with mobility and global issues (i.e., clusters 3 and 4). Such results demonstrate the strength of the global approach, although future work should investigate other classification approaches to improve the overall accuracy. For example, machine learning approaches with a K-fold cross-validation loop could improve the accuracy of the classification process. The general approach also needs to be tested on a much larger group of patients and by using traditional clinical testing to render it more usable. We could then be able to determine, based on that evaluation, what will be the impact of the symptomatology of the patient's ADL, and as such predict their ability to perform everyday tasks.

CONCLUSION

PD affects the motor repertoire of patients to different extents. This study demonstrated that four major performance profiles appear to exist: patients with normal performance, patients affected slightly in fine motor tasks, patients affected in mobility tasks and patients affected in all domains of mobility. This study demonstrated that it is possible to predict the mobility performance of any patient, based on personal clinical features. Although future research is needed to refine the clustering method, as well as performance prediction suiting clinical evaluations, these results appear promising, and may lead to more personalized treatment by identifying and targeting symptoms that specifically impede a particular patient's motor performance.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comité d'éthique de la recherche vieillissement-neuroimagerie du CIUSSS du Centre-Sud-de-l'Île-de-Montréal. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KL developed the algorithms, designed the analysis, and drafted the manuscript. EG and SB collected the data and provided significant feedback on the study analysis and the paper.

CD conceived the experiment, helped in data interpretation, and reviewed the paper. PB was significantly involved in the interpretation of the data and the review of the manuscript.

REFERENCES

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recogn.* 46, 243–256. doi: 10.1016/j.patcog.2012.07.021

Berganzo, K., Tijero, B., González-Eizaguirre, A., Somme, J., Lezcano, E., Gabilondo, I., et al. (2016). Motor and non-motor symptoms of Parkinson's disease and their impact on quality of life and on different clinical subgroups. *Neurología* 31, 585–591. doi: 10.1016/j.nrleng.2014.10.016

Boraud, T., Bezard, E., Bioulac, B., and Gross, C. E. (2001). Dopamine agonist-induced dyskinesias are correlated to both firing pattern and frequency alterations of pallidal neurones in the MPTP-treated monkey. *Brain* 124, 546–557. doi: 10.1093/brain/124.3.546

Carignan, B., Daneault, J.-F., and Duval, C. (2011). Assessing drug-induced Dyskinesia in the clinic, the laboratory and the natural environment of patients. *J. Park. Dis.* 1, 329–337. doi: 10.3233/JPD-2011-11054

Colosimo, C., Martinez-Martin, P., Fabbrini, G., Hauser, R. A., Merello, M., Miyasaki, J., et al. (2010). Task force report on scales to assess dyskinesia in Parkinson's disease: critique and recommendations. *Mov. Disord.* 25, 1131–1142. doi: 10.1002/mds.23072

Eisinger, R. S., Hess, C. W., Martinez-Ramirez, D., Almeida, L., Foote, K. D., Okun, M. S., et al. (2017). Motor subtype changes in early Parkinson's disease. *Parkinsonism Relat. Disord.* 43, 67–72. doi: 10.1016/j.parkreldis.2017.07.018

Eisinger, R. S., Martinez-Ramirez, D., Ramirez-Zamora, A., Hess, C. W., Almeida, L., Okun, M. S., et al. (2019). Parkinson's disease motor subtype changes during 20 years of follow-up. *Parkinsonism Relat. Disord.* doi: 10.1016/j.parkreldis.2019.05.024

Erro, R., Picillo, M., Amboni, M., Savastano, R., Scannapieco, S., Cuoco, S., et al. (2019). Comparing postural instability and gait disorder and akinetic-rigid subtyping of Parkinson disease and their stability over time. *Eur J Neurol.* 1212–1218. doi: 10.1111/ene.13968

Fahn, S., Elton, R. L., and Updrs Program Members (1987). "Unified Parkinsons rating scale," in *Recent Developments in Parkinsons Disease,* eds S. Fahn, C. D. Marsden, M. Goldstein, and D. B. Calne (Florham Park: Macmillan Healthcare Information), 153–163.

Foltynie, T., Brayne, C., and Barker, R. A. (2002). The heterogeneity of idiopathic Parkinson's disease. *J. Neurol.* 249, 138–145. doi: 10.1007/PL00007856

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., et al. (2008). Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord.* 23, 2129–2170. doi: 10.1002/mds.22340

Goubault, E., Nguyen, H. P., Bogard, S., Blanchet, P. J., Bezard, E., Vincent, C., et al. (2018). Cardinal motor features of Parkinson's disease coexist with peak-dose choreic-type drug-induced dyskinesia. *J. Parkinsons Dis.* 8, 323–331. doi: 10.3233/JPD-181312

Goubault, E., Nguyen, H. P., Bogard, S., Blanchet, P. J., Bézard, E., Vincent, C., et al. (2019). Remnants of cardinal symptoms of Parkinson's disease, not dyskinesia, are problematic for dyskinetic patients performing activities of daily living. *Front. Neurol.* 10:256. doi: 10.3389/fneur.2019.00256

Guttman, M., Slaughter, P. M., Theriault, M. E., Deboer, D. P., and Naylor, C. D. (2003). Burden of parkinsonism: a population-based study. *Mov. Disord.* 18, 313–319. doi: 10.1002/mds.10333

Health Canada and Interdepartmental Committee on Aging and Seniors Issues (2001). *Canada's Aging Population.* 3.

Health Canada/Parkinson Society Canada. (2003). *Parkinson's Disease: Social and Economic Impact.* 1–6. Available online at: https://www.parkinson.ca/wp-content/uploads/PD_Social-Econ_Impact_EN.pdf

Hennig, C. (2015). What are the true clusters? *Pattern Recogn. Lett.* 64, 53–62. doi: 10.1016/j.patrec.2015.04.009

Hughes, A. J., Daniel, S. E., Kilford, L., and Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* 55, 181–184. doi: 10.1136/jnnp.55.3.181

Klein, R. M., and Bell, B. (1982). Self-care skills: behavioral measurement with Klein-Bell ADL scale. *Arch. Phys. Med. Rehabil.* 63, 335–338.

Krystkowiak, P., and Defebvre, L. (2002). [Rush dyskinesia scale or Goetz scale]. *Rev. Neurol.* (Paris) 158, 622–624.

Lozano, A. M., Lang, A. E., Galvez-Jimenez, N., Miyasaki, J., Duff, J., Hutchinson, W. D., et al. (1995). Effect of GPi pallidotomy on motor function in Parkinson's disease. *Lancet* 346, 1383–1387. doi: 10.1016/S0140-6736(95)92404-3

Muslimovic, D., Post, B., Speelman, J. D., Schmand, B., and De Haan, R. J. (2008). Determinants of disability and quality of life in mild to moderate Parkinson disease. *Neurology* 70, 2241–2247. doi: 10.1212/01.wnl.0000313835.33830.80

Regnault, A., Boroojerdi, B., Meunier, J., Bani, M., Morel, T., and Cano, S. (2019). Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *J. Neurol.* 266, 1927–1936. doi: 10.1007/s00415-019-09348-3

Rui, X., and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Netw.* 16, 645–678. doi: 10.1109/TNN.2005.845141

Shulman, L. M., Armstrong, M., Ellis, T., Gruber-Baldini, A., Horak, F., Nieuwboer, A., et al. (2016). Disability rating scales in parkinson's disease: critique and recommendations. *Mov. Disord.* 31, 1455–1465. doi: 10.1002/mds.26649

Sveinbjornsdottir, S. (2016). The clinical symptoms of Parkinson's disease. *J. Neurochem.* 139(Suppl. 1), 318–324. doi: 10.1111/jnc.13691

Statistical Modeling of Lower Limb Kinetics During Deep Squat and Forward Lunge

Joris De Roeck[1], J. Van Houcke[1], D. Almeida[2], P. Galibarov[3], L. De Roeck[1] and Emmanuel A. Audenaert[1,4,5,6]*

[1] Department of Human Structure and Repair, Ghent University, Ghent, Belgium, [2] Centre for Rapid and Sustainable Product Development, Polytechnic Institute of Leiria, Leiria, Portugal, [3] AnyBody Technology A/S, Aalborg, Denmark, [4] Department of Orthopaedic Surgery and Traumatology, Ghent University Hospital, Ghent, Belgium, [5] Department of Trauma and Orthopaedics, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom, [6] Department of Electromechanics, Op3Mech Research Group, University of Antwerp, Antwerp, Belgium

Purpose: Modern statistics and higher computational power have opened novel possibilities to complex data analysis. While gait has been the utmost described motion in quantitative human motion analysis, descriptions of more challenging movements like the squat or lunge are currently lacking in the literature. The hip and knee joints are exposed to high forces and cause high morbidity and costs. Pre-surgical kinetic data acquisition on a patient-specific anatomy is also scarce in the literature. Studying the normal inter-patient kinetic variability may lead to other comparable studies to initiate more personalized therapies within the orthopedics.

Methods: Trials are performed by 50 healthy young males who were not overweight and approximately of the same age and activity level. Spatial marker trajectories and ground reaction force registrations are imported into the Anybody Modeling System based on subject-specific geometry and the state-of-the-art TLEM 2.0 dataset. Hip and knee joint reaction forces were obtained by a simulation with an inverse dynamics approach. With these forces, a statistical model that accounts for inter-subject variability was created. For this, we applied a principal component analysis in order to enable variance decomposition. This way, noise can be rejected and we still contemplate all waveform data, instead of using deduced spatiotemporal parameters like peak flexion or stride length as done in many gait analyses. In addition, this current paper is, to the authors' knowledge, the first to investigate the generalization of a kinetic model data toward the population.

Results: Average knee reaction forces range up to 7.16 times body weight for the forwarded leg during lunge. Conversely, during squat, the load is evenly distributed. For both motions, a reliable and compact statistical model was created. In the lunge model, the first 12 modes accounts for 95.26% of inter-individual population variance. For the

maximal-depth squat, this was 95.69% for the first 14 modes. Model accuracies will increase when including more principal components.

Conclusion: Our model design was proved to be compact, accurate, and reliable. For models aimed at populations covering descriptive studies, the sample size must be at least 50.

Keywords: lower limb kinetics, inverse dynamics, principal component analysis, musculoskeletal model, validation analysis

INTRODUCTION

In biomechanics, the safety and efficiency of novel surgical techniques as well as the development of biocompatible products ultimately rely on its capability of being tested on humans through clinical trials. The complete development chain of a new surgical technique or implant and their introduction into clinic practice is both time-consuming and economically demanding. Next to it, it is known that patient-specific surgery planning or implant design can improve the long-time outcome of an implant (Pietsch et al., 2013; Spencer-Gardner et al., 2016). This fact is due to the high anatomical variability between individuals and the different functional activities, which have a significant effect in the ratio of the force components on the lower limb between subjects (Kutzner et al., 2010) and on the functional alignment of the prosthetic components of a lower limb implant (Smoger, 2016; Spencer-Gardner et al., 2016). Within this context, methodologies such as statistical models of the human anatomy as well as kinematics or kinetics that account for the anatomical inter-variability of the population combined with biomechanical simulation studies can provide non-invasive pre-surgical clinical output.

Lower limb kinetics can be estimated based on musculoskeletal models and ground force plates using inverse dynamics (Carbone et al., 2012; Galloway et al., 2012; Vaitkus and Várady, 2015; Bagwell et al., 2016). These techniques do not often account for patient-specific variability as they use scaled generic models (Worsley et al., 2011; Vaitkus and Várady, 2015), while it was already widely shown that the geometry of the musculoskeletal models is very sensitive to muscle force predictions (Carbone et al., 2012). In addition, and to the authors knowledge, the available studies merely consider very limited population samples which may not be representative of the total variability of the lower limb anatomy. Lastly, the available literature lacks completeness as, to date, no study has considered a statistical model of the full lower limb, namely, on demanding tasks such as the deep squat and the forward lunge.

Hence, in order to create the foundations for the development and optimization of the design or the durability of orthopedic implants, it is mandatory to generate appropriate loading conditions that represent inter-patient variability across the population (Honari and Taylor, 2013; Bischoff et al., 2014). Patient-specific finite element analyses are the state-of-the-art technique to infer quantitative information on a specific design or performance of an arthroscopic implant (Shu et al., 2018). Taylor et al. (2012) found most studies to be focusing on variations on the morphological and bone properties rather than the consequences of variability because of loading. Furthermore, it has been proved that the application of single-representative models can be extended to account for variability by either parametrically or probabilistically varying the loading/boundary conditions. These approaches allow model generation which can significantly extend the coverage of the anatomical variability and ultimately create a powerful tool to assess the performance of medical devices (Taylor et al., 2012).

Recent developments in medical imaging significantly increased the accuracy of the three-dimension computational anatomical representation, enhancing the anatomical differences within a determined population (Almeida et al., 2016; Audenaert et al., 2019). Hence, combining the use of magnetic resonance imaging (MRI) with musculoskeletal models will provide us an insight on lower limb kinematics on patient-specific anatomies. The statistical analysis of kinematic time series by means of dimensionality reduction techniques such as principal component analysis (PCA) or independent component analysis is not novel *per se* (Daffertshofer et al., 2004; Galloway et al., 2012), but the inclusion of patient-specific anatomies is believed to more accurately represent inter-patient kinetic variability. Such approach, hereby presented, will allow for a large population of kinetic data to be generated without the time and the expense of collecting the motion capture data of hundreds of patients. Simultaneously, it will open the door to the generation of large simulated populations for use in clinical outcome simulation studies, injury biomechanics, musculoskeletal disease models, or implant design optimization (Henak et al., 2013; Zhang et al., 2016).

While the gait cycle has been the most researched activity in the current literature, it is not particularly demanding for the lower limb joints. For the purpose of implant wear testing, implant fixation, and joint stability analysis, there are other more challenging activities commonly performed in daily living that might be of particular interest (Hartmann et al., 2013). Clinical, experimental, and computational studies have clearly reported increased complication risk and wear rate under high contact stress conditions (Kang et al., 2008; O'Brien et al., 2015; de Ruiter et al., 2017).

In sum, the purpose of this study is to build statistical models of deep squatting and forward lunging for applications in pre-clinical testing of orthopedic implants and surgery in an asymptomatic adult population and ultimately to analyze and validate the inter-individual variations in lower limb kinetics.

MATERIALS AND METHODS

Participants

Fifty-three asymptomatic volunteers participated in the study. In order to eliminate sex and race differences and reduce the potential influence of age and body mass index (BMI), only healthy Caucasian men who were not overweight and aged between 17 and 25 years are included. The admission requirement is practicing sports for at least 2 h a week. The subjects were asked to perform five times a smooth maximal-depth squat and a right forward lunge step with a predetermined frequency and fluency after a short training. In addition, the volunteers underwent full lower limbs MRI. An ethics committee (Ghent University Hospital, Belgium) approved these investigations (EC2014/0286). The characteristics of our study population are listed in **Table 1**. Because of missing data, there was no complete data acquisition for the squat among the three subjects.

In both examinations, 28 reflective markers are stuck on the skin on palpable anatomical landmarks. The application of skin markers to investigate kinetics is obvious but rather inaccurate. By contrast, using more accurate measurements with implants would raise ethical concerns.

Instrumentation

Our motion capture acquisition strategy was based on a similar study by Deluzio et al. (Deluzio and Astephen, 2007). Spatial marker trajectory data and the corresponding force registrations are imported into the Anybody Modeling System (AMS version 7.1.0, Anybody Technology, Aalborg, Denmark) (Damsgaard et al., 2006) as well as geometric data from a 3-Tesla MAGNETOM Trio-Tim System MRI device (Siemens AG, Erlangen, Germany).

Musculoskeletal Modeling

Motion capture musculoskeletal models were personalized with subject-specific bone geometry that was incorporated in a simulation model from the Twente Lower Extremity Model (TLEM 2.0) dataset (Carbone et al., 2015). An overview of

the musculoskeletal model input is presented in **Figure 1**. In the simulation output, the forces are described in three fixed perpendicular planes.

Data Processing

The output data from musculoskeletal models are numerous, multivariate, and multidimensional (Deluzio and Astephen, 2007; Lai et al., 2009). In contrast to some gait studies that modeled kinematic and kinetic data together, we used only kinetic data (Deluzio and Astephen, 2007; Reid et al., 2010; Galloway et al., 2012). We think that integrating linear quantities (forces) and rotation quantities (angles) is like comparing apples and oranges. On top of that, the kinetic data in Anybody is generated by an inverse dynamics approach starting from the kinematic data.

The beginning and end frames of all motion lab recordings are not useful due to irrelevant transients. Analogously, the peak evolution will vary from the center of the recorded data. Hence, data alignment and trimming are essential prior to incorporating the subjects' motion recordings into a statistical model. These operations are executed using standard implementations in MATLAB (MathWorks, Natick, MA, United States).

The frame recorded with the peak knee flexion angle is defined as 50% progress of the motion. Trimming is based on knee flexion. For the lunge, the best is to consider only the closed chain part. As such, recordings where the right foot is not on the right force plate are left aside. Several arbitrary ways to execute an open-chain motion could be an important source of noise η. Noise is defined as artifacts when processing the input data to the output data (Lai et al., 2009). On top of data, we used only information from the leg that was the most loaded. So, in contrast to the squat data, a lot of waveform data are not used for the lunge.

Interpolation is performed to ensure that the measurements are running synchronized in real time. All trimmed measurements are subdivided into 0–50–100 proceedings, corresponding to the onset, the middle, and the finish of motion, respectively. Each set of kinetic data is arranged in a feature vector and concatenated into a training matrix. The training data matrix X contains observations in the rows and subjects in the columns as described in Eqs. [1] to [3].

$$X = \left[x_1, x_2, x_3, \ \ldots, x_i, \ldots, x_{p-1}, x_p\right] \tag{1}$$

TABLE 1 | Demographic and anthropometric characteristics of the study population.

Demographic descriptor	Mean (95% CI*)	Normal values
Height (cm)	181.79 (180.08–183.51)	Not applicable
Weight (kg)	71.75 (69.63–73.88)	Not applicable
Body mass index (kg/m^2)	21.70 (21.16–22.23)	18.5–25 (Waxman, 2004)
Sport activity (hours a week)	3.40 (2.76–4.03)	Not applicable
Center-edge angle (°)	28.41 (27.19–29.63)	25–39° (Audenaert et al., 2012; Ghaffari et al., 2018)
Alpha angle (°)	64.61 (62.38–66.84)	<55° (Audenaert et al., 2012; Ghaffari et al., 2018)
Centrum-collum-diaphyseal angle or neck-shaft angle (°)	129.24 (127.99–130.49)	125–135° (Audenaert et al., 2012)
Femoral anteversion angle (°)	9.40 (7.30–11.49)	<15° (Audenaert et al., 2012)

*Confidence interval of the mean.
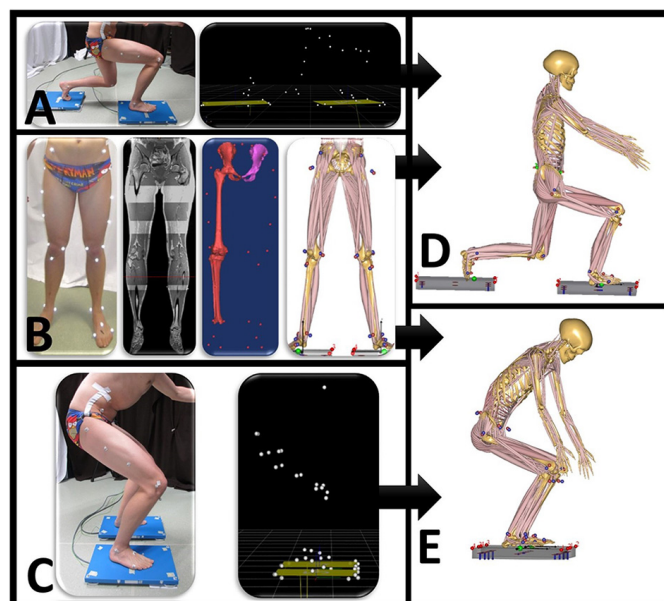
FIGURE 1 | Overview of data input for the motion capture musculoskeletal simulation model. **(A)** Motion is performed when standing on two force plates. Motion capture data synchronized with ground reaction forces are exported as .c3d file. **(B)** Twenty-eight reflective markers are placed on anatomical bony landmarks. A MRI scan of the full lower limb is performed. Segmentation of pelvis, thigh, and shank with corresponding positions of marker landscape. **(C)** Motion capture squat model. Anybody squat **(D)** and lunge **(E)** model.
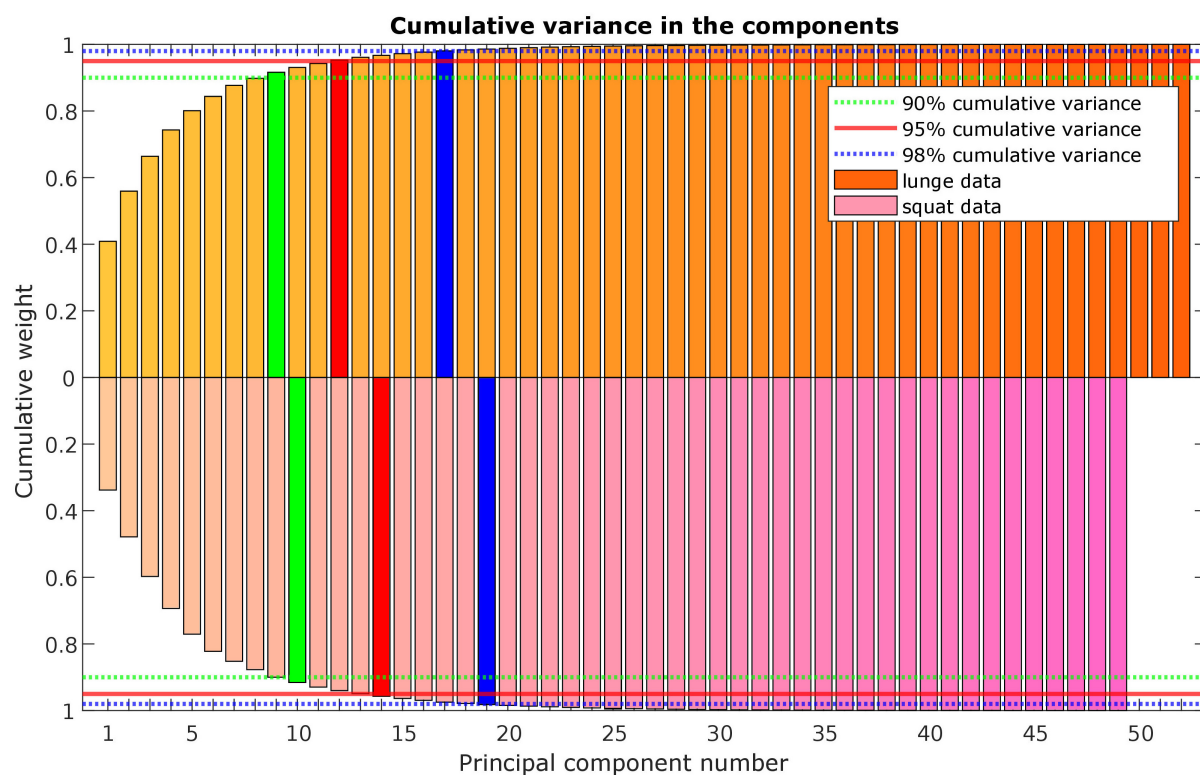
FIGURE 2 | Scree plot with the cumulative variance of the modes (or principal components) in the lunge (orange) and squat (purple) kinetic model.

FIGURE 3 | Relation between the kinetic waveform simulation output and the squat progress for each individual sample in gray. Mean values of the measurements in green ±2 standard deviations of the first mode in red and blue. The first mode accounts for 33.80% of the inter-subject population variance. Note the different y axis calibrations.
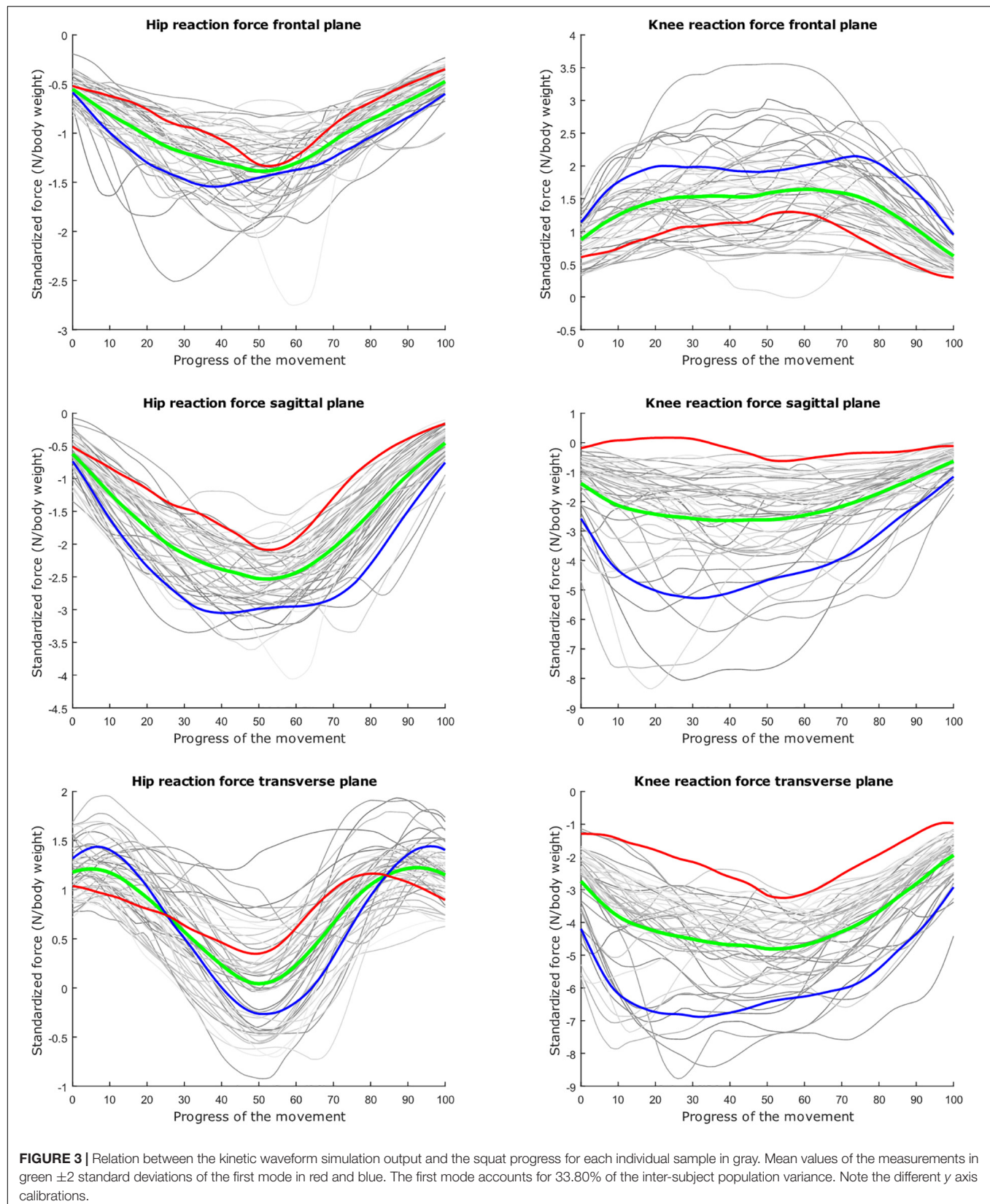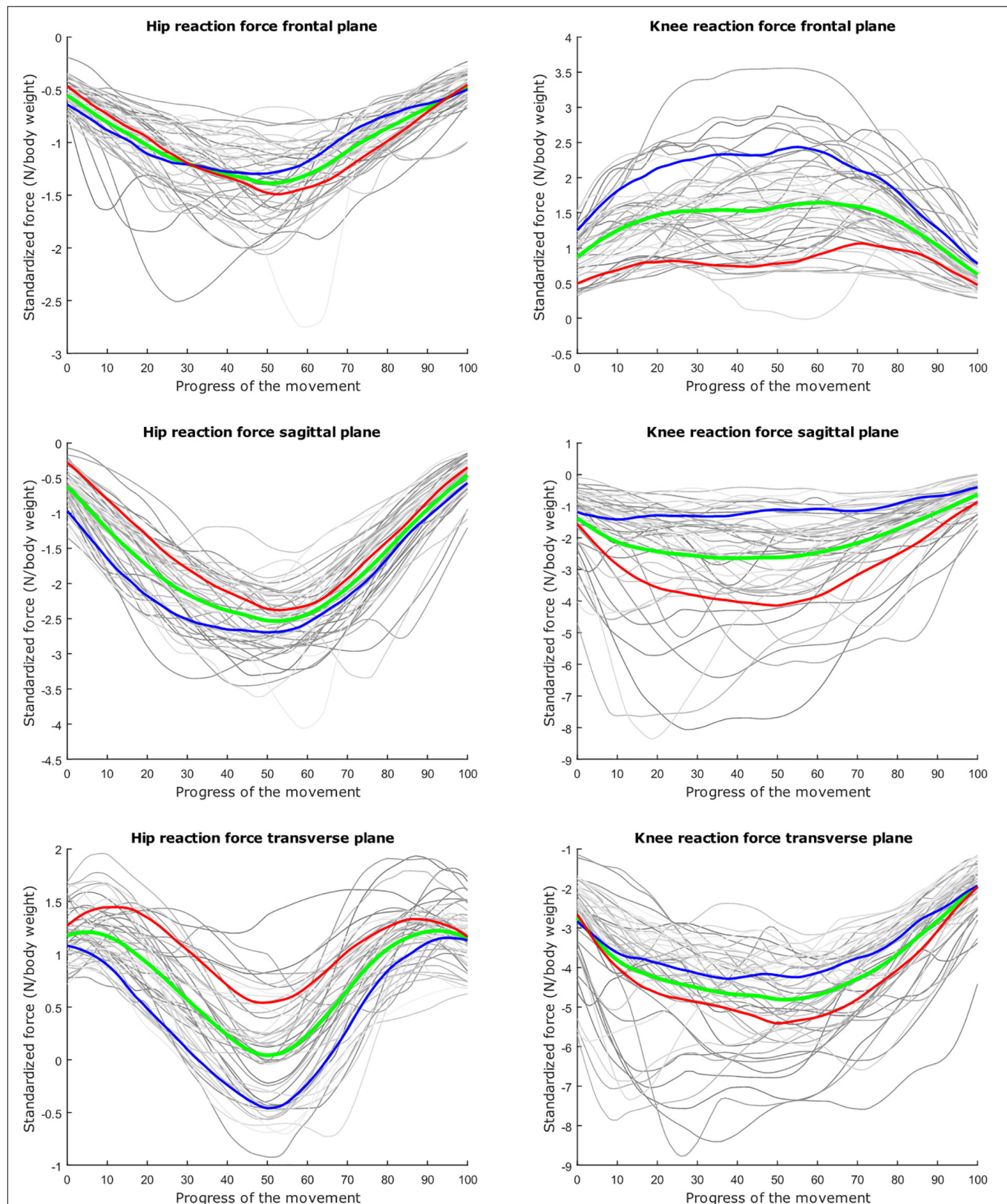
FIGURE 4 | Mean values of joint reaction forces during deep squatting in green ±2 standard deviations of the second mode in red and blue. The second mode accounts for 14.05% of the inter-subject population variance.
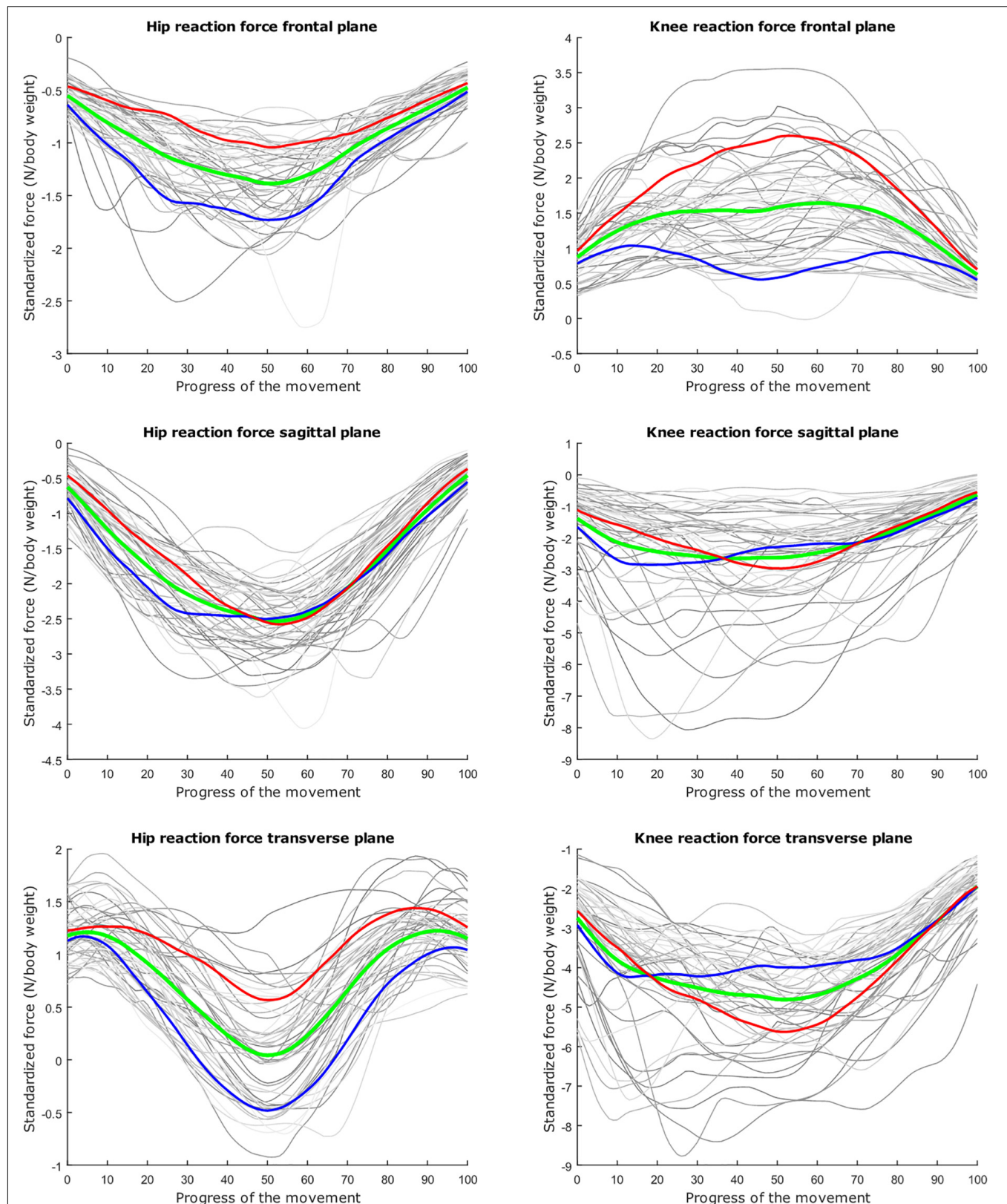
FIGURE 5 | Mean values of joint reaction forces during deep squatting in green ±2 standard deviations of the third mode in red and blue. The third mode accounts for 11.88% of the inter-subject population variance.
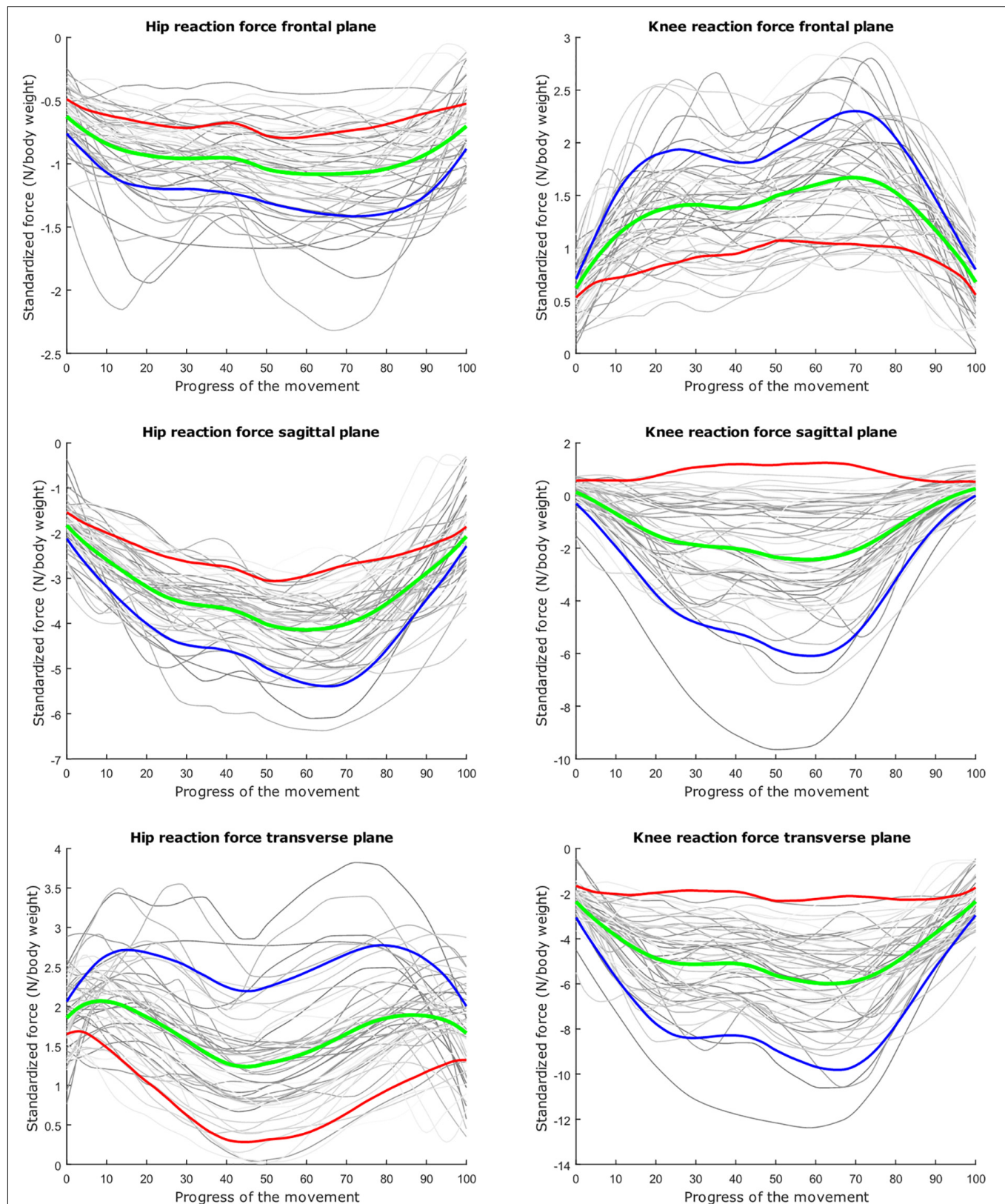
FIGURE 6 | Mean values of joint reaction forces during lunging in green ±2 standard deviations of the first mode in red and blue. The first mode accounts for 40.87% of the inter-subject population variance.

FIGURE 7 | Mean values of joint reaction forces during lunging in green ±2 standard deviations of the second mode in red and blue. The second mode accounts for 15.07% of the inter-subject population variance.
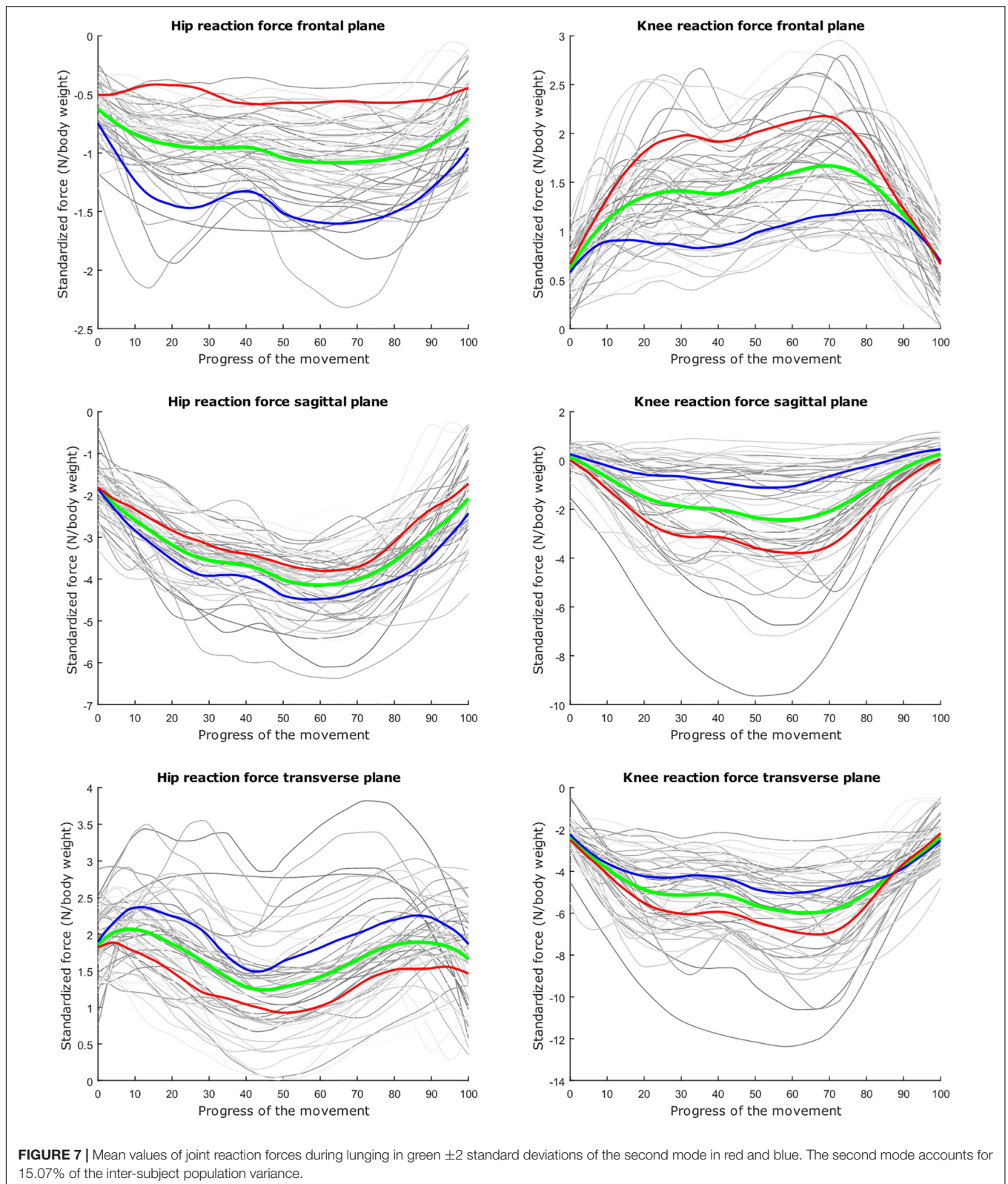
An observation expresses several dynamic parameters on a certain progress of the aligned lunge or squat motion from 0 to 100%. For each participant i, the kinetic model input data are taken from the musculoskeletal model output. The kinetic variables are implemented into a subject vector x_i for the ith subject (out of p). p represents the number of training samples,
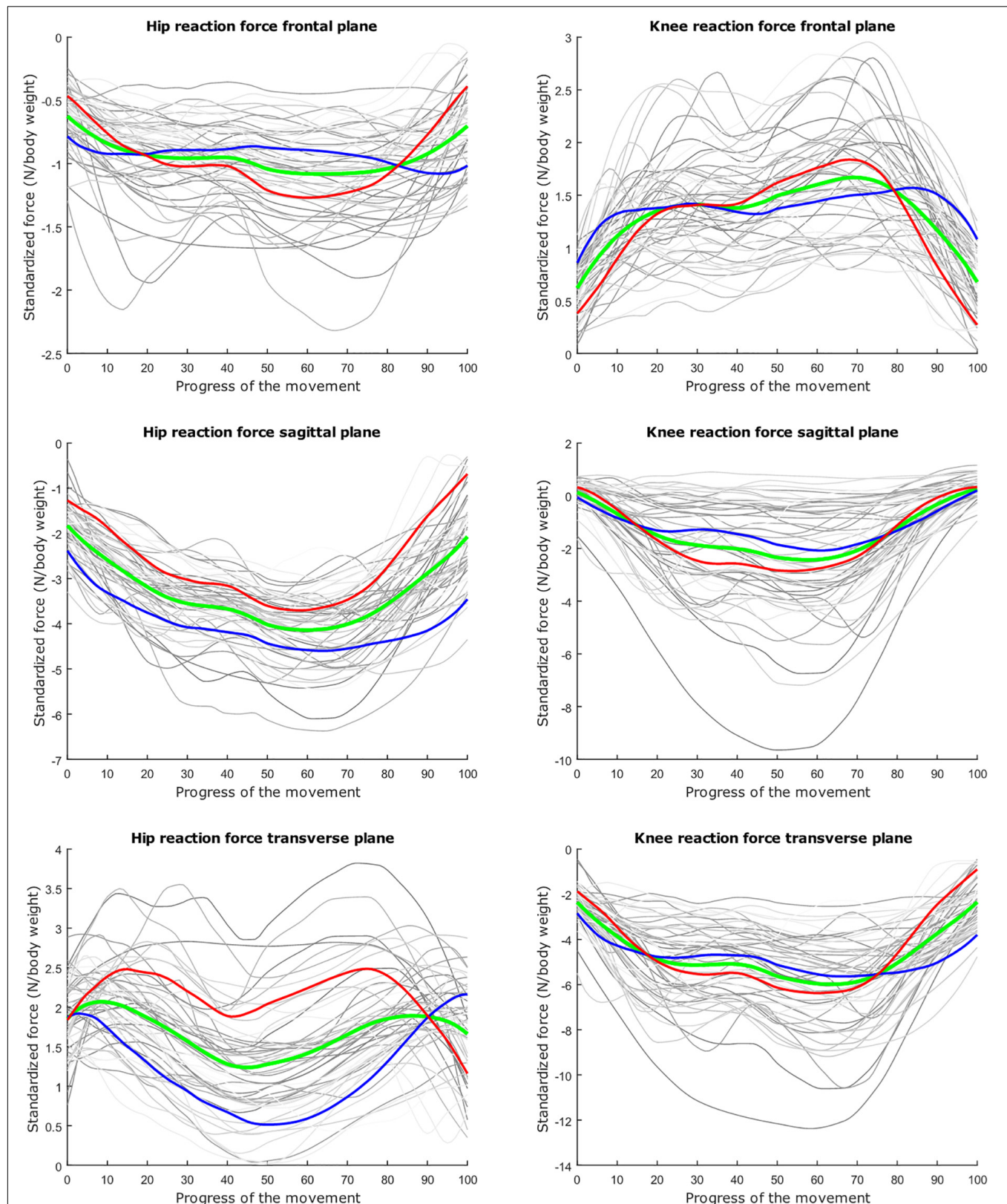
FIGURE 8 | Mean values of joint reaction forces during lunging in green ±2 standard deviations of the third mode in red and blue. The third mode accounts for 10.46% of the inter-subject population variance.

TABLE 2 | Validation analyses of the squat and lunge statistical kinetic model. We consider squat and lunge models that capture 80, 90, 95, and 98% of inter-individual population variance.

Validation summary	Squat model				Lunge model			
% of inter-variability in the population	80%	90%	95%	98%	80%	90%	95%	98%
Model accuracy RMSE (median ±IQR**) (BW)	0.0149 ±0.0122	0.0107 ±0.0087	0.0075 ±0.0064	0.0054 ±0.0048	0.0248 ±0.0210	0.0162 ±0.0167	0.0132 ±0.0126	0.0082 ±0.0083
Dimensionality*	6	10	14	19	5	9	13	17
Model specificity RMSE (median ±IQR**) (BW)	0.1582±0.0943	0.1581±0.0948	0.1583±0.0946	0.1584±0.0943	0.1291±0.0831	0.1310±0.0815	0.1314±0.0809	0.1320±0.0803

*i.e., the number of modes. **IQR, interquartile range.

being 53 for the lunge and 50 for the squat.

$$x_i = \begin{bmatrix} HJRF_x, & HJRF_y, & HJRF_z, KJRF_x, & KJRF_y, & KJRF_z, \end{bmatrix}^T \quad (2)$$

The input matrices JRF_{axis} consist all of 101 observations o.

$$JRF_{\text{axis}} = \begin{bmatrix} \text{measurement 1} \\ \text{measurement 2} \\ \ldots \\ \ldots \\ \text{measurement 51 (at maximal right knee flexion)} \\ \ldots \\ \text{measurement 101} \end{bmatrix}$$

$$(3)$$

D serves as a diagonal matrix with row-wise standard deviations d_o for each observation o. The total number of observations is the multiplication of the number of dynamic variables and aligned time instances.

$$D = \begin{bmatrix} d_1 & 0 & 0 \ldots \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ \ldots & 0 & 0 & d_{606} \end{bmatrix} \quad (4)$$

After normalization by row-wise standard deviation in [4] and [5] as well as mean centering in [6], a residual matrix R is created. R comprises the entry data for the model M as a measure of dispersion.

$$\tilde{X} = \begin{bmatrix} & \ldots & \\ \ldots & \tilde{x}_i & \ldots \\ & \ldots & \end{bmatrix} = D^{-1}X \quad (5)$$

$$R = \begin{bmatrix} & \ldots & \\ \ldots & \hat{x}_i & \ldots \\ & \ldots & \end{bmatrix} \text{ with} \hat{x}_i = \tilde{x}_i - \bar{x} \quad (6)$$

PCA is a powerful dimensionality reduction technique developed by Karl Pearson. It is not a method to investigate the center size of the data but the common variability. PCA is mathematically defined as an orthogonal linear transformation. PCA transforms the data; as such, most of the variance of the data will come to lie in the first components. This allows us to create statistical models. Altogether PCA is a reliable tool in capturing the salient features of waveform data (Robbins et al., 2013; Jolliffe and Cadima, 2016).

Using this for a statistical model, it enables to generate population data from a small set of clinical data. The kinetic model should represent waveform data as a linear combination of vectors, representing the primary modes of variation in experimental data (Jolliffe et al., 2002; Saliba et al., 2018). Eigenvalues and eigenvectors have been created by singular value decomposition.

$$R = U \, x \, L \, x \, A^T \quad (7)$$

In Eq. [7], U and A are the left and right singular vectors, so $U^T.U = I$ and $A^T.A = I$ because of orthogonality. I refers to the unity matrix. L is a diagonal matrix that contains the square roots

of the eigenvalues $\sqrt{\lambda_k}$ belonging to $R^T x R$, as $k\{1, 2, 3, \ldots, p\}$. A^T contains the eigenvectors of $R^T x R$, whereas U has the scaled versions of the principal component (PC) scores u_{ok}. Here, o corresponds to the observation and k to the PC number. The PC scores are mentioned in Eq. [8].

$$z_{ok} = u_{ok}\sqrt{\lambda_k} \qquad (8)$$

The PC scores from a single waveform quantify the contribution of each feature. The variance of the scores for the kth eigenvalue of $R^T x R$ amounts to $\frac{\lambda_k}{n-1}$, as λ_k represents the variance of the kth PC, whereas n is the number of observations.

The cumulative variance of each mode k is defined as

$$\text{Compactness}\ (M) = \frac{1}{p-1}\sum_{m=1}^{k}\lambda_m \qquad (9)$$

The PC weight matrix W in Eq. [10] involves the correlation coefficients between components and test subject data.

$$W = L\, x\, A^T \qquad (10)$$

A set of patient data can be approximately reconstructed by using t selected PCs in Eq. [11].

$$\hat{x}_{ij} = D\sqrt{p-1}\sum_{i=1}^{t} u_{oi}\sqrt{l_i}a_{ki} + \bar{x} \qquad (11)$$

As mentioned before, D represents the diagonal matrix of row-wise standard deviations and p stands for the subject count. $\sqrt{l_i}$ is the ith diagonal element of L, also from the singular value decomposition (Jolliffe et al., 2002; Galloway et al., 2012; Jolliffe and Cadima, 2016).

Model Validation

Validation is defined as the process of ensuring that the dimensionality-reduced PCA model accurately represents real-world kinetics. Probably the most important problem arising with this process is the choice of the optimal number of the principal components to be retained. PCA projects the input data from a high dimensional space into a subspace of lower dimension, which can then further be divided into two separate subspaces: the *kinetic data subspace*, preserving the essence of the original kinetic data as lossless as possible, and the *noise subspace*, corresponding to the remaining tail of principal components associated with the smallest eigenvalues. Given the complexity of the problem of optimally defining the threshold between signal and noise principal components, the literature on the topic is overwhelming and beyond the scope of this work. Reliable results in distinguishing components that express meaningful correlations among variables as opposed to trivial components, explaining noise, have been provided using the Monte Carlo permutation test (Peres-Neto et al., 2005). The principal components were tested for representing valid correlations as opposed to residual error using the following two criteria: rank of roots and equality of roots (Jackson, 2005; Vasco, 2012).

Further, four quantitative model parameters are investigated. "Goodness" measures are chosen according to the statistical shape modeling study of Styner et al. (2003) in which there is also a PCA dimensionality reduction algorithm. This study is, to the authors' best knowledge, the first to provide such an approach, implemented for a kinetic model.

Model Accuracy

$$\text{Accuracy}\ (M) = \frac{1}{p}\sum_{i=1}^{p}\|\hat{x}_i(M) - x_i\|^2 \qquad (12)$$

The first validation test that analyzes relevant information is retained by the model or otherwise states how well the original data can be reconstructed from the model given the number of principal components retained. Here, the root-mean-square error (RMSE) is computed in Eq. [12] as the average absolute difference between the original training data and the reconstructed data for models with 80, 90, 95, and 98% variance of the original data.

Model Compactness

The model will be compact enough if it can describe the variance in kinetic measurements with a minimal number of modes. Eq. [9] is used to describe the compactness with the cumulative variance for a certain number of modes.

Model Generalization

$$\text{Generalization}\ (M) = \frac{1}{T_g}\sum_{i=1}^{T_g}\|\hat{x}_i(M_i') - x_i\|^2 \qquad (13)$$

The model generalization quantifies the ability of models to represent new instances. The generalization ability is evaluated by performing a series of leave-one-out tests on the training data. The question here is: how many training samples are necessary to approach the population precisely? The generalization ability is therefore a means for *post hoc* sample size evaluation. If having enough training samples, we expect the model to be able to describe unseen data quite accurately (Wang and Shi, 2017). The generalization value can be interpreted as the median out-of-sample accuracy value.

The generalization evolution gives the RMSE between the excluded subject data and the best-matched 95% variance model M' values of randomly selected training data by ascending number of training samples in the model M'. The higher the T_g test value in Eq. [13], the higher the precision of the median generalization value. Here the number of models created for each number of training samples amounts to $T_g = 10{,}000$.

Model Specificity

$$\text{Specificity}\ (M) = \frac{1}{T_s}\sum_{i=1}^{T_s}\|\hat{x}_i'(M) - x_i\|^2 \qquad (14)$$

A population model is able to generate new data. The model specificity measures the soundness of new instances randomly generated by the developed model M. Models with 80, 90, 95, and 98% of variance are tested. $\hat{x}_i'(M)$ refers to a randomly generated subject.
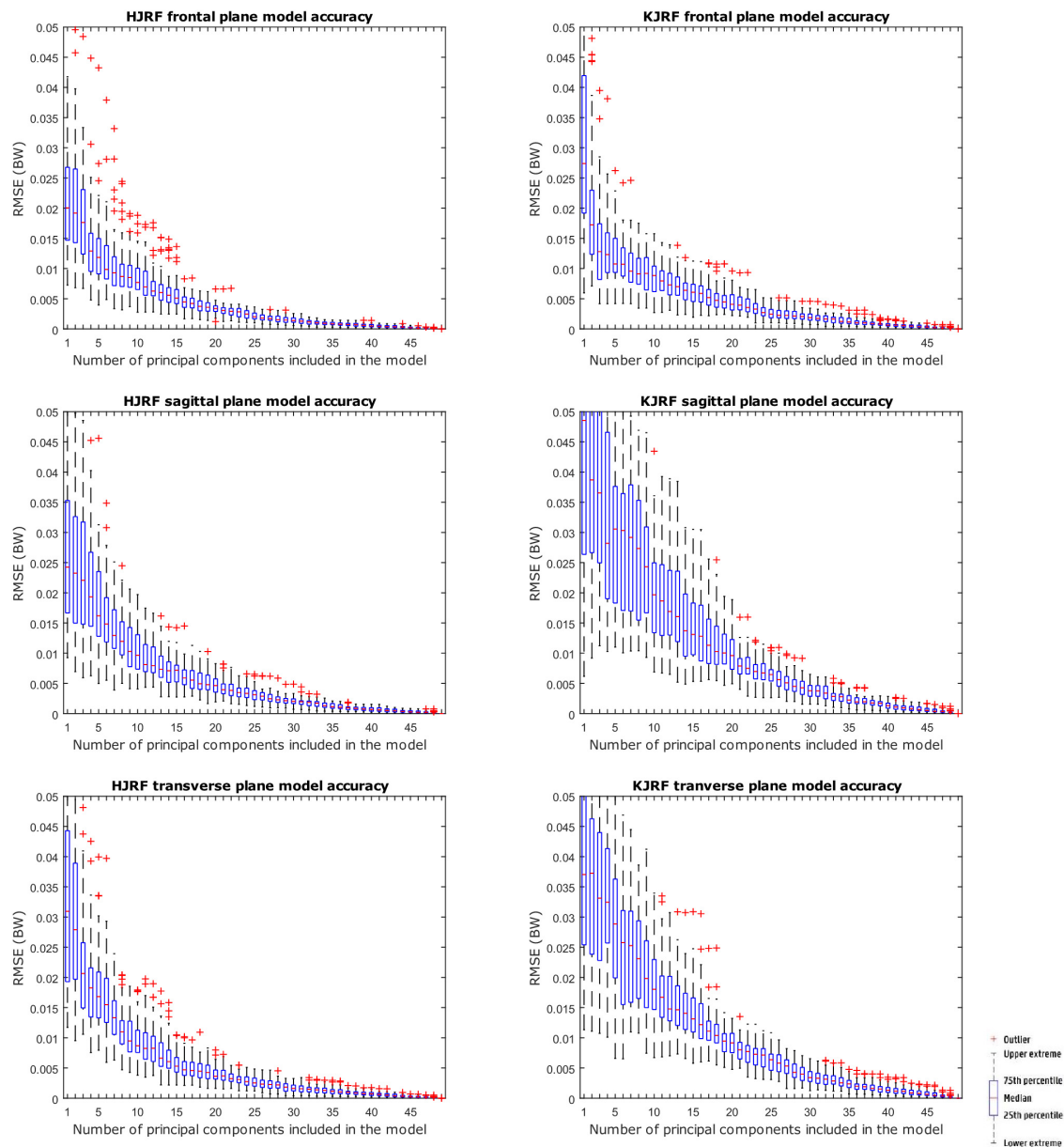
FIGURE 9 | RMSE for the original squat training data versus reconstructed squat data with an increasing number of principal components on the *x* axis.

We assume that the PCs of the model are normally distributed (Jolliffe et al., 2002; Jackson, 2005; Galloway et al., 2012). The specificity estimator is defined in Eq. [14]. For each observation o, an imaginary subject i is defined by choosing random normal distributed values $n \in N(0, 1)$ for each mode m in the model M as in Eq. [15].

$$\hat{x}'_{i,o}(M) = \bar{x}_o + d_o \sum_{m=1}^{p} n.\sqrt{\lambda_m}.z_m = \bar{x} + d_o \sum_{m=1}^{p} n.u_m \quad (15)$$

The RMSE is defined as the error between the virtually subject data and the most similar sample in the training dataset. The specificity value can be interpreted as the median approximation error of T_s generated subjects. The higher the T_s test value, the higher the precision of the specificity. Here the number of models created is set to $T_s = 1,000,000$.

RESULTS

The average hip and knee peak flexion angles are, respectively 95° and 104° for the lunge and 107° and 112° for the squat motion, respectively. The average peak hip joint reaction force (HJRF) amounts to 3.08 times body weight (BW) for the maximum-depth squat and 4.76 BW for lunging. The means

TABLE 3 | Choosing the optimal amount of principal components for the squat kinetic datasets.

PC	Eigenvalue	Percentage of variance	Cumulative variance	Rank of roots	Equality of roots
1	204.85	33.80	33.80	0.001	0.001
2	85.11	14.04	47.85	0.001	0.001
3	71.98	11.88	59.73	0.001	0.001
4	58.21	9.61	69.33	0.001	0.001
5	46.81	7.73	77.06	0.001	0.001
6	31.26	5.16	82.22	0.001	0.001
7	18.13	2.99	85.21	0.001*	0.001
8	15.18	2.50	87.71	1	0.001
9	13.82	2.28	89.99	1	0.001
10	9.55	1.58	91.57	1	0.001
11	8.38	1.38	92.95	1	0.001
12	6.37	1.05	94.00	1	0.001
13	5.57	0.92	94.92	1	0.001
14	4.65	0.77	95.69	1	0.005*
15	3.97	0.66	96.35	1	0.078

*Type I error probability is set to 0.05. Rank of roots measure suggests that seven principal components (PCs) are statistically significant in meaningfully describing the dataset, corresponding to 85% of data variance, whereas the equality of roots suggests that 14 PCs are to be included (representing 95.7% of data variance). *p < 0.05.*

for peak knee joint reaction force (KJRF) are still higher: 4.52 BW for squat motion and 7.16 BW for the lunge. The trimmed original waveform data from HJRF and KJRF of our musculoskeletal model are represented by gray curves in **Figures 3–8**.

A statistical model of kinetic output data from the AMS was made for deep squatting and another one for lunging. **Figure 2** displays the cumulative variance of modes in the statistical model. The variances of the first three modes in the squat model are illustrated in **Figures 3–5**. Together they represent 59.73% of the population variance. For the lunge, the first three modes accounts for 66.40% of population variance. More about these modes are detailed in **Figures 6–8**. In **Table 2**, the in-sample model accuracy and the specificity median RMSE are described for each model. The boxplots in **Figure 9** illustrate the in-sample squat model accuracy for the ascending number of the components included. There is a boxplot for every variable in the model. The median and the interquartile range of the RMSE, when compared to the initial data, decrease as more PCs are included in the reconstructed data, as expected. The model accuracies from the lunge model are quite similar but around one IQR RMSE higher. In contrast, the lunge model is a bit more compact. For calculating the out-of-sample accuracy based on leave-one-out tests, we based on the lunge data because it has the most test data. The results of the model compactness and the statistical findings of the permutation testing related to the number of the principal components used are demonstrated in **Table 3**.

Regarding **Figure 10**, for each training data input amount going from 4 to 52, 10,000 models were created, including 95% population variance, to reconstruct an excluded subject. Out-of-sample accuracy RMSE from the reconstructed data versus the original excluded data are given on the *y* axis in box-and-whisker diagrams. The boxplots are log–log-scaled in order to visualize the downward trend of the out-of-sample accuracy. Also plotted is a horizontal line of the in-sample model accuracy of our 95% model. The out-of-sample accuracies are less than 0.1 BW, except for the KJRF in the transverse plane. From 50 test subjects up, the out-of-sample accuracies are clearly stagnating for the HJRF in the frontal and the transverse plane as well as for the KJRF in the sagittal plane.

DISCUSSION

The validation analysis confirmed that our models have a high degree of compactness and accuracy. Many types of noise are in the higher components. The PCA technique has adequately allowed rejection of the error variance from the model. The meaningful variance is obviously divided over the first 12 or 14 components. This multidimensionality describes the silent features in the data and, eventually, they could be linked to the varying characteristics of the study population. A common source of meaningless variance originates from data alignment. It is impossible to avoid this because we do not want to introduce supplementary noise in the data by aligning them more stringently. Since all subjects have a BMI lower than 25, skin shift errors during movements are limited (Cappozzo et al., 1996).

According to the lunge, the model only describes the closed-chain part of motion for two reasons. First, femoroacetabular impingement and joint reaction forces are more pronounced at higher flexion (Audenaert et al., 2012). Secondly, while creating a model from the onset of the lunge back to the original position, the model would be no longer compact enough because there is too much degree of freedom when moving a leg in the air.

The dominant mode is supposed to describe the overall variance (Jolliffe et al., 2002), as is clearly apparent in the lunge model. In the squat model, the overall variance is limited for the HJRF in the frontal and the sagittal plane. This is due to low hip flexion and rotation moments because the center of gravity will lie almost perfectly between the hip joints (and not the knee joints during squat), in contrast to the lunge case. For this statement, we based on Schwab et al. (2006). They found that, for young adults, the femoral head position appears to be a reliable indicator for the gravity line in the sagittal plane during stance (Schwab et al., 2006).

The second mode of the squat model indicates that a high HJRF component in the transverse plane results in high KJRF components in the frontal and sagittal plane in order to counterbalance the downward force at the hip. The third mode correlates the depth of squatting with the joint reaction force components in the frontal plane. For the lunge, the association of the frontal joint reaction force components is mainly summarized in the second mode. Finally, according to our interpretation, the third mode of the lunge model may take alignment errors into account.

The RMSE for model accuracy are far below 0.05 BW, as opposed to similar studies. The specificity was almost equal for models with 80, 90, 95, and 98% of variance. It questions
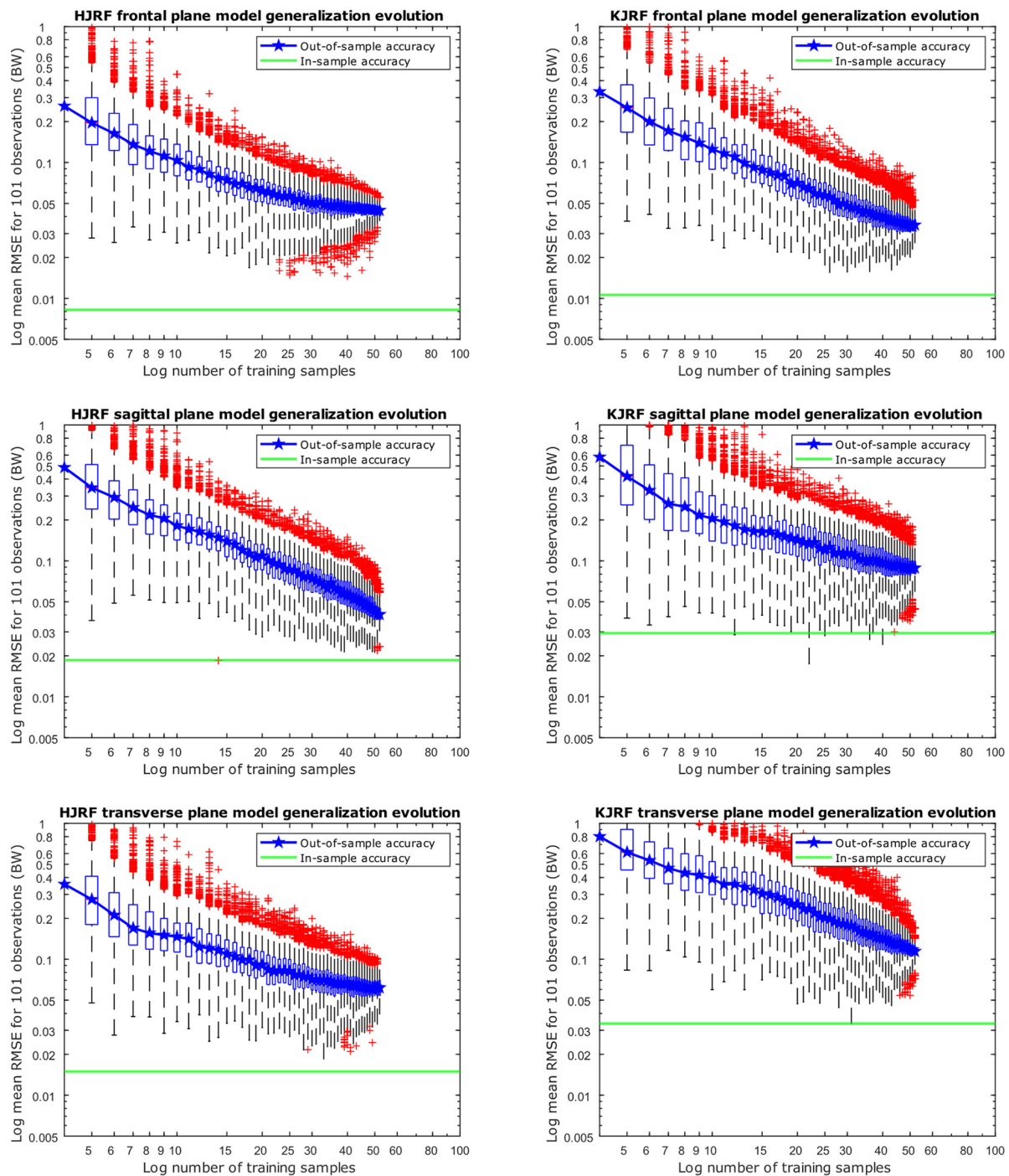
FIGURE 10 | Accuracy evolution of kinetic lunge data with log–log scaling (boxplot with root-mean-square error of the reconstructed data with 95% variance versus the original training data) for different levels of prior knowledge expressed as amounts of training data in a kinetic model. The green horizontal line indicates the in-sample target accuracy.

the relevance of taking the model specificity into account in this setting. According to the generalization evolution, we could conclude that, minimally, 50 samples are enough to provide reliable models at 0.1 BW precision for both squat and lunge motion. Nevertheless, we recommend exceeding this threshold

number because the in-sample accuracy is still lower, especially for the squat. Note that gender, age group, BMI group, and race differences are not included here. Therefore, it is very likely that, in more heterogeneous populations like the elderly, 50 samples will be too low to ensure reliable models.

Unfortunately, electromyography data are not collected during this study. This could give information about muscle activation and muscle strength. Motor unit action potentials could be registered non-invasively by using surface electromyography. It has been stated several times that the muscle activation patterns depend on several aspects like training level and osteoarthritis (Benedetti et al., 2003; Knoop et al., 2012; Mei et al., 2017). The integration of electromyography and kinetic data could help to declare aberrant kinetic patterns.

By applying correlation matrix PCA to obtain uncorrelated maximum-variance linear combinations and given that there is only kinetic data with limited scaling differences, some more PCs are required to account for the same amount of covariance compared to classical covariance matrix PCA (Jolliffe et al., 2002; Jolliffe and Cadima, 2016). This makes the selection of PCs in the kinetic data subspace even more crucial to ensure model validation properties like accuracy, compactness, generalization, and specificity, which is the major drawback of PCA (Jolliffe et al., 2002; Peres-Neto et al., 2005; Vasco, 2012). To handle this, there are numerous methods described in the literature, but there is no consensus yet. We objectified our selection strategy based on eigenvalues by considering the validation measures for different cutoffs. On top of that, for the model generalization and specificity abilities, we assume multivariate normal distribution which is seldom true (Vasco, 2012).

The most important limitation of the present work, however, relates to the population under investigation, namely, young male, Belgian adolescents and the unknown extent of which findings can be extrapolated to other populations. Nevertheless, in general terms, we expect our results to be representative by extension for a Western European population.

CONCLUSION

We created two models that describe kinetics from both hip and knee joint, contrary to the limited number of studies available with PCA analyses of waveform data considering the knee only (Deluzio and Astephen, 2007; Reid et al., 2010; Galloway et al., 2012). Since all muscles from the knee, except from the M. popliteus (Paulsen and Waschke, 2011), are biarticular and the body should be seen as a whole, a model with the HJRF as well as KJRF is preferable. We proved that such a

model for 95% of population variance was compact and very accurate (<0.015 BW). To describe the population at <0.1 BW precision, our small sample size was still sufficient. Using t-tests to investigate differences in PC scores in comparing studies will enable the creation of personalized hip and tibial implants with specific weight-bearing properties, resulting in prolonged longevity. In our opinion, this is a very important feature since total knee arthroplasty and total hip arthroplasty are increasingly utilized to treat more physically active patients.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Commission for Medical Ethics, UZ Gent. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Almeida, D. F., Ruben, R. B., Folgado, J., Fernandes, P. R., Audenaert, E., Verhegghe, B., et al. (2016). Fully automatic segmentation of femurs with medullary canal definition in high and in low resolution CT scans. *Med. Eng. Phys.* 38, 1474–1480. doi: 10.1016/j.medengphy.2016.09.019

Audenaert, E. A., Pattyn, C., Steenackers, G., De Roeck, J., Vandermeulen, D., and Claes, P. (2019). Statistical shape modelling of skeletal anatomy for sex discrimination: their training size, sexual dimorphism and asymmetry. *Front. Bioeng. Biotechnol.* 7:302. doi: 10.3389/fbioe.2019.00302

Audenaert, E. A., Peeters, I., Vigneron, L., Baelde, N., and Pattyn, C. (2012). Hip morphological characteristics and range of internal rotation in femoroacetabular impingement. *Am. J. Sports Med.* 40, 1329–1336. doi: 10.1177/0363546512441328

Bagwell, J. J., Snibbe, J., Gerhardt, M., and Powers, C. M. (2016). Hip kinematics and kinetics in persons with and without cam femoroacetabular impingement during a deep squat task. *Clin. Biomech.* 31, 87–92. doi: 10.1016/j.clinbiomech.2015.09.016

Benedetti, M. G., Catani, F., Bilotta, T. W., Marcacci, M., Mariani, E., and Giannini, S. (2003). Muscle activation pattern and gait biomechanics after total knee replacement. *Clin. Biomech.* 18, 871–876. doi: 10.1016/s0268-0033(03)00146-3

Bischoff, J. E., Dai, Y., Goodlett, C., Davis, B., and Bandi, M. (2014). Incorporating population-level variability in orthopedic biomechanical analysis: a review. *J. Biomech. Eng.* 136:021004. doi: 10.1115/1.4026258

Cappozzo, A., Catani, F., Leardini, A., Benedetti, M. G., and Croce, U. D. (1996). Position and orientation in space of bones during movement: experimental artefacts. *Clin. Biomech.* 11, 90–100. doi: 10.1016/0268-0033(95)00046-1

Carbone, V., Fluit, R., Pellikaan, P., van der Krogt, M. M., Janssen, D., Damsgaard, M., et al. (2015). TLEM 2.0 – A comprehensive musculoskeletal geometry

dataset for subject-specific modeling of lower extremity. *J. Biomech.* 48, 734–741. doi: 10.1016/j.jbiomech.2014.12.034

Carbone, V., van der Krogt, M. M., Koopman, H. F., and Verdonschot, N. (2012). Sensitivity of subject-specific models to errors in musculo-skeletal geometry. *J. Biomech.* 45, 2476–2480. doi: 10.1016/j.jbiomech.2012.06.026

Daffertshofer, A., Lamoth, C. J., Meijer, O. G., and Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clin. Biomech.* 19, 415–428. doi: 10.1016/j.clinbiomech.2004.01.005

Damsgaard, M., Rasmussen, J., Christensen, S. T., Surma, E., and De Zee, M. (2006). Analysis of musculoskeletal systems in the anybody modeling system. *Simul. Model. Pract. Theory* 14, 1100–1111. doi: 10.1016/j.simpat.2006.09.001

de Ruiter, L., Janssen, D., Briscoe, A., and Verdonschot, N. (2017). The mechanical response of a polyetheretherketone femoral knee implant under a deep squatting loading condition. *Proc. Instit. Mech. Eng. H J. Eng. Med.* 231, 1204–1212. doi: 10.1177/0954411917738805

Deluzio, K. J., and Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. *Gait Posture* 25, 86–93. doi: 10.1016/j.gaitpost.2006.01.007

Galloway, F., Worsley, P., Stokes, M., Nair, P., and Taylor, M. (2012). Development of a statistical model of knee kinetics for applications in pre-clinical testing. *J. Biomech.* 45, 191–195. doi: 10.1016/j.jbiomech.2011.09.009

Ghaffari, A., Davis, I., Storey, T., and Moser, M. (2018). Current concepts of femoroacetabular impingement. *Radiol. Clin. North Am.* 56, 965–982.

Hartmann, H., Wirth, K., and Klusemann, M. (2013). Analysis of the load on the knee joint and vertebral column with changes in squatting depth and weight load. *Sports Med.* 43, 993–1008. doi: 10.1007/s40279-013-0073-6

Henak, C. R., Anderson, A. E., and Weiss, J. A. (2013). Subject-specific analysis of joint contact mechanics: application to the study of osteoarthritis and surgical planning. *J. Biomech. Eng.* 135:021003. doi: 10.1115/1.4023386

Honari, G., and Taylor, J. S. (2013). Commentary on Crawford, et al., The role of patch testing in the evaluation of orthopedic implant-related adverse effects: current evidence does not support broad use. *Dermatitis* 24, 108–111. doi: 10.1097/der.0b013e3182979722

Jackson, J. E. (2005). *A user's Guide to Principal Components.* Hoboken, NJ: John Wiley & Sons.

Jolliffe, I. T., Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., et al. (2002). *Principal Component Analysis*, 2nd Edn. Cham: Springer.

Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374:20150202. doi: 10.1098/rsta.2015.0202

Kang, L., Galvin, A. L., Brown, T. D., Jin, Z., and Fisher, J. (2008). Quantification of the effect of cross-shear on the wear of conventional and highly cross-linked UHMWPE. *J. Biomech.* 41, 340–346. doi: 10.1016/j.jbiomech.2007.09.005

Knoop, J., Dekker, J., Klein, J. P., van der Leeden, M., van der Esch, M., Reiding, D., et al. (2012). Biomechanical factors and physical examination findings in osteoarthritis of the knee: associations with tissue abnormalities assessed by conventional radiography and high-resolution 3.0 Tesla magnetic resonance imaging. *Arthritis Res. Ther.* 14:R212. doi: 10.1186/ar4050

Kutzner, I., Heinlein, B., Graichen, F., Bender, A., Rohlmann, A., Halder, A., et al. (2010). Loading of the knee joint during activities of daily living measured in vivo in five subjects. *J. Biomech.* 43, 2164–2173. doi: 10.1016/j.jbiomech.2010.03.046

Lai, D. T., Begg, R. K., and Palaniswami, M. (2009). Computational intelligence in gait research: a perspective on current applications and future challenges. *IEEE Trans. Inf. Technol. Biomed.* 13, 687–702. doi: 10.1109/TITB.2009.2022913

Mei, Q., Gu, Y., Fu, F., and Fernandez, J. (2017). A biomechanical investigation of right-forward lunging step among badminton players. *J. Sports Sci.* 35, 457–462. doi: 10.1080/02640414.2016.1172723

O'Brien, S. T., Luo, Y., and Brandt, J.-M. (2015). In-vitro and in-silico investigations on the influence of contact pressure on cross-linked polyethylene wear in total knee replacements. *Wear* 332, 687–693. doi: 10.1016/j.wear.2015.02.048

Paulsen, F., and Waschke, J. (2011). *Sobotta Atlas of Human Anatomy*, 15th Edn. Amsterdam: Elsevier.

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* 49, 974–997. doi: 10.1016/j.csda.2004.06.015

Pietsch, M., Djahani, O., Hochegger, M., Plattner, F., and Hofmann, S. (2013). Patient-specific total knee arthroplasty: the importance of planning by the surgeon. *Knee Surg. Sports Traumatol. Arthrosc.* 21, 2220–2226. doi: 10.1007/s00167-013-2624-7

Reid, S. M., Graham, R. B., and Costigan, P. A. (2010). Differentiation of young and older adult stair climbing gait using principal component analysis. *Gait Posture* 31, 197–203. doi: 10.1016/j.gaitpost.2009.10.005

Robbins, S. M., Astephen Wilson, J. L., Rutherford, D. J., and Hubley-Kozey, C. L. (2013). Reliability of principal components and discrete parameters of knee angle and moment gait waveforms in individuals with moderate knee osteoarthritis. *Gait Posture* 38, 421–427. doi:10.1016/j.gaitpost.2013.01.001

Saliba, C. M., Clouthier, A. L., Brandon, S. C. E., Rainbow, M. J., and Deluzio, K. J. (2018). Prediction of knee joint contact forces from external measures using principal component prediction and reconstruction. *J. Appl. Biomech.* 34, 419–423. doi: 10.1123/jab.2017-0262

Schwab, F., Lafage, V., Boyce, R., Skalli, W., and Farcy, J. P. (2006). Gravity line analysis in adult volunteers: age-related correlation with spinal parameters, pelvic parameters, and foot position. *Spine* 31, E959–E967.

Shu, L., Yamamoto, K., Yao, J., Saraswat, P., Liu, Y., Mitsuishi, M., et al. (2018). A subject-specific finite element musculoskeletal framework for mechanics analysis of a total knee replacement. *J. Biomech.* 77, 146–154. doi: 10.1016/j.jbiomech.2018.07.008

Smoger, L. M. (2016). *Statistical Modeling to Investigate Anatomy and Function of the Knee.* Denver: Thesis.

Spencer-Gardner, L., Pierrepont, J., Topham, M., Baré, J., McMahon, S., and Shimmin, A. (2016). Patient-specific instrumentation improves the accuracy of acetabular component placement in total hip arthroplasty. *Bone Joint J.* 98, 1342–1346. doi: 10.1302/0301-620x.98b10.37808

Styner, M. A., Rajamani, K. T., Nolte, L. P., Zsemlye, G., Szekely, G., Taylor, C. J., et al. (2003). Evaluation of 3D correspondence methods for model building. *Inf. Process. Med. Imaging* 18, 63–75. doi: 10.1007/978-3-540-45087-0_6

Taylor, M., Barrett, D. S., and Deffenbaugh, D. (2012). Influence of loading and activity on the primary stability of cementless tibial trays. *J. Orthopaedic Res.* 30, 1362–1368. doi: 10.1002/jor.22056

Vaitkus, M., and Várady, T. (eds) (2015). "A general framework for constrained mesh parameterization," in *Proceedings of the 31st Spring Conference on Computer Graphics*, (New York, NY: ACM).

Vasco, M. (2012). Permutation tests to estimate significances on Principal Components Analysis. *Comput. Ecol. Softw.* 2:103.

Wang, J., and Shi, C. (2017). Automatic construction of statistical shape models using deformable simplex meshes with vector field convolution energy. *Biomed. Eng. Online* 16:49. doi: 10.1186/s12938-017-0340-0

Waxman, A. (2004). *WHO Global Strategy on Diet, Physical Activity and Health.* Geneva: World Health Organization.

Worsley, P., Stokes, M., and Taylor, M. (2011). Predicted knee kinematics and kinetics during functional activities using motion capture and musculoskeletal modelling in healthy older people. *Gait Posture* 33, 268–273. doi: 10.1016/j.gaitpost.2010.11.018

Zhang, J., Fernandez, J., Hislop-Jambrich, J., and Besier, T. F. (2016). Lower limb estimation from sparse landmarks using an articulated shape model. *J. Biomech.* 49, 3875–3881. doi: 10.1016/j.jbiomech.2016.10.021

A Machine Learning Approach to Estimate Hip and Knee Joint Loading Using a Mobile Phone-Embedded IMU

Arne De Brabandere[1]*, Jill Emmerzaal[2,3], Annick Timmermans[3], Ilse Jonkers[2], Benedicte Vanwanseele[2] and Jesse Davis[1]

[1] Department of Computer Science, KU Leuven, Leuven, Belgium, [2] Department of Movement Sciences, KU Leuven, Leuven, Belgium, [3] Faculty of Rehabilitation Sciences, Hasselt University, Hasselt, Belgium

Hip osteoarthritis patients exhibit changes in kinematics and kinetics that affect joint loading. Monitoring this load can provide valuable information to clinicians. For example, a patient's joint loading measured across different activities can be used to determine the amount of exercise that the patient needs to complete each day. Unfortunately, current methods for measuring joint loading require a lab environment which most clinicians do not have access to. This study explores employing machine learning to construct a model that can estimate joint loading based on sensor data obtained solely from a mobile phone. In order to learn such a model, we collected a dataset from 10 patients with hip osteoarthritis who performed multiple repetitions of nine different exercises. During each repetition, we simultaneously recorded 3D motion capture data, ground reaction force data, and the inertial measurement unit data from a mobile phone attached to the patient's hip. The 3D motion and ground reaction force data were used to compute the ground truth joint loading using musculoskeletal modeling. Our goal is to estimate the ground truth loading value using only the data captured by the sensors of the mobile phone. We propose a machine learning pipeline for learning such a model based on the recordings of a phone's accelerometer and gyroscope. When evaluated for an unseen patient, the proposed pipeline achieves a mean absolute error of 29% for the left hip and 36% for the right hip. While our approach is a step in the direction of using a minimal number of sensors to estimate joint loading outside the lab, developing a tool that is accurate enough to be applicable in a clinical context still remains an open challenge. It may be necessary to use sensors at more than one location in order to obtain better estimates.

Keywords: machine learning, inertial measurement units, joint loading, patient monitoring, hip osteoarthrithis

1. INTRODUCTION

Hip osteoarthritis (OA) patients exhibit changes in kinematics and kinetics that affect the contact forces of the hip and knee joints during walking and daily activities. It is believed that these changes are important in the progression of OA (Felson, 2013) and that monitoring these changes during daily life could provide valuable information to clinicians. For example, a patient's joint loading

measured across different exercises can serve as an indication for the number of exercise repetitions that the patient needs to complete when rehabilitating after hip arthroplasty surgery. Despite the importance of joint loading monitoring, it is difficult to systematically and widely measure joint loading in a clinical environment. First, acquiring these measurements requires a lab environment consisting of optoelectronic cameras and ground reaction force plates. The cost and space required for such a setup makes this impractical to install in a clinician's practice. Second, it would be infeasible to analyze a large number of patients in a lab since collecting and processing the data is a time-consuming task. Third, in order to calculate joint contact forces, one would need to use a musculoskeletal modeling workflow, which requires expert knowledge.

Because of these drawbacks, clinicians could greatly benefit from a mobile system that is able to provide accurate joint loading estimates based on cheaper sensors. Ideally, such a system would be based on inexpensive, wearable sensors that the patients can easily use at the clinician's practice or even at home. Inertial measurement unit (IMU) sensors and electromyography (EMG) sensors are ideal candidates for this purpose as they are relatively cheap and have been applied successfully in a wide range of human movement analysis tasks (Zhang et al., 2011; Camomilla et al., 2018; De Brabandere et al., 2018; Op De Beéck et al., 2018). Designing such a system requires collecting data in a lab setting where a subject performs the relevant exercises while simultaneously recording data from the cheap portable sensors and the expensive, standard lab sensors. This enables either hand-crafting a model or applying a data-driven approach such as machine learning to learn a model that relates the data produced by the portable sensors to the ground truth joint loading estimated from the lab equipment. These predictive models can then be deployed outside the lab as they can make predictions about a subject's joint loading solely based on the data captured by cheaper sensors.

Related studies have proposed different models for estimating joint loading from wearable sensors. de Vries et al. (2012) proposed a neural network model which estimates several loading variables for the shoulder joint based on kinematics and EMG data. The kinematics were measured using four IMU sensors. While the model can be used in an ambulatory setting, it still requires a relatively large number of sensors. Moreover, the EMG measurement is somewhat intrusive as it requires attaching 13 electrodes to the person's body. Other work by Wesseling et al. (2018) proposed a model for estimating hip and knee joint contact forces based on IMU kinematics and ground reaction force (GRF) data. They found that the IMU kinematics were sufficient to estimate the hip contact forces reliably, which enables using the model outside a lab. However, the knee contact force model required both the IMU and GRF data. Hence, this has the same drawbacks as the lab sensors for calculating joint contact forces as it is challenging to measure GRF data in the wild. Stetter et al. (2019) proposed a model for predicting knee joint loading using two IMU sensors, one on the upper leg and one on the lower leg. However, similar to de Vries et al. (2012) and Wesseling et al. (2018), they evaluated the model on data from healthy subjects only. Applying the same model to patients may not

work due to altered movement patterns. Other studies considered similar problems, such as estimating the daily cumulative joint loading (Robbins et al., 2009) and ground reaction forces (Guo et al., 2017; Karatsidis et al., 2017; Wouda et al., 2018).

The goal of this paper is to predict the joint loading of the left and right hip and knee based on IMU data collected from a mobile phone. First, we collect data using three types of sensors simultaneously: a hip-mounted phone, optoelectronic motion capture cameras and ground reaction force plates. We use the latter two to calculate the ground truth joint loading using a musculoskeletal modeling workflow. Second, we employ machine learning to automatically construct a model that can predict the ground truth joint loading on the basis of the IMU data collected from the mobile phone. Our approach confers two advantages over prior work. First, by relying on a mobile phone it both builds off an omnipresent technology and minimizes the number of required sensors. Hence, clinicians and possibly even patients will not need to rely on expensive specialized equipment. Second, we focus on hip OA patients instead of healthy subjects. Since clinicians see patients with abnormal movement patterns, we train and evaluate the model using data collected from a representative patient group.

2. METHODS

2.1. Subjects

For this study, 20 patients with unilateral end-stage hip osteoarthritis were recruited from a local hospital (Ziekenhuis Oost Limburg, Belgium). They were included based on the following criteria: aged between 55 and 75 years; unilateral hip osteoarthritis; awaiting joint replacement surgery; Body Mass Index $\leq$ 30kg $\cdot$ m^{-2}; able to walk 10 m; no cortiosteroid injection 3 months prior to inclusion; no joint replacements and no other musculoskeletal or neurological disorders that would affect movement pattern. Participants provided written informed consent prior to the start of the measurements. Out of these 20 patients, we select only those for which the mobile phone measurements were recorded correctly throughout the whole protocol, which corresponds to a subset of 10 patients. The ethical committee of the academic hospital Leuven approved the study (reference no. s-59857).

2.2. Protocol

Each patient performed multiple repetitions of nine types of exercises. **Table 1** shows the number of repetitions per exercise. The exercise types are defined as follows:

- **Walk**: level walking at a self-selected speed, one repetition corresponds to one stride;
- **Ascend stairs** and **descend stairs**: at a self selected speed, without hand-held support on a standardized 4-step staircase, one repetition corresponds to one stride;
- **Sit down** and **stand up**: the height of the chair was standardized to participants knee height;
- **Forward lunge** and **side lunge**: step length standardized at 70% leg length;

TABLE 1 | Number of exercise repetitions per subject.

Subject ID	W	AS	DS	SD	SU	FL	SL	SOL	SQOL
1	13	9	10	10	10	10	10	10	9
2	10	9	9	10	10	8	10	10	10
3	10	10	7	11	11	10	10	10	10
4	10	10	10	10	10	10	9	10	9
5	10	10	9	10	10	9	10	10	10
6	10	12	8	10	10	10	9	10	10
7	10	10	10	10	10	10	10	10	10
8	10	9	10	10	10	10	10	11	10
9	13	10	12	10	10	10	10	10	12
10	10	10	10	10	10	10	10	10	10

W, walk; AS, ascend stairs; DS, descend staris; SD, sit down; SU, stand up; FL, forward lunge; SL, side lunge; SOL, stand on one leg; SQOL, squat on one leg.

- **Stand on one leg** (approx. 2 s) and **squat on one leg**: hands fixed at the side.

2.3. Joint Loading

We measure the patients' hip and knee contact force to define the ground truth joint loading that we aim to estimate. While contact forces can be measured directly using instrument prostheses, we instead use a combination of experimental data and musculoskeletal modeling (Fregly et al., 2012) since the direct method is an invasive procedure. Moreover, this method requires total joint replacement, which would limit the number of patients we can analyze. The remainder of this section describes the procedure for calculating the contact forces. Validation studies by Wesseling et al. (2016) and Zargham et al. (2019) have shown that this procedure results in accurate estimates.

The experimental data was collected using 13 optoelectronic cameras (Vicon, Oxford Metrics, UK, 100 Hz) and three ground reaction force plates embedded in the floor (AMTI, Watertown, MA, USA, 1,000 Hz). Each participant was equipped with 38 reflective markers on bony landmarks conforming to the full-body plug-in walk model (Oxford Metrics). The single markers on the body segments were substituted by rigid three marker clusters. The marker trajectories and ground reaction force data were used as input in a standard musculoskeletal modeling workflow applied in OpenSim 3.3 (Delp et al., 2007). First, the generic OpenSim model `gait2392` (Delp et al., 1990) was used. We added a degree of freedom in the knee joint (i.e., ab/adduction) and implemented a functional knee axis of rotation (Meireles et al., 2017). The model was scaled to match the height and weight of the participant. Joint kinematics were derived from marker trajectories using inverse kinematics analysis with a Kalman smoothing algorithm (De Groote et al., 2008). Subsequently, joint moments were calculated with the inverse dynamic analysis using the calculated joint angles and measured ground reaction forces. Muscle force and muscles activation were determined using static optimization. Lastly, the joint contact forces were calculated using the vector sum of the estimated muscle forces and joint reaction forces (Steele et al., 2012).

Since our goal is to build a workflow that estimates the joint loading for one repetition of an exercise, we aggregate the contact forces by extracting the joint impulse y. This variable is defined as the integral of the contact force signal, relative to the subject's body weight:

$$y = \frac{\int_0^T CF_t \, dt}{m \cdot 9.81}$$

where CF_t is the joint contact force at time t, T is the duration of one exercise repetition, and m is the body mass. We compute the joint impulse for the left and right hip and knee.

2.4. Input Signals

For the input data of our joint impulse estimation models, we use inertial measurement unit (IMU) sensors since they are easy to use outside the lab. IMU sensors are often used in human motion analysis for this reason (Bussmann et al., 2001; Weyand et al., 2001; Alvarez et al., 2008; Camomilla et al., 2018). In addition, they are relatively inexpensive to buy compared to the lab equipment needed to calculate joint contact forces.

In this study, we use the IMU sensors from a mobile phone (Samsung Galaxy J5 2017). During the whole exercise protocol, the phone continuously recorded the 3D acceleration (a_x, a_y, a_z) and 3D angular velocity (g_x, g_y, g_z), both with a sampling rate of 50 Hz[1]. The phone was attached to the patient's left hip using a velcro strap around the patients' hips. While our goal is to predict the joint loading on both sides, we also wanted to use a simple setup with the minimal number of sensors. Hence, we only use one sensor and always attach it on the same side of the body. Since people usually wear their phone in a pocket, we chose the left hip to mimic that placement. The phone was attached such that the IMU's reference frame corresponded to the anterior-posterior (x), proximal-distal (y) and lateral-medial (z) direction of the person's left leg.

Since the signals change over time, each signal is represented as a time series, i.e., a sequence of values. For example, the a_x acceleration signal corresponds to the time series $[a_x(t_0), a_x(t_1), ..., a_x(t_n)]$ where t_i is the ith time stamp of an exercise repetition.

2.5. Synchronization

Whereas the optoelectronic cameras and the ground reaction force plates were connected to the computer that was used for measuring the joint contact forces, the mobile phone sensor recordings were collected directly on the phone. Since the computer and the mobile phone recorded the data independently, their recordings were not synchronized through a single clock. In order to link to correct parts of the sensor data to the joint contact forces, both systems' time stamps have to be aligned. This can be done by incrementing the time stamps of the phone by the lag between the two clocks, i.e., the difference between the computer's clock time and the phone's clock time.

---

[1] We recorded the data from the `TYPE_ACCELEROMETER` and `TYPE_GYROSCOPE` sensor types of Android. As shown in the API documentation (https://developer.android.com/guide/topics/sensors/sensors_motion). Android corrects for drift in the sensor measurements.

Unfortunately, the exact lag was unknown at the time of data collection. Finding this lag manually would require to check for each possible lag whether the joint contact force signal is aligned with the phone's signals and select the lag that results in the best alignment. Additionally, since the lags vary across the different collection sessions due to drift in the phone's clock time, this would have to be repeated for each subject. Therefore, we align the signals automatically using an approach based on the cross-correlation coefficient between the signals. Specifically, we compute the cross-correlation for each possible lag, i.e., xcorr$(P_{l:l+n_{CF}}, CF)$ for each lag $l \in [0, n_P - n_{CF}]$, which corresponds to the lags l for which all time stamps of the contact force data CF are between the start and end of the phone data P. **Figure 1** illustrates the synchronization approach.

2.6. Pipeline

Figure 2 shows our machine learning pipeline for predicting the joint impulse based on the phone's signals. The input and output of the pipeline are defined as follows:

- **Input:** The measurements of the phone's IMU collected during a single exercise repetition.
- **Output:** The joint impulse at the left hip, right hip, left knee or right knee. Each location corresponds to one target variable, i.e., the goal is to predict one value per location. Since the four locations may require different models, we develop a separate pipeline (with the same building blocks) for each target.

The pipeline consists of three building blocks. First, the *feature extraction* block converts the raw phone signals into a format that is suitable for learning a model. This format consists of features that summarize the phone signals, e.g., by extracting the mean of the a_x signal. Each feature summarizes the data of one exercise repetition, i.e., one window of data. The process of defining a set of relevant features is called feature construction. Section 2.6.1 describes this process in more detail. Next, the *normalization* block normalizes the feature values in order to make the predictions more robust. Section 2.6.2 lists several normalization procedures. Finally, the *model* block turns the (normalized) feature values into a prediction for the left/right hip/knee joint impulse. Since the relation between phone-based features and joint impulse is unknown from a biomechanics perspective, we use machine learning to automatically learn a model from a dataset labeled with ground-truth joint impulses. Sections 2.6.3 and 2.6.4 describe the learning settings and methods for training these models.

2.6.1. Feature Construction

The input of the pipeline consists of the measurements collected by the phone's sensors. However, the high dimensionality of the raw phone signals prevents using these signals directly for training a model. Therefore, we follow a feature-based approach (Fulcher, 2018) and convert the raw phone signals into a low-dimensional feature representation which captures the relevant characteristics of the signals.

We use the TSFuse Python package with the minimal feature extraction settings to generate a feature representation. This package extracts a set of statistical features (e.g., mean, median,

variance,...) from both the original signals and additional signals derived from these signals. To derive new signals, TSFuse combines multiple signals using different transformations (e.g., the resultant of three signals). We refer to De Brabandere et al. (2019) for the complete list of transformations as well as the feature construction algorithm which builds features using these transformations.

Since this construction method uses the target data to remove irrelevant features, the feature construction method was repeated for each cross-validation fold (see section 2.7). In our experiments, TSFuse constructed the same 63 features in each fold. **Supplementary Table 1** shows an overview of the constructed features.

2.6.2. Normalization Procedures

Normalizing the feature values may be required from a machine learning and biomechanics perspective. From a machine learning perspective, standardizing the feature values to a similar range is necessary for certain types of models, including the regularized linear model of our pipeline (section 2.6.4). From a biomechanics perspective, other studies using accelerometer data have shown that individual differences (e.g., body height, body mass, movement pattern, ...) may influence the signals and thus affect the feature values as well (Op De Beéck et al., 2018).

For the features, we consider the following normalization procedures:

- No normalization: use the original feature values.
- Dataset-level standardization: standardize each feature using the mean and standard deviation as computed over the complete dataset. This procedure only accounts for differences in the range of the features.
- Subject-level standardization: standardize each feature using the mean and standard deviation as computed separately for each subject. This procedure also accounts for differences between subjects.

For the target data, we only consider (1) the original joint impulses relative to subject's body weight, and (2) the standardized impulses using the mean and standard deviation over the complete training set. We do not consider standardizing based on each subject's joint impulses since that would require ground truth joint loading measurements for the test data, which is understandable as the model does not have these when applied to an unseen subject.

2.6.3. Learning Settings

Since some exercises have completely different movements, the joint impulse can not be modeled in the same way for each exercise. The model could detect the exercise type itself by training the model using the complete dataset. However, given the small dataset size, we simplify the learning task by training multiple models, each focusing on only one or a few similar exercises. Specifically, we consider the following two learning settings:

FIGURE 1 | Illustration of our synchronization method for one of the subjects. The original signals **(left)** are the resultant acceleration measured by the phone and the left hip contact force. The cross-correlation between these signals **(middle)** is computed for all possible lags, i.e., all lags for which the hip contact force signal still ends before the acceleration signal ends. The location of the highest peak then corresponds to the time difference between the signals, which can be used to align the signals **(right)**.

FIGURE 2 | Joint loading estimation pipeline.

- **One exercise (OE)**
  This setting splits the data per exercise type and evaluates models for each exercise type separately.
- **Similar exercises (SE)**
  This setting splits the dataset in groups of similar exercises: walk, ascend stairs and descend stairs; sit down and stand up; forward lunge and side lunge; and stand on one leg and squat on one leg.

Grouping multiple exercises in the SE setting increases the number of training examples compared to the OE setting, which may help selecting relevant features and setting good parameters for the model. We hypothesize that the SE setting yields more accurate models as a result of the increased training set size. To evaluate this hypothesis, section 3.4 compares both settings.

2.6.4. Learning Methods
To estimate the joint loading based on the phone's data, we train regularized linear regression models using the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996). This method performs both regularization and feature selection by including the ℓ_1-norm of the weights in the cost function. Given the small dataset of this study, this method is suitable as it is able to select relevant features from a large number of features (p) when the number of training examples (n) is small ($n \ll p$). In our experiments, we use the `Lasso` implementation of `scikit-learn` (Pedregosa et al., 2011) with the default parameters, which sets the regularization constant `alpha` to 1.

We compare the linear regression models to a naïve baseline model which predicts the average joint impulse of all exercise repetitions in the training data. As the baseline requires no learning, achieving a lower prediction error is a minimal requirement for the linear model to do better than the current best approach for monitoring the joint loading of patients. This approach uses the population average as a "joint loading profile" for monitoring an individual patient. The naïve baseline estimates the population average from a specific group of subjects, in this case a sample of hip OA patients.

2.7. Evaluation

We evaluate the pipeline's performance on unseen (i.e., future) data with respect to two scenarios: (1) applying the model to an *unseen patient*, and (2) applying the model to a *seen patient*, i.e., a patient for whom some labeled data is already available. The first scenario is relevant when a doctor without lab access applies the model to one of his patients. Since this patient's movement patterns may be different compared to the patients for whom the model was trained, we hypothesize that the second scenario may improve the predictions by including labeled data of the patient in the training data. To evaluate these scenarios, we employ the following cross-validation procedures:

- **Leave-one-subject-out cross-validation**
  This cross-validation procedure evaluates how accurate the pipeline works for an *unseen* patient. In each fold, we hold out all data of a single patient and train the model using all other patients' data. The error averaged over all folds corresponds to the prediction error that a doctor without lab access can expect.
- **Leave-one-exercise-type-out cross-validation**
  This cross-validation procedure evaluates how accurate the pipeline works for a *seen* patient. This procedure splits the data based on the exercise type. In each fold, the test data consists of all repetitions of one exercise type performed by one subject. The training data consists of all other exercises performed by the same subject as well as all data of the other subjects. Note that we do not consider leave-one-repetition-out cross-validation: since all trials of each exercise were performed consecutively, the dependency between trials may be too strong and result in overly optimistic errors.

For both cross-validation schemes, we evaluate the models by reporting the relative mean absolute error (MAE%) of the estimated joint impulses $\hat{y}_i$ w.r.t. the ground truth joint impulses y_i over all exercise repetitions i. This metric represents the average relative deviation from the actual joint impulses over all exercise repetitions performed by a patient. The MAE% is defined as follows:

$$\text{MAE\%} = \sum_i^N \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

3. RESULTS

This section evaluates the proposed joint impulse prediction pipeline. We evaluate the pipeline using both cross-validation procedures in section 3.1 (leave-one-subject-out) and section 3.2 (leave-one-exercise-type-out). For the pipeline's building blocks, we use dataset-level standardization for the feature values and train the models using the SE setting. Our comparison in sections 3.3 and 3.4 shows that this normalization procedure and learning setting were found to be optimal choices for our dataset.

3.1. Error for Unseen Patients

Table 2 shows the MAE% for the joint impulse at each of the four locations. Overall, the linear model outperforms the baseline for the hip joint loading. However, the knee joint loading seems

TABLE 2 | MAE% evaluated using leave-one-subject-out cross-validation.

Exercise	Method	Left hip	Right hip	Left knee	Right knee
Walk	Baseline	0.439	0.460	0.286	0.291
	Linear	**0.168**	**0.155**	0.430	0.406
Ascend stairs	Baseline	0.158	0.077	0.193	0.076
	Linear	0.158	0.077	0.193	0.183
Descend stairs	Baseline	0.184	0.340	0.207	0.164
	Linear	0.184	**0.227**	0.319	0.478
Sit down	Baseline	0.360	0.324	0.372	0.214
	Linear	0.360	0.324	**0.279**	0.214
Stand up	Baseline	0.296	0.204	0.269	0.142
	Linear	0.296	0.204	0.269	0.142
Forward lunge	Baseline	0.280	0.300	0.208	0.337
	Linear	**0.263**	**0.265**	**0.178**	**0.256**
Side lunge	Baseline	0.277	0.293	0.254	0.314
	Linear	0.325	0.330	**0.243**	**0.269**
Stand on one leg	Baseline	0.461	0.469	0.531	0.352
	Linear	0.531	**0.315**	0.744	0.401
Squat on one leg	Baseline	0.278	1.031	0.291	1.986
	Linear	**0.251**	1.081	**0.223**	**1.811**
Overall	Baseline	0.314	0.417	0.297	0.483
	Linear	**0.290**	**0.360**	0.321	**0.482**

The errors which outperform the baseline are highlighted in bold.

TABLE 3 | MAE% for the two cross-validation schemes.

	Left hip	Right hip	Left knee	Right knee
Leave-one-subject-out	**0.290**	**0.360**	0.321	0.482
Leave-one-exercise-type-out	0.296	0.407	**0.295**	0.482

For each location, the lowest error is highlighted in bold, if one cross-validation method outperforms the other.

harder to estimate as the linear model is marginally more accurate than the baseline for right knee and even less accurate than the baseline for the left knee. Evaluating the error for each exercise type separately shows that the results are different across different exercise types. The hip joint impulse estimations for walking show the largest improvement over the baseline compared to the other exercises.

3.2. Error for Seen Patients

Table 3 compares the leave-one-subject-out cross-validation scheme with the leave-one-exercise-type-out cross-validation. We hypothesized that the leave-one-exercise-type-out cross-validation could improve the predictions by including data of the patient in the test data. Unfortunately, the leave-one-exercise-type-out errors are close to the leave-one-subject-out errors and for the left and right hip, the leave-one-subject-out models often outperform the leave-one-exercise-type-out models.

3.3. Comparison of Normalization Procedures

In section 2.6.2, we hypothesized that normalization procedures can improve the error of the models by scaling features

TABLE 4 | Overall MAE% (averaged over all locations, i.e., left/right hip/knee) for different combinations of the normalization procedures.

		Target normalization	
		No	Yes
	No	0.439	0.433
Feature normalization	Dataset-level	**0.363**	0.391
	Subject-level	0.371	0.391

The table shows the results for the linear models with the SE setting evaluated using leave-one-subject-out cross-validation. The lowest error over all combinations is highlighted in bold.

to a similar range and removing inter-individual differences. **Table 4** compares all possible combinations of the normalization procedures for both the features and the target data by reporting the overall MAE% averaged over all locations (left and right hip and knee) for each of combination. The best performing combination is the dataset-level feature standardization and no target normalization. Surprisingly, subject-level standardization does not result in more accurate models compared to dataset-level standardization.

3.4. Comparison of Learning Settings

In section 2.6.3, we hypothesized that the SE setting yields more accurate results as this setting increases the number of training examples by combining multiple exercises. To evaluate this hypothesis, **Table 5** compares the SE setting with the OE setting. Overall, the SE models are more accurate than the OE models for all locations except for the left knee. For the hip joint impulse, the SE models show the largest improvement for walking. However, the results are slightly less accurate for other exercises (e.g., forward lunge and side lunge) which indicates that the SE models are suitable for walking but not for other exercises.

4. DISCUSSION

The goal of this study was to explore the possibility of using a minimal number of sensors for predicting joint loading in hip OA patients. We proposed a machine learning pipeline that requires only the IMU data collected from a mobile phone. In this section, we discuss our choices for the different building blocks of this pipeline. We then discuss the differences in the obtained errors with respect to the joints and exercise types. Finally, we discuss the accuracy vs. ease-of-use trade-off of our approach and suggest future directions with respect to this trade-off.

4.1. Building Blocks of the Machine Learning Pipeline

The proposed machine learning pipeline required making several decisions for the different building blocks. For the *feature extraction* block, we used an automated approach (De Brabandere et al., 2019) to define the feature representation. For the *normalization* block, we compared different normalization procedures and found that dataset-level feature standardization was important. For the *model* block, we only considered a linear regression model, since the small dataset

TABLE 5 | MAE% of the similar exercises (SE) models and one exercise (OE) models.

Exercise	Setting	Left hip	Right hip	Left knee	Right knee
Gait	OE	0.439	0.460	0.286	0.291
	SE	**0.168**	**0.155**	0.430	0.406
Ascend stairs	OE	0.158	0.077	0.193	0.076
	SE	0.158	0.077	0.193	0.183
Descend stairs	OE	0.184	0.340	0.207	0.164
	SE	0.184	**0.227**	0.319	0.478
Sit down	OE	0.360	0.324	0.380	0.214
	SE	0.360	0.324	**0.279**	0.214
Stand up	OE	0.296	0.204	0.269	0.142
	SE	0.296	0.204	0.269	0.142
Forward lunge	OE	0.202	0.265	0.174	0.264
	SE	0.263	0.265	0.178	**0.256**
Side lunge	OE	0.277	0.293	0.254	0.314
	SE	0.325	0.330	**0.243**	**0.269**
Stand on one leg	OE	0.462	0.292	0.582	0.352
	SE	0.531	0.315	0.744	0.401
Squat on one leg	OE	0.196	1.167	0.242	2.058
	SE	0.251	**1.081**	**0.223**	**1.811**
Overall	OE	0.296	0.407	0.295	0.482
	SE	**0.290**	**0.360**	0.321	0.482

The SE errors which outperform the OE errors are highlighted in bold.

size prevented us from using non-linear models. Whereas we used the pipeline for predicting the joint impulses of the hip and knee, it could be relevant for other locations as well. It could also be interesting to explore whether this pipeline (potentially with a non-linear model) can be used for other types of exercises and for other types of sensors as input.

4.2. Errors Across Different Joints

The results of **Table 2** show a clear difference in accuracy for the hip and knee joints. The obtained results indicate that the proposed pipeline is able to predict the hip impulse, but it remains hard to outperform the naïve baseline for the knee impulse. Perhaps placing the IMU closer to the target joint might lead to better results in predicting knee contact forces. An IMU sensor on the hip might not capture the higher linear accelerations and angular velocities that are found on the segments connected to the knee joint. Considering the body's ability to attenuate shock, the acceleration signal amplitude has already weakened when reaching the IMU placed at hip level (Kavanagh and Menz, 2008). Placing an IMU on the shank could better capture these initial loading shocks (distal part of the shank), or higher acceleration signals (middle part of the shank). However, which placement is best to obtain better joint loading predictions should be investigated. Therefore, different IMU placements should be investigated to examine if personalizing the placement based on the type of patient (i.e., hip or knee osteoarthritis patient) leads to better joint loading prediction results during these types of exercises.

Similar to the difference between the hip and knee, there are also different errors for the left and right side. Interestingly, this difference does not only hold for the linear models, but also for the baselines, which suggests that there may be a larger variability in the joint impulses for the right side compared to the left side. One possible reason could be the side of the hip that was affected. However, this is unlikely as the right hip was affected for 6 patients and the left hip for 4 patients. To evaluate whether the difference between left and right is significant, we performed a two-sided paired t-test for the overall MAE% of the baseline. That is, we tested whether $[x_1, \ldots, x_{10}]$ is significantly different from $[y_1, \ldots, y_{10}]$ where x_i is subject i's MAE% of the baseline for the left hip/knee and y_i is subject i's MAE% of the baseline for the right hip/knee. The resulting p-values are 0.3694 (for the hip) and 0.4458 (for the knee) meaning that the difference between left and right was not significant. Most likely, the difference is due to the small sample size (only 10 subjects) and does not hold in general.

4.3. Errors Across Different Exercise Types

The errors of the linear model are different for the different exercise types, which suggests that predicting joint impulses is easier for some exercises compared to others. Given that there are large differences in the movements between exercises, the differences in the joint impulse prediction errors can depend on the relation between the data collected by the phone and the contact force at each point in time. **Figure 3** shows these contact forces along with the resultant acceleration for each exercise type. For those exercises for which both the left and right hip joint impulse predictions are better than the baseline (walk and forward lunge), the contact force signal shows two main peaks for which also the resultant acceleration has a clear peak. The potential relation between the height of these acceleration peaks and the height of the contact force peaks could make the prediction of the joint impulses easier.

4.4. Selected Features

The linear models in the *model* block were trained using L1-regularization (lasso), which select a small number of features out of the 63 generated features. Since the selected features are different for each fold, it is hard to visualize which features are used in the models given the large number of models (10 subjects, 4 groups of exercises and 4 locations result in 160 models). Instead, we run stability selection (Meinshausen and Bühlmann, 2010) for each group of exercises and for each location to get an idea of which features were selected most often. Stability selection repeatedly trains Lasso models (with random subsampling) and reports each feature's importance as the percentage of models in which the feature was used. **Supplementary Table 2** shows the top 5 features with the corresponding importance scores, for each location and each group of exercises. One observation is that sum and length features are commonly used. Since the joint impulse is defined as an integral of the contact force, it is expected that this feature is important to capture the duration of the exercise repetition. Unfortunately, it is hard to interpret the importance of the other selected features. Future work could explore using more specific (manually handcrafted) features when the goal is to get a better insight in the learned models.

4.5. Trade-Off Between Accuracy and Ease-of-Use

This study explores a trade-off between accuracy and ease-of-use. The most accurate model would be the one that uses all lab equipment needed for calculating joint contact forces using a musculoskeletal modeling workflow. However, this model would also be the most inconvenient as it requires the patient to come to the lab (which is probably not located in the hospital), attach 38 reflective markers to the patient and analyze the collected data in order to calculate the joint contact forces from the collected measurements.

Our model only requires attaching a mobile phone to the patient's left hip[2]. Given that a patient consultation typically takes approximately only 15 min, using a small number of sensors is an important requirement for developing a joint loading estimation tool. In addition, using a mobile phone reduces the cost of such a tool, since clinicians most likely already own a mobile phone and only need to install an app to apply the model.

However, given the results of this work, we recognize that using a mobile phone may be an easy solution, but unfortunately, one that is not accurate enough for valid clinical use. A better compromise between accuracy and ease-of-use would be to use a combination of IMU sensors at more than one location. This would allow having a better view of the patient's movements. Still, the number of sensors should be kept to a minimum in order to keep the tool practical. More research is needed to evaluate which locations are most suitable.

Even though the results are far from perfect, we argue that our phone-based model is a step in the right direction in estimating joint loading in a clinical setting using a very limited amount of sensors. Especially the results for predicting the joint impulse during level walking are interesting, where distinct reduction in mean absolute error from the baseline can be seen (MAE% from 43.9 to 16.8%). When monitoring a patient during daily life, this result is promising as walking is one of the most commonly performed daily activities and might be responsible for the majority of the joint loading during a day. The improvement over the baseline indicates that clinicians are able to obtain more accurate estimates of a patient's joint load compared to using a population average. In the future, a "hip OA" profile using population averages could shift to a "personal" profile using a more individualized estimate of joint loading. This in turn could help clinicians align a person's exercise prescription to their individual loading profile based on more accurate methods which could improve their rehabilitation. Given that joint contact forces are believed to be important in the initiation and progression of OA (Felson, 2013), this might be a promising tool in the rehabilitation setting to asses patients' joint impulses during walking over time and adjust the rehabilitation and exercise prescription accordingly.

---

[2]Alternatively, the patient could wear the phone in his left pocket, but further research is needed to evaluate whether this does not decrease the accuracy of the estimations. Both ways of wearing the phone are convenient in a clinician's practice.
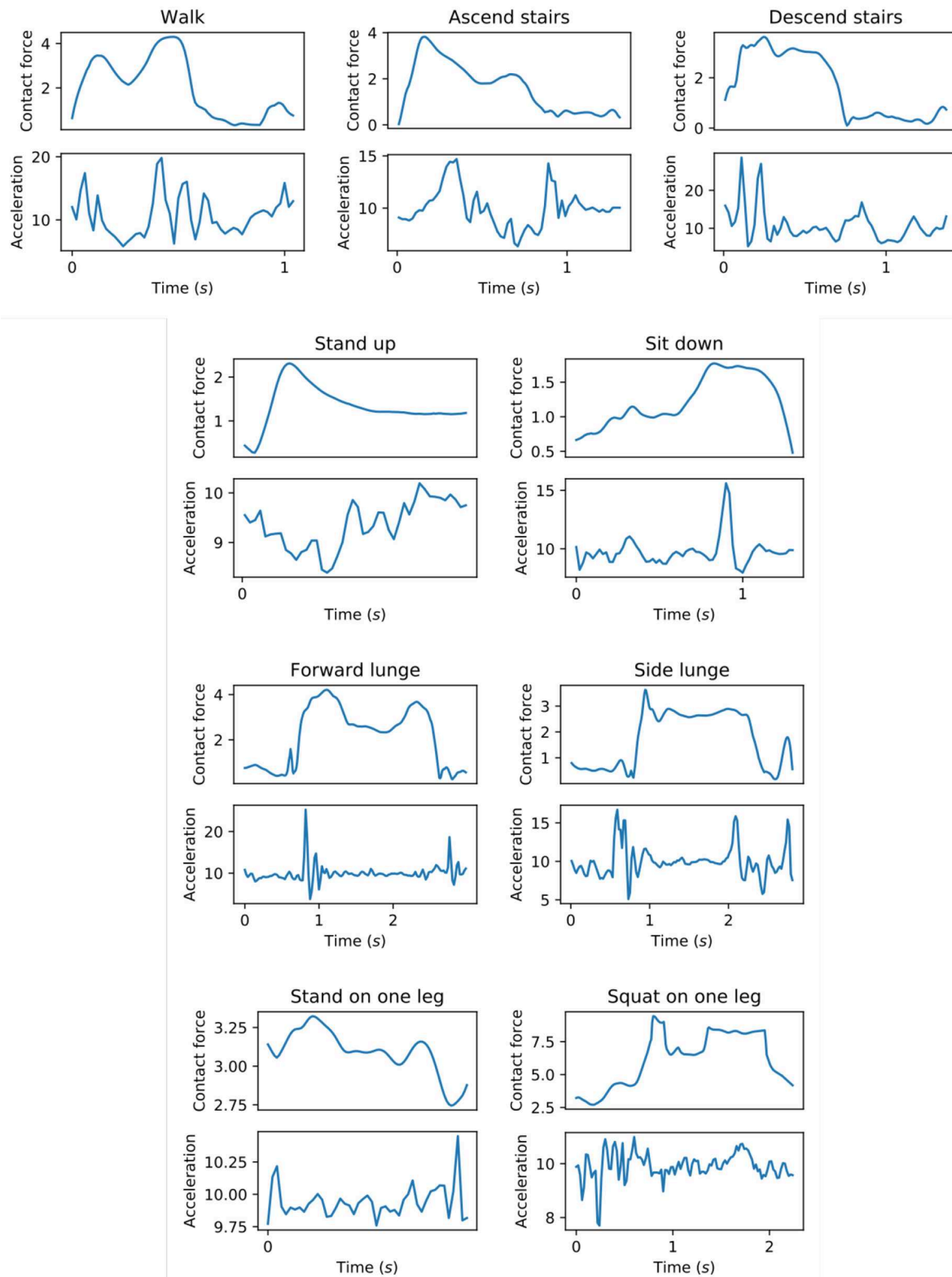
FIGURE 3 | Hip contact force (*N/kg*) along with the resultant acceleration (*m/s*^{2}) as measured by the mobile phone. For each exercise type, the figure shows a single repetition performed by one subject.

5. CONCLUSION AND FUTURE WORK

This work presented a machine learning pipeline to estimate the hip and knee joint impulse based on a mobile phone. In terms of the mean absolute error, we found that the proposed pipeline is able to slightly outperform a population average baseline for the hip (left hip: 29.0% for the linear model vs. 31.4% for the baseline; left hip: 36.0 vs. 41.7% for the right hip), but not for the knee. Our approach has two key advantages over existing methods for predicting joint loading. First, the proposed pipeline only requires a mobile phone as input. Second, we trained and evaluated the pipeline using data of patients instead of healthy subjects, which is relevant with respect to the setting in which the proposed pipeline is applicable, i.e., monitoring patients.

However, even though our phone-based model is a step in the direction of estimating joint contact forces using a minimal number of sensors, the current approach still has several limitations that need to be addressed in future work. First, the overall error of our approach should be reduced further in order to be applicable in a clinical context. One possibility is to use multiple sensors, but still only a few such that the model remains easy to use. Related work by Wesseling et al. (2018) has shown that a combination of six IMU kinematic variables can estimate hip joint loading but that for accurate knee joint loading estimates both kinematic variables and ground reaction forces are needed. Future work can investigate how to extract sufficiently informative features from a minimal number of sensors. For example, extracting joint angles could improve the prediction error (McLean et al., 2003), but this requires at least two sensors. Second, while we always attached the phone at a fixed position, the phone's orientation could be slightly different due to variations across experiments with different subjects. This means that our learned models are evaluated on data that may have been collected using a slightly different reference frame for the sensor measurements. Hence, our model should be robust against minor perturbations of the phone's orientation, but not against attaching the phone at different locations. Future work should develop models that are robust against variations in the position of the sensors as well. This can be done by collecting data with sensors at different locations and using machine learning to train a model that works for various locations. Third, we decided to only use linear models, since non-linear models did not improve the results for this small dataset of 10 patients. Training non-linear models using data from more patients can potentially detect non-linear relations between the sensor data and the joint impulse. Moreover, additional data can improve the model's accuracy by learning from a larger number of training examples.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee Research UZ/KU Leuven. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AD, JE, AT, IJ, BV, and JD conceived, designed and coordinated the study. JE collected the data. AD and JD developed the machine learning pipeline. AD and JE analyzed the data and initially drafted the manuscript. AT, IJ, BV, and JD provided useful suggestions in the preparation of the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00320/full#supplementary-material

REFERENCES

Alvarez, D., González, R. C., López, A., and Alvarez, J. C. (2008). 'Comparison of step length estimators from wearable accelerometer devices," in *Encyclopedia of Healthcare Information Systems*, eds N. Wickramasinghe, and E. Geisler (Hershey, PA: IGI Global), 244–250. doi: 10.4018/978-1-59904-889-5.ch034

Bussmann, J., Martens, W., Tulen, J., Schasfoort, F., Van Den Berg-Emons, H., and Stam, H. (2001). Measuring daily behavior using ambulatory accelerometry: the activity monitor. *Behav. Res. Methods Instrum. Comput.* 33, 349–356. doi: 10.3758/BF03195388

Camomilla, V., Bergamini, E., Fantozzi, S., and Vannozzi, G. (2018). Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review. *Sensors* 18:873. doi: 10.3390/s18030873

De Brabandere, A., De Beeck, T. O., Schütte, K. H., Meert, W., Vanwanseele, B., and Davis, J. (2018). Data fusion of body-worn accelerometers and heart rate to predict vo2max during submaximal running. *PLoS ONE* 13:e0199509. doi: 10.1371/journal.pone.0199509

De Brabandere, A., Robberechts, P., Op De Beéck, T., and Davis, J. (2019). "Automating feature construction for multi-view time series data," in *Automating Data Science Workshop @ ECML PKDD 2019* (Würzburg).

De Groote, F., De Laet, T., Jonkers, I., and De Schutter, J. (2008). Kalman smoothing improves the estimation of joint kinematics and kinetics in marker-based human gait analysis. *J. Biomech.* 41, 3390–3398. doi: 10.1016/j.jbiomech.2008.09.035

de Vries, W., Veeger, H., Baten, C., and van der Helm, F. (2012). Determining a long term ambulatory load profile of the shoulder joint: Neural networks

predicting input for a musculoskeletal model. *Hum. Movement Sci.* 31, 419–428. doi: 10.1016/j.humov.2011.07.010

Delp, S. L., Anderson, F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., et al. (2007). Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE Trans. Biomed. Eng.* 54, 1940–1950. doi: 10.1109/TBME.2007.901024

Delp, S. L., Loan, J. P., Hoy, M. G., Zajac, F. E., Topp, E. L., and Rosen, J. M. (1990). An interactive graphics-based model of the lower extremity to study orthopaedic surgical procedures. *IEEE Trans. Biomed. Eng.* 37, 757–767. doi: 10.1109/10.102791

Felson, D. T. (2013). Osteoarthritis as a disease of mechanics. *Osteoarthrit. Cartilage* 21, 10–15. doi: 10.1016/j.joca.2012.09.012

Fregly, B. J., Besier, T. F., Lloyd, D. G., Delp, S. L., Banks, S. A., Pandy, M. G., et al. (2012). Grand challenge competition to predict *in vivo* knee loads. *J. Orthopaedic Res.* 30, 503–513. doi: 10.1002/jor.22023

Fulcher, B. D. (2018). "Feature-based time-series analysis," in *Feature Engineering for Machine Learning and Data Analytics*, eds G. Dong, and H. Liu (Boca Raton, FL: CRC Press), 87–116. doi: 10.1201/9781315181080-4

Guo, Y., Storm, F., Zhao, Y., Billings, S., Pavic, A., Mazzá, C., et al. (2017). A new proxy measurement algorithm with application to the estimation of vertical ground reaction forces using wearable sensors. *Sensors* 17:2181. doi: 10.3390/s17102181

Karatsidis, A., Bellusci, G., Schepers, H., de Zee, M., Andersen, M., and Veltink, P. (2017). Estimation of ground reaction forces and moments during gait using only inertial motion capture. *Sensors* 17:75. doi: 10.3390/s170 10075

Kavanagh, J. J. and Menz, H. B. (2008). Accelerometry: a technique for quantifying movement patterns during walking. *Gait Post.* 28, 1–15. doi: 10.1016/j.gaitpost.2007.10.010

McLean, S. G., Su, A., and van den Bogert, A. J. (2003). Development and validation of a 3-D model to predict knee joint loading during dynamic movement. *J. Biomech. Eng.* 125, 864–874. doi: 10.1115/1.1634282

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B* 72, 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Meireles, S., De Groote, F., Van Rossom, S., Verschueren, S., and Jonkers, I. (2017). Differences in knee adduction moment between healthy subjects and patients with osteoarthritis depend on the knee axis definition. *Gait Post.* 53, 104–109. doi: 10.1016/j.gaitpost.2017.01.013

Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., and Davis, J. (2018). "Fatigue prediction in outdoor runners via machine learning and sensor fusion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London: ACM), 606–615. doi: 10.1145/3219819.3219864

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/v12/pedregosa11a.html

Robbins, S. M., Birmingham, T. B., Jones, G. R., Callaghan, J. P., and Maly, M. R. (2009). Developing an estimate of daily cumulative loading for the knee: examining test-retest reliability. *Gait Post.* 30, 497–501. doi: 10.1016/j.gaitpost.2009.07.118

Steele, K. M., DeMers, M. S., Schwartz, M. H., and Delp, S. L. (2012). Compressive tibiofemoral force during crouch gait. *Gait Post.* 35, 556–560. doi: 10.1016/j.gaitpost.2011.11.023

Stetter, B. J., Ringhof, S., Krafft, F. C., Sell, S., and Stein, T. (2019). Estimation of knee joint forces in sport movements using wearable sensors and machine learning. *Sensors* 19:3690. doi: 10.3390/s19173690

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Wesseling, M., De Groote, F., Meyer, C., Corten, S., Simon, J.-P., Desloovere, K., and Jonkers, I. (2016). Subject-specific musculoskeletal modelling in patients before and after total hip arthroplasty. *Comput. Methods Biomech. Biomed. Eng.* 19, 1683–1691. doi: 10.1080/10255842.2016.1181174

Wesseling, M., van der Straaten, R., Timmermans, A., and Jonkers, I. (2018). "Estimating joint loading using inertial measurement units and ground reaction forces," in *European Orthopaedic Research Society* (Galway).

Weyand, P. G., Kelly, M., Blackadar, T., Darley, J. C., Oliver, S. R., Ohlenbusch, N. E., et al. (2001). Ambulatory estimates of maximal aerobic power from foot-ground contact times and heart rates in running humans. *J. Appl. Physiol.* 91, 451–458. doi: 10.1152/jappl.2001.91.1.451

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., Van Beijnum, B.-J. F., et al. (2018). Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors. *Front. Physiol.* 9:218. doi: 10.3389/fphys.2018.00218

Zargham, A., Afschrift, M., De Schutter, J., Jonkers, I., and De Groote, F. (2019). Inverse dynamic estimates of muscle recruitment and joint contact forces are more realistic when minimizing muscle activity rather than metabolic energy or contact forces. *Gait Post.* 74, 223–230. doi: 10.1016/j.gaitpost.2019.08.019

Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., and Yang, J. (2011). A framework for hand gesture recognition based on accelerometer and EMG sensors. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum.* 41, 1064–1076. doi: 10.1109/TSMCA.2011.2116004

Check for
updates

Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning

Johannes Burdack[1], Fabian Horst[1*], Sven Giesselbach[2,3], Ibrahim Hassan[1,4], Sabrina Daffner[5] and Wolfgang I. Schöllhorn[1,6]

[1] Department of Training and Movement Science, Institute of Sport Science, Johannes Gutenberg-University, Mainz, Germany, [2] Knowledge Discovery, Fraunhofer-Institute of Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany, [3] Competence Center Machine Learning Rhine-Ruhr (ML2R), Dortmund, Germany, [4] Faculty of Physical Education, Zagazig University, Zagazig, Egypt, [5] Qimoto, Doctors' Surgery for Sport Medicine and Orthopedics, Wiesbaden, Germany, [6] Department of Wushu, School of Martial Arts, Shanghai University of Sport, Shanghai, China

Human movements are characterized by highly non-linear and multi-dimensional interactions within the motor system. Therefore, the future of human movement analysis requires procedures that enhance the classification of movement patterns into relevant groups and support practitioners in their decisions. In this regard, the use of data-driven techniques seems to be particularly suitable to generate classification models. Recently, an increasing emphasis on machine-learning applications has led to a significant contribution, e.g., in increasing the classification performance. In order to ensure the generalizability of the machine-learning models, different data preprocessing steps are usually carried out to process the measured raw data before the classifications. In the past, various methods have been used for each of these preprocessing steps. However, there are hardly any standard procedures or rather systematic comparisons of these different methods and their impact on the classification performance. Therefore, the aim of this analysis is to compare different combinations of commonly applied data preprocessing steps and test their effects on the classification performance of gait patterns. A publicly available dataset on intra-individual changes of gait patterns was used for this analysis. Forty-two healthy participants performed 6 sessions of 15 gait trials for 1 day. For each trial, two force plates recorded the three-dimensional ground reaction forces (GRFs). The data was preprocessed with the following steps: GRF filtering, time derivative, time normalization, data reduction, weight normalization and data scaling. Subsequently, combinations of all methods from each preprocessing step were analyzed by comparing their prediction performance in a six-session classification using Support Vector Machines, Random Forest Classifiers, Multi-Layer Perceptrons, and Convolutional Neural Networks. The results indicate that filtering GRF data and a supervised data reduction (e.g., using Principal Components Analysis) lead to increased prediction performance of the machine-learning classifiers. Interestingly, the weight normalization and the number of data points (above a certain minimum) in the time normalization does

not have a substantial effect. In conclusion, the present results provide first domain-specific recommendations for commonly applied data preprocessing methods and might help to build more comparable and more robust classification models based on machine learning that are suitable for a practical application.

INTRODUCTION

Human movements are characterized by highly non-linear and multi-dimensional interactions within the motor system (Chau, 2001a; Wolf et al., 2006). In this regard, the use of data-driven techniques seems to be particularly suitable to generate predictive and classification models. In recent years, different approaches based on machine-learning techniques such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) or Random Forest Classifiers (RFCs) have been suggested in order to support the decision making of practitioners in the field of human movement analysis, e.g., in classifying movement patterns into relevant groups (Schöllhorn, 2004; Figueiredo et al., 2018). Most machine-learning applications in human movements are found in human gait using biomechanical data (Schöllhorn, 2004; Ferber et al., 2016; Figueiredo et al., 2018; Halilaj et al., 2018; Phinyomark et al., 2018). Although it is generally striking that there are more and more promising applications of machine learning in the field of human movement analysis, the applications are very diverse and differ in their objectives, samples and classification tasks. In order to fulfill the application requirements and to ensure the generalizability of the results, a number of stages are usually carried out to process the raw data in classifications using machine learning. Typically, machine-learning classifications of gait patterns consist of a preprocessing and a classification stage (Figueiredo et al., 2018). The preprocessing stage can be distinguished in feature extraction, feature normalization, and feature selection. The classification stage includes cross validation, model building and validation, as well as evaluation. Different methods have been used for each stage and there is no clear consensus on how to proceed in each of these stages. This is particularly the case for the preprocessing stages of the measured raw data before the classification stage, where there are hardly any recommendations, standard procedures or systematic comparisons of different steps within the preprocessing stage and their impact on the classification accuracy (Slijepcevic et al., 2020). The following six steps, for example, can be derived from the preprocessing stage: (1) Ground reaction force (GRF) filtering, (2) time derivative, (3) time normalization, (4) data reduction, (5) weight normalization, and (6) data scaling.

(1) There are a number of possible noise sources in the recording of biomechanical data. Noise can be reduced by careful experimental procedures, however, cannot be completely removed (Challis, 1999). So far there is less known about optimal filter-cut-off frequencies in biomechanical gait analysis (Schreven et al., 2015). Apart from a limited certainty about an optimal range of filter cut-off frequencies of the individual GRF components, the effect of GRF filtering on the prediction performance of machine-learning classifications has not been reported.

(2) In the majority of cases, time-continuous waveforms or time-discrete gait variables are measured and used for the classification (Schöllhorn, 2004; Figueiredo et al., 2018). Although, some authors also used time derivatives or data in the frequency or frequency-time domain from time-continuous waveforms (Schöllhorn, 2004; Figueiredo et al., 2018). A transformation, which has barely been applied so far, is the first-time derivative of the acceleration, also known as jerk (ΔtGRF) (Flash and Hogan, 1985). However, ΔtGRF might describe human gait more precisely than velocity and acceleration, especially when the GRF is measured. ΔtGRF can be determined directly by calculating the first-time derivative of the GRF measured by force plates.

(3) Feature normalization has been applied in order to achieve more robust classification models (Figueiredo et al., 2018). A normalization in time is commonly applied to normalize the biomechanical waveforms as percentage of the step, stride or stance phase (Kaczmarczyk et al., 2009; Alaqtash et al., 2011a,b; Eskofier et al., 2013; Zhang et al., 2014). It is differentiated among other things between 101 points in time (Eskofier et al., 2013), 1000 points in time (Slijepcevic et al., 2017) or the percentage occurrence per step cycle (Su and Wu, 2000).

(4) The purpose of data reduction is to reduce the amount of data to the most relevant features. A dimensionally reduction is often performed in order to determine which data is to be retained and which can be discarded. The use of dimension reduction can speed up computing time or reduce storage costs for data analysis. However, it should be noted that these feature selection approaches can not only reduce computation costs, but could also improve the classification accuracy (Phinyomark et al., 2018). Beside the unsupervised selection of single time-discrete gait variables (Schöllhorn, 2004; Begg and Kamruzzaman, 2005), typical methods for reducing the dimensionality of the data is, for example, the Principal Component Analysis (Deluzio and Astephen, 2007; Lee et al., 2009; Eskofier et al., 2013; Badesa et al., 2014).

(5) Another way of feature normalization is weight or height normalization. Weight and height normalizations in amplitude are a frequently used method to control for inter-individual differences in kinetic and kinematic variables (Wannop et al., 2012). To what extent the multiplication by a constant factor influences the

classification has not yet been investigated to the best of our knowledge.

(6) A third way of feature normalization is data scaling. Data scaling is often performed to normalize the amplitude of one or different variable time courses (Mao et al., 2008; Laroche et al., 2014). The z-score method is mainly used (Begg and Kamruzzaman, 2005; Begg et al., 2005). In machine learning, scaling to a variable or variable waveform the interval [0, 1] or [-1, 1] is common in order to minimize amplitude-related weightings when training the classifiers (Hsu et al., 2003). To the best of our knowledge, it has not yet been investigated whether it makes a difference to scale over a single gait trial or over all trials of one subject in one session.

In summary, there is a lack of domain-specific standard procedures and recommendations, especially for the various data preprocessing steps commonly applied before machine-learning classifications. Therefore, the aim of this analysis is to compare different commonly applied data preprocessing steps and examine their effect on the classification performance using different machine-learning classifiers (ANN, SVM, RFC). A systematic comparison is of particular interest for deriving domain-specific recommendations, finding best practice models and the optimization of machine-learning classifications of human gait data. The analysis is based on the classification problem described by Horst et al. (2017), who investigated intra-individual gait patterns across different time-scales over 1 day.

MATERIALS AND METHODS

Sample and Experimental Protocol

The publicly available dataset on intra-individual changes of gait patterns by Horst et al. (2017, 2019a) and two unpublished datasets (Daffner, 2018; Hassan, 2019) following the same experimental protocol were used for this analysis. In total, the joint dataset consisted of 42 physically active participants (22 females, 20 males; 25.6 ± 6.1 years; 1.72 ± 0.09 m; 66.9 ± 10.7 kg) without gait pathology and free of lower extremity injuries. The study was conducted in accordance with the Declaration of Helsinki and all participants were informed of the experimental protocol and provided their written consent. The approval of the ethics committee of the Rhineland-Palatinate Medical Association in Mainz has been received.

As presented in **Figure 1**, the participants performed 6 sessions (S1–S6) of 15 gait trials in each session, while there was no intervention between the sessions. After the first, third and fifth session, the participants had a break of 10 min until the beginning of the subsequent session. Between S2 and S3 was a break of 30 min and between S4 and S5 the break was 90 min. The participants were instructed to walk a 10 m-long path at a self-selected speed barefooted. For each trial, three-dimensional GRFs were recorded by means of two Kistler force plates of type 9287CA (Kistler, Switzerland) at a frequency of 1,000 Hz. The Qualisys Track Manager 2.7 software (Qualisys AB, Sweden) managed the recording. During the investigation, the laboratory environment was kept constant and each subject was analyzed by the same assessor only. Before the first session, each participant carried out 20 familiarization trials to get used to the experimental setup and to determine a starting point for a walk across the force plates. Before each of the following sessions, five familiarization trials were carried out to take into account an effect of practice and to control the individual starting position. In addition, the participants were instructed to look toward a neutral symbol (smiley) on the opposite wall of the laboratory to direct their attention away from targeting the force plates and ensure a natural gait with upright posture. The description of the experimental procedure can be found as well in the original study (Horst et al., 2017).

Data Preprocessing

The stance phase of the right and left foot was determined using a vertical GRF threshold of 20 Newton. Different combinations of commonly used data preprocessing steps, which typically precede machine-learning classifications of biomechanical gait patterns have been compared (**Figure 2**). Within the introduced stage of preprocessing, the following six data preprocessing steps were investigated: (1. GRF filtering) comparing filtered and unfiltered GRF data. The method described by Challis (1999) was used to determine the optimal cut-off frequencies (f_c) for the respective gait trials. The optimal filter frequencies were calculated for each foot and each of the three dimensions in each gait trial separately. (2. Time derivative) comparing the recorded GRF and ΔtGRF, the first-time derivative of the GRF. ΔtGRF was calculated by temporally derivating the GRF for each time interval. (3. Time normalization) comparing
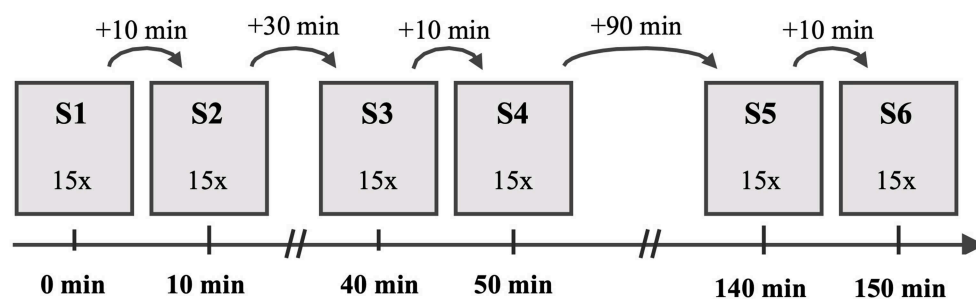
FIGURE 1 | Experimental procedure with the chronological order of the six sessions (S1–S6) and the duration of the rest periods between subsequent sessions.

FIGURE 2 | Combinations of commonly used data preprocessing steps before machine-learning classifications. (1) Data points per foot and dimension. (2) Time-continuous waveforms without reduction (TC), time-discrete gait variables by an unsupervised reduction (TD), and principle components by a supervised reduction using Principal Component Analysis (PCA). (3) Z-transformation combined with scaling from [−1, 1] over single trials (ST) or all trials (AT). f_c: individual optimal filter cut-off frequency. (4) ΔtGRF: first-time derivative of GRF.

the number of time points for the time normalization to the stance phase. Each variable was time normalized to 11, 101 and 1,001 data points, respectively. (4. Data reduction) comparing non-reduced, time-continuous waveforms (TC), time-discrete gait variables (TD) and principle components by a reduction using Principal Component Analysis (PCA) applied to the time-continuous waveforms. The PCA (Hotelling, 1933) is a statistical procedure that uses an orthogonal transformation from a set of observations of potentially correlated variables into a set of values of linearly uncorrelated variables, the so called "principal components." In this transformation, the first principal component explains the largest possible part of the variance. Each subsequent principal component again explains the largest part of the remaining variance, with the restriction that subsequent principal components are orthogonal to the preceding principal components. In our experiment, the resulting features, i.e., the principle components explaining 98% of the total variance, were used as input feature vectors for the classification. The time-discrete gait variables of the fore-aft and medio-lateral shear force were the minimum and the the maximum values as well as their occurrence during the stance phase, and of the vertical force the minimum and the two local maxima values as well as their occurrence during the stance phase. This resulted in 28 time-discrete gait variables for GRF data and 24 time-discrete gait variables for ΔtGRF data. (5. Weight normalization) comparing whether weight normalization to the body weight of every session was performed or not. The normalization to the body weight before every season would exclude the impact of any changes in the body mass during the investigation. (6. Data scaling) comparing different data scaling techniques. Scaling is a common procedure for data processing prior to classifications of gait data (Chau, 2001a,b). It was carried out to ensure an equal contribution of all variabilities to the prediction performance and to avoid dominance of

variables with greater numeric range (Hsu et al., 2003). On the one hand, this involved a z-transformation over all trials and one over each single trial combined with a scaling to the range of [−1, 1] (Hsu et al., 2003), determined over all trials or over each single trial. The combination of these amplitude normalization methods result in four different scaling methods.

The data preprocessing was managed within Matlab R2017b (MathWorks, USA) and all combinations of each methods of each data preprocessing and classification step were performed in the current analysis in the order described in **Figure 2**. In total, the analysis included 1,152 different combinations of data preprocessing and classification step methods (1,152 = 2 GRF filtering * 2 Time derivative * 3 Time normalization * 3 Data reduction * 2 Weight normalization * 4 Data scaling * 4 Classifier). In the two methods TD and PCA for data reduction, the data scaling could not be applied for all methods. In many cases, all values of a time-discrete gait variable or a principle component were identical [**Figure 2**: Data Scaling z: ST or [−1, 1]: ST] and thus no variance occurred, which is necessary for the calculation of the data scaling. Only, the data scaling over all trials from one subject [**Figure 2**: Data scaling: z: AT, [−1, 1]: AT] could be performed for all three methods of data reduction. In order to keep the number of considered combinations the same for all methods of a data preprocessing step, only the data scaling of all attempts of one subject [**Figure 2**: Data scaling: z: AT, [−1, 1]: AT] was considered for the descriptive and statistical analysis in the results section. This scaling also led to by far the best performance scores. Consequently, 288 different combinations of data preprocessing and classification step methods (288 = 2 GRF filtering * 2 Time derivative * 3 Time normalization * 3 Data reduction * 2 Weight normalization * 1 Data scaling * 4 Classifier) were compared quantitatively with each other on basis of the performance scores.

TABLE 1 | Length of the resulting input feature vectors depending on different combinations of preprocessing methods.

Data reduction	Time normalization	Time derivative	GRF filtering	Length of input feature vector
TC	11	GRF; ΔtGRF	No; Yes	66 = 11 * 3 * 2
	101	GRF; ΔtGRF	No; Yes	606 = 101 * 3 * 2
	1,001	GRF; ΔtGRF	No; Yes	6006 = 1001 * 3 * 2
TD	11; 101; 1,001	GRF	No; Yes	28 = 7 * 2 * 2
		ΔtGRF	No; Yes	24 = 6 * 2 * 2
PCA	11	GRF	No	46 (44, 47)
			Yes	47 (44, 48)
		ΔtGRF	No	53 (49, 55)
			Yes	38 (43, 46)
	101	GRF	No	78 (73, 83)
			Yes	72.5 (69, 79)
		ΔtGRF	No	239 (210, 268)
			Yes	108 (97, 119)
	1,001	GRF	No	79 (73, 84)
			Yes	72 (68, 79)
		ΔtGRF	No	369 (341, 386)
			Yes	108 (97, 120)

*TC, time-continuous waveforms for three dimensions (*3) and two steps (*2); TD, time-discrete gait variables of minima and maxima of the three dimensions (GRF: 7; ΔtGRF: 6) for two steps (*2) and their relative occurrences (*2); PCA, Median and interquartile distance of the number of principle components.*

Data Classification

The intra-individual classification of gait patterns was based on the 90 gait trials (90 = 6 sessions × 15 trials) of each participant. For each trial, a concatenated vector of the three-dimensional variables of both force plates was used for the classification. Due to the different time normalization and data reduction methods, the resulting length of the input feature vectors differed (**Table 1**).

The classification based on the following four supervised machine-learning classifiers with an exhaustive hyper-parameter search: (1) Support Vector Machines (SVMs) (Boser et al., 1992; Cortes and Vapnik, 1995; Müller et al., 2001; Schölkopf and Smola, 2002) using a linear kernel and a grid search to determine the best cost parameter $C = 2^{-5}, 2^{-4.75}, \ldots, 2^{15}$. (2) Random Forest Classifiers (RFCs) (Breiman, 2001) with the Gini coefficient as decision criterion. Different numbers of trees (n_estimators = 200, 225, …, 350) and maximal tree depth (n_depth = 4, 5, …, 8) were determined empirically via grid search. (3) Multi-Layer Perceptrons (MLPs) (Bishop, 1995) with one hidden layer of size 2^6 (= 64 neurons) and 2,000 iterations with the weight optimization algorithm Adam ($\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 10^{-8}$). The learning rate regularization parameter α (= $10^{-1}, 10^{-2}, \ldots, 10^{-7}$) was determined via grid search in the cross-validation. (4) Convolutional Neural Networks (CNNs) (LeCun et al., 2015) consisting of three convolutional layers and one fully connected layer. The first convolutional layers contained 24 filters with a kernel size of 8, a stride of 2 and a padding of 4. The second contained 32 filters with a kernel size of 8, a stride of 2 and a padding of 4. The third convolutional layer contained 48 filters with a kernel size of 6, a stride of 3 and a padding of 3. After each convolutional layer a ReLU activation was performed and after a fully connected layer a SoftMax was used to obtain probability of each of the classes. This architecture follows CNNs previously used for the classification of GRF data (Horst et al., 2019b). The ability to distinguish gait patterns of one test session from gait patterns of other test sessions was investigated in a multi-class classification (six-session classification) setting. For the evaluation of the prediction performance, the F1-, precision- and recall-scores were calculated over a stratified 15-fold cross validation configuration. 78 of 90 parts of the data were used for training, 6 of 90 parts were used as a validation set and the remaining 6 of 90 parts was reserved for testing. The 6 samples per test split were evenly distributed across all session partitions and are excluded from the complete training and validation process. Only 6 samples were selected for the test split because we wanted to guarantee as much training data as possible. In order to get meaningful results, the Training Validation Test splitting was stratified repeated 15 times so that each of the 90 gait trials was exactly once in the test set. The classification was performed within Python 3.6.3 (Python Software Foundation, USA) using the scikit-learn toolbox (0.19.2) (Pedregosa et al., 2011) and PyTorch (1.2.0) (Paszke et al., 2019).

The evaluation was carried out by calculating the performance indicators (accuracy, F1-score, precision and recall) defined by the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Please note that since this is a balanced data set for multi-class classification, the accuracy corresponds exactly to the recall.

Statistical Analysis

For the comparison of the different combinations of the described preprocessing steps, the mean performance scores were compared statistically. Each mean value combined all combinations of preprocessing steps where the preprocessing method was part of. The Shapiro-Wilk test showed that none of the examined variables violated the normal distribution assumption ($p \geq 0.109$). For the comparison of all combinations of the preprocessing methods, paired-samples t-test and repeated-measures ANOVAs were calculated for the variables of time derivative, GRF filtering and weight normalization. For the ANOVAs *post hoc* Bonferroni corrected paired-samples t-tests were calculated for the variables of time normalization, data reduction and classifier. Furthermore, the effect sizes d and η_p^2 were calculated; d and η_p^2 are considered a small effect for $|d| = 0.2$ and $\eta_p^2 < 0.06$, a medium effect for $|d| = 0.5$ and

TABLE 2 | Mean F1-score for each individual participant depending on each preprocessing method and machine-learning classifier.

	GRF filtering		Time derivative		Time normalization			Data reduction			Weight normalization		Machine-learning classifier			
	No	Yes	GRF	ΔtGRF	11	101	1001	TC	TD	PCA	No	Yes	SVM	RFC	MLP	CNN
S01	38.3	42.0	41.4	39.0	34.9	42.4	43.2	40.2	25.7	54.5	39.9	40.4	46.4	45.0	36.3	32.9
S02	24.5	29.4	27.9	26.0	23.2	28.4	29.2	24.5	20.6	35.7	27.2	26.7	30.0	30.7	23.9	23.2
S03	36.9	43.5	40.6	39.8	36.1	44.3	40.2	40.9	27.6	52.1	40.1	40.3	44.8	43.9	38.2	33.9
S04	42.9	50.0	48.2	45.1	41.5	48.8	49.5	48.8	36.5	53.6	45.9	47.3	51.4	56.2	42.0	36.8
S05	49.9	50.2	52.0	48.1	47.1	51.1	52.0	50.2	36.5	63.6	49.6	50.6	56.7	56.2	45.5	41.9
S06	38.4	39.8	39.3	38.9	32.0	42.3	43.2	38.9	28.0	49.8	39.3	38.9	42.8	44.6	38.2	30.8
S07	31.5	40.5	35.4	36.7	30.5	40.0	37.6	34.6	28.7	44.8	36.2	35.9	39.2	41.7	32.8	30.3
S08	42.7	49.0	46.4	45.4	41.4	49.0	47.0	47.3	38.2	51.8	45.7	46.0	49.0	52.1	44.9	37.6
S09	43.2	47.2	46.1	44.3	39.1	49.8	46.7	43.5	34.0	58.1	45.3	45.1	51.2	49.7	41.2	38.7
S10	41.3	40.3	41.2	40.4	34.2	44.1	44.1	44.4	27.5	50.5	40.5	41.1	43.8	43.2	42.2	33.9
S11	38.5	40.7	42.5	36.7	35.3	42.8	40.8	42.0	27.6	49.3	39.5	39.7	44.0	45.1	35.2	34.2
S12	34.1	31.9	36.2	29.8	27.9	35.4	35.7	35.3	22.9	40.9	33.5	32.6	36.7	34.9	34.1	26.3
S13	31.7	34.5	34.4	31.8	28.6	36.9	33.8	32.5	27.8	39.0	32.8	33.4	36.9	36.6	31.2	27.6
S14	33.9	34.0	38.1	29.8	28.3	37.3	36.2	35.7	24.4	41.7	34.1	33.8	36.3	35.4	35.7	28.3
S15	39.9	45.3	46.8	38.4	36.8	46.7	44.2	42.5	31.2	54.0	42.8	42.3	48.7	46.2	39.5	35.8
S16	32.0	32.9	32.9	31.9	27.5	34.5	35.3	33.8	23.3	40.2	32.7	32.2	34.3	34.6	34.6	26.3
S17	29.3	30.0	31.7	27.6	22.6	33.4	32.6	30.0	21.7	36.9	29.7	29.7	33.0	31.5	29.4	24.8
S18	24.4	26.9	25.6	25.8	22.8	27.0	27.3	28.3	17.9	30.9	25.7	25.7	27.6	26.7	27.5	21.0
S19	27.3	28.7	31.5	24.5	25.0	29.4	29.5	26.8	22.7	34.5	27.9	28.0	31.0	30.6	25.5	24.8
S20	29.3	34.0	32.4	31.0	26.6	33.7	34.7	31.9	25.7	37.4	31.8	31.6	34.4	36.3	30.0	26.2
S21	27.7	29.6	30.9	26.4	25.4	30.8	29.7	28.1	22.4	35.2	28.6	28.8	31.2	33.1	26.4	24.0
S22	32.3	33.6	36.4	29.5	28.6	34.5	35.7	34.1	24.2	40.5	33.2	32.7	33.6	35.3	35.6	27.2
S23	31.7	35.0	34.6	32.1	28.5	35.0	36.5	33.8	25.6	40.7	33.1	33.6	34.9	39.0	32.8	26.6
S24	35.4	43.3	40.1	38.6	33.9	41.3	42.9	39.7	32.1	46.3	39.7	39.1	42.8	43.8	39.2	31.6
S25	34.7	41.9	39.3	37.4	34.7	41.6	38.7	37.1	33.0	44.8	38.3	38.4	40.2	43.8	36.8	32.5
S26	47.6	49.9	53.6	43.8	42.3	51.3	52.5	52.2	41.7	52.1	48.5	48.9	52.9	56.3	47.0	38.6
S27	31.5	31.8	30.4	32.8	26.5	34.2	34.2	33.6	24.6	36.7	31.6	31.6	32.8	35.2	31.5	27.0
S28	35.9	45.5	41.9	39.4	33.5	45.1	43.5	43.0	29.9	49.2	40.9	40.5	43.5	43.9	42.3	33.0
S29	32.2	36.1	33.1	35.2	30.1	36.5	35.8	36.3	22.6	43.6	34.5	33.9	36.8	35.7	36.0	28.1
S30	31.1	33.1	35.4	28.9	28.3	32.6	35.4	35.1	21.6	39.0	31.9	32.3	33.4	37.1	32.5	25.4
S31	51.3	53.7	54.5	50.5	44.5	56.0	57.0	58.6	36.6	62.2	52.4	52.5	56.8	58.3	53.5	41.3
S32	43.0	45.9	47.4	41.5	38.6	46.0	48.7	49.5	31.1	52.7	44.7	44.2	47.9	50.3	44.0	35.5
S33	35.7	41.4	39.7	37.4	32.1	40.4	43.1	41.6	23.5	50.3	38.2	38.9	42.7	41.6	39.2	30.7
S34	49.8	51.8	53.8	47.8	44.5	53.2	54.7	52.1	39.0	61.4	50.7	50.9	54.1	57.5	51.3	40.4
S35	38.4	45.4	45.3	38.8	35.1	45.5	45.5	45.6	25.7	53.7	42.2	41.8	45.8	47.7	42.2	32.2
S36	36.9	39.3	41.0	35.3	32.9	40.7	40.7	39.5	29.8	45.1	37.9	38.3	41.2	43.1	36.8	31.3
S37	30.9	33.7	35.8	28.9	27.7	33.5	35.8	33.8	20.3	42.9	32.2	32.5	35.3	33.3	34.3	26.4
S38	35.1	38.2	39.0	34.3	30.9	38.7	40.3	37.8	26.5	45.6	36.7	36.6	39.5	40.9	37.1	29.1
S39	41.6	43.2	46.1	38.7	39.0	43.1	45.1	47.4	28.1	51.7	42.4	42.4	44.3	48.6	42.8	33.9
S40	41.4	48.9	48.8	41.5	37.1	47.9	50.4	48.1	30.3	56.9	45.1	45.2	48.9	50.5	46.8	34.4
S41	38.4	43.2	43.9	37.6	34.7	42.9	44.7	44.4	28.1	49.7	40.6	41.0	42.6	48.3	40.5	31.7
S42	27.2	29.4	31.3	25.4	25.5	28.2	31.4	29.7	21.6	33.0	28.3	28.3	29.2	31.4	28.3	24.3
M	**36.2**	**39.6**	**39.8**	**36.0**	**32.8**	**40.4**	**40.6**	**39.4**	**27.8**	**46.5**	**37.9**	**37.9**	**41.2**	**42.3**	**37.3**	**31.0**
SD	**6.7**	**7.3**	**7.3**	**6.8**	**6.3**	**7.3**	**7.3**	**7.8**	**5.6**	**8.3**	**6.8**	**7.0**	**7.8**	**8.3**	**6.8**	**5.3**

*The mean precision and mean recall (= accuracy) scores for each individual participant depending on each preprocessing method and machine-learning classifier can be found in **Supplementary Tables S1, S2**.*
Each mean value combines all combinations of preprocessing steps where the preprocessing method was part of (n = 42).

$0.06 < \eta_p^2 < 0.14$ and a large effect for $|d| = 0.8$ and $\eta_p^2 > 0.14$ (Cohen, 1988). The p-value at which research is considered worth to be continued (Fisher, 1922) has been set to $p = 0.05$. To determine a best practice model, all combinations of data preprocessing methods were ranked according to their mean performance scores over 15-fold cross validation and the rank sum was calculated.

RESULTS

Average Performance of Different Data Preprocessing Methods

The analysis compares 288 different combinations of data preprocessing methods based on the resulting F1-score. **Table 2** displays the mean F1-score for each individual participant over the 15-fold cross validation (**Supplementary Tables S1**, **S2** show the mean precision and recall values).

Figure 3 shows the mean F1-scores over all participants. It is noticeable that the highest mean F1-scores were achieved using PCA, while the normalization to 101 and 1,001 data points or the weighting has only a minor effect on the F1-score. The time normalization to only 11 data points and the reduction to time-discrete gait variables gave particularly low mean classification scores. Concerning the machine-learning classifiers, the RFCs achieved the highest mean F1-scores followed by the SVMs, MLPs, and CNNs.

GRF Filtering

A paired-samples t-test was performed to determine if there were differences in F1-score in unfiltered GRF data compared to f_c– filtered GRF data across all participants. The mean F1-score of the filtered GRF data ($M = 39.6\%$, $SD = 7.3\%$) was significantly higher than that of the unfiltered GRF data ($M = 36.2\%$, $SD = 6.7\%$). The effect size, however, was small $[t_{(41)} = 8.200, p < 0.001, |d| = 0.492]$.

Time Derivative

A paired-samples t-test was conducted to compare the F1-score of GRF and ΔtGRF across all participants. The mean F1-score of GRF ($M = 39.8\%$, $SD = 7.3\%$) was significantly higher than that of ΔtGRF ($M = 36.0\%$, $SD = 6.8\%$) and showed a medium effect size $[t_{(41)} = 8.162, p < 0.001, |d| = 0.540]$.

Time Normalization

A repeated-measures ANOVA determined that there is a significant global effect with large effect size of F1-score between time normalization to 11, 101 and 1,001 data points $[F_{(2.000, 82.000)} = 367.115, p < 0.001, \eta_p^2 = 0.900]$. *Post hoc* paired-samples t-test with Bonferroni correction revealed that there is no significant difference $[t_{(41)} = -0.741, p = 0.463, |d| = 0.031]$ between a time normalization to 101 ($M = 40.4\%$, $SD = 7.3\%$) data points and 1,001 data points ($M = 40.6\%$, $SD = 7.3\%$). However, the time normalization to 101 data points performed significantly better $[t_{(41)} = 22.397, p < 0.001, |d| = 1.118]$ than time normalized to 11 data points ($M = 32.8\%$, $SD = 6.3\%$). Also the time normalization to 1,001 data points performed significantly better than to 11 data

points $[t_{(41)} = 21.789\ p < 0.001, |d| = 1.150]$. Both effect sizes are considered as large.

Data Reduction

A one-way repeated-measures ANOVA was conducted to compare the F1-scores of PCA ($M = 54.9\%$, $SD = 8.5\%$), TC ($M = 50.9\%$, $SD = 8.8\%$), and TD ($M = 37.5\%$, $SD = 6.5\%$). The Huynh-Feldt corrected results showed a highly significant main effect with a large effect size $[F_{(1.594, 65.365)} = 378.372, p < 0.001, \eta_p^2 = 0.902]$. Bonferroni corrected *post hoc* paired-samples t-tests showed that PCA performed significantly better than TC $[t_{(41)} = 14.540, p < 0.001, |d| = 0.884]$ and TD $[t_{(41)} = 22.658, p < 0.001, |d| = 2.635]$. The effect size for both comparisons is considered as large. Furthermore, TC performed also significantly better than TD with a large effect size $[t_{(41)} = 16.516, p < 0.001, |d| = 1.701]$.

Weight Normalization

A paired-samples t-test was conducted to compare the F1-scores of weight-normalized and non-weight-normalized data across all participants. There was no significant difference $[t_{(41)} = -0.644, p = 0.523, |d| = 0.006]$ in the F1-scores for non-weight-normalized data ($M = 37.9\%$, $SD = 6.8\%$) and weight-normalized data ($M = 37.9\%$, $SD = 7.0\%$).

Machine-Learning Classifier

A repeated-measures ANOVA with Huynh-Feldt correction showed a highly significantly global effect with large effect size $[F_{(1.130, 103.478)} = 240.138, p < 0.001, \eta_p^2 = 0.854]$ between the predicted F1-scores by the SVMs ($M = 41.2\%$, $SD = 7.8\%$), RFCs ($M = 42.3\%$, $SD = 8.3\%$), MLPs ($M = 37.3\%$, $SD = 6.8\%$), and CNNs ($M = 31.0\%$, $SD = 5.3\%$). *Post hoc* Bonferroni corrected paired-samples t-test revealed that the RFCs performed significantly better, with a small effect size, than the SVMs $[t_{(41)} = 3.531, p = 0.001, |d| = 0.140]$, with a medium effect size than the MLPs $[t_{(41)} = 9.459, p < 0.001, |d| = 0.664]$ and with a large effect size than the CNNs $[t_{(41)} = 20.780, p < 0.001, |d| = 1.625]$. Also the SVMs performed significantly better than the MLPs with a medium effect $[t_{(41)} = 8.115, p < 0.001, |d| = 0.534]$ and significantly better than the CNNs with a large effect $[t_{(41)} = 23.811, p < 0.001, |d| = 1.530]$. Furthermore, the MLPs performed significantly better than the CNNs with a large effect $[t_{(41)} = 13.725, p < 0.001, |d| = 1.035]$.

Best Practice Combinations of Different Data Preprocessing Methods

In addition to the mean F1-scores for each method of all preprocessing and classification steps, **Table 3** shows the 30 combinations with the highest overall F1-scores, including precision and recall (the complete list including precision and recall can be found in **Supplementary Table S3**). It is particularly noticeable that the first 18 ranks were all achieved using PCA for data reduction. Furthermore, the first eight ranked combinations used GRF data. The first twelve ranked combinations were classified with SVMs, while the highest F1-score was 13th with MLP, 27th with RFC and 57th with CNN.

Table 4 shows the rank scores of all classifications performed for the 288 combinations of the different preprocessing steps
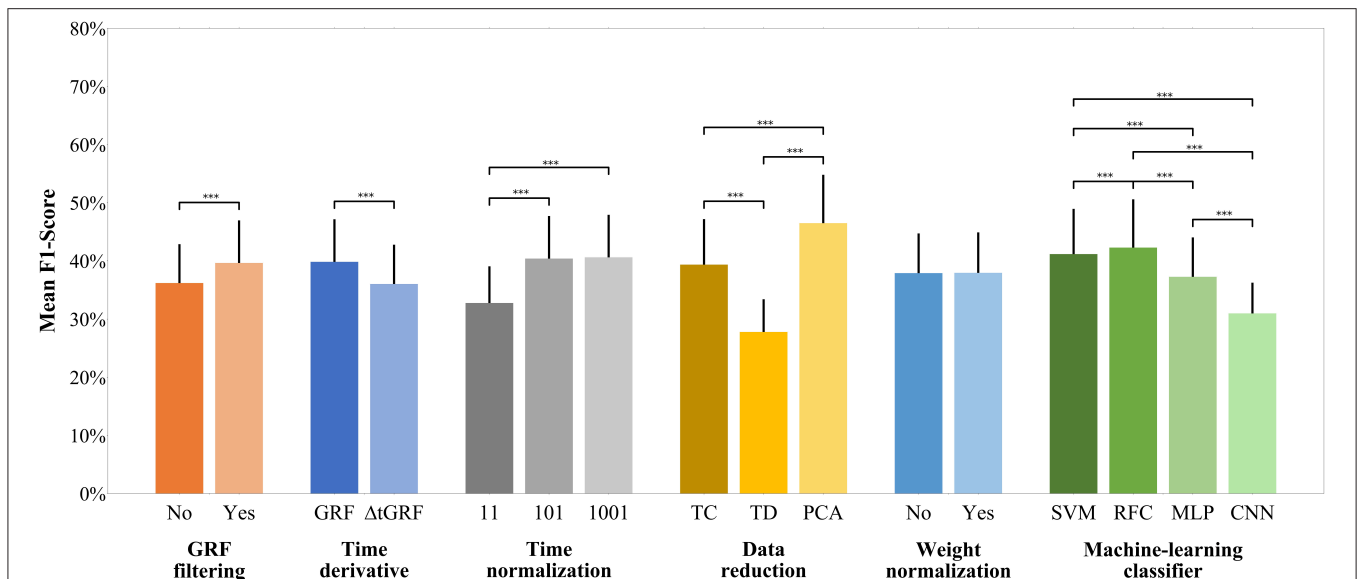
FIGURE 3 | F1-score of each preprocessing step across all participants. The y-axis shows the mean F1-score achieved. The bar charts show the mean value and the standard deviation depending on the respective preprocessing step. The parentheses show a statistically significant effect. Random Baseline = 16.7%; ***$p \leq 0.001$.

according to the F1-scores (**Supplementary Tables S4**, **S5** display the rank score depending on precision and recall). The PCA achieved a particularly high rank score with 87.5% of the maximum achievable rank score. In addition, the GRF with 73.9% and the GRF filtering with 73.4% finished with high rank scores. Again, there are no or only minor differences within the weight normalization and the time normalization to 101 and 1,001 data points. Among the classifiers, the RFCs achieved the highest rank score, just ahead of the SVMs and MLPs and quite far in advance of CNNs.

DISCUSSION

A growing number of promising machine-learning applications could be found in the field of human movement analysis. However, these approaches differ in terms of objectives, samples, and classification tasks. Furthermore, there is a lack of standard procedures and recommendations within the different methodological approaches, especially with respect to data preprocessing steps usually performed prior to machine-learning classification. In this regard, the current analysis comprised a systematic comparison of different preprocessing steps and their effects on the prediction performance of different machine-learning classifiers. The results revealed first domain-specific recommendations for the preprocessing of GRF data prior to machine-learning classifications. This includes, for example, benefits of filtering GRF data and supervised data reduction techniques (e.g., PCA) compared to non-reduced (time-continuous waveforms) or unsupervised data reduction techniques (time-discrete gait variables). On the other hand, the results indicate that the normalization to a constant factor (weight normalization) and the number of data points (above a certain minimum) used during time normalization seem to have

little influence on the prediction performance. Furthermore, the first-time derivative (ΔtGRF) could not achieve advantages over the GRF in terms of prediction performance.

In general, the present results can help to find domain-specific standard procedures for the preprocessing of data that may enable to improve machine-learning classifications in human movement analysis make different approaches better comparable in the future. It should be noted, however, that the results presented are based solely on prediction performance and do not provide information about the effects on the trained models.

GRF Filtering

The present results indicate that the filtered GRF data led to significantly higher mean F1-scores and rank scores than the unfiltered GRF data. The results were especially striking for the classifications of ΔtGRF data. While no clear trend could be derived for the best-ranked combinations of GRF data, most of the best-ranked combinations of ΔtGRF data were filtered. To our knowledge, this analysis was the first that investigated whether a filter (using an optimal filter cut-off frequency) affects the prediction performance of GRF data in human gait (Schreven et al., 2015). The present findings suggest that machine-learning classification should use filtered GRF data. However, it should be noted that the estimation of the optimal filter cut-off frequency using the method described by Challis (1999) is only one out of several possibilities to set a cut-off frequency. Because the individual filter cut-off frequencies were separately calculated for trial and each variable, so it is not yet possible to recommend a generally valid unique cut-off frequency.

Time Derivative

With respect to the feature extraction using the first-time derivative, our analysis revealed that the GRF achieved significantly higher F1-scores compared to the ΔtGRF. In

TABLE 3 | Top 30 combinations of preprocessing methods, ranked by the mean F1-score over the 15-fold cross validation (*n* = 42).

Rank	GRF filtering	Time derivative	Time normalization	Data reduction	Weight normalization	Machine-learning classifier	M	SD
1	No	GRF	1,001	PCA	No	SVM	54.4	9.8
2	No	GRF	101	PCA	Yes	SVM	54.2	10.3
3	No	GRF	1,001	PCA	Yes	SVM	54.1	11.2
4	Yes	GRF	1,001	PCA	No	SVM	54.0	10.3
5	Yes	GRF	101	PCA	No	SVM	53.9	10.3
6	No	GRF	101	PCA	No	SVM	53.8	9.8
7	Yes	GRF	1,001	PCA	Yes	SVM	53.7	11.6
8	Yes	GRF	101	PCA	Yes	SVM	53.6	11.3
9	Yes	ΔtGRF	1,001	PCA	No	SVM	53.5	10.6
10	Yes	ΔtGRF	101	PCA	No	SVM	53.2	10.3
11	Yes	ΔtGRF	101	PCA	Yes	SVM	53.2	10.8
12	Yes	ΔtGRF	1,001	PCA	Yes	SVM	53.2	10.6
13	No	GRF	1,001	PCA	No	MLP	53.0	9.7
14	Yes	GRF	101	PCA	No	MLP	52.7	9.2
15	Yes	GRF	1,001	PCA	No	MLP	52.7	10.0
16	No	GRF	101	PCA	Yes	MLP	52.6	10.2
17	Yes	GRF	1,001	PCA	Yes	MLP	52.6	10.2
18	No	GRF	1,001	PCA	Yes	MLP	52.3	9.6
19	Yes	ΔtGRF	101	TC	Yes	RFC	52.1	10.4
20	No	GRF	101	PCA	No	MLP	52.1	9.3
21	Yes	GRF	101	PCA	Yes	MLP	52.1	10.5
22	Yes	ΔtGRF	101	TC	Yes	MLP	51.6	9.6
23	Yes	ΔtGRF	1,001	PCA	Yes	MLP	51.6	9.4
24	Yes	ΔtGRF	101	PCA	No	MLP	51.6	10.6
25	Yes	ΔtGRF	101	TC	No	RFC	51.5	10.8
26	Yes	ΔtGRF	1,001	PCA	No	MLP	51.4	9.3
27	Yes	ΔtGRF	1,001	TC	Yes	RFC	51.4	10.7
28	Yes	ΔtGRF	101	PCA	Yes	MLP	51.4	10.5
29	Yes	ΔtGRF	101	TC	No	MLP	51.1	9.9
30	Yes	ΔtGRF	1,001	TC	No	RFC	51.1	10.5

*(1) The rounded percentage means and standard deviations of the F1-scores are shown; therefore, identical values may occur in the table. However, there are no pairwise identical values, so the ranking is unique. (2) A table including precision and recall (= accuracy) can be found in **Supplementary Table S3**.*

addition, the highest prediction F1-scors were also achieved with the GRF. However, it needs to be noted that the highest F1-score using ΔtGRF data were <1% lower than the highest F1-score using GRF data. Because the time derivative alone did not increase the prediction performance, it might be helpful to aggregate different feature extraction methods to improve classification models (Slijepcevic et al., 2020).

Time Normalization
The time normalization to 101 and 1,001 data points was significantly better than that to only 11 data points. These results are in line with current research, where 101 and 1,001 values are commonly used (Eskofier et al., 2013; Slijepcevic et al., 2017). Three of the four best ranks were achieved using the time normalization to 1,001 data points, but these were only slightly higher than those time normalized to 101 data points. In both methods, the best prediction performances where achieved in combination with PCA. In terms of computational

costs, it is advisable to weigh up to what extent relatively small improvements in the prediction performance justify the additional time required for classification. Furthermore, if computational cost is an important factor, a time normalization to fewer data points (above a certain minimum) could also be useful, since the results showed only little influence on the prediction performance.

Data Reduction
This analysis showed that PCA, which is frequently used in research (Figueiredo et al., 2018; Halilaj et al., 2018; Phinyomark et al., 2018), also achieves the highest F1-scores and ranks, compared with time-continuous waveforms and time-discrete gait variables. The highest F1-score of a machine-learning model based on time-continuous waveforms was 2.3% lower than that of PCA. Machine-learning models solely according to time-discrete characteristics is not recommended based on these analysis results. In line with Phinyomark et al. (2018),

TABLE 4 | Rank scores of all combinations of preprocessing methods depending on their mean F1-score over the 15-fold cross validation (n = 42).

	GRF filtering		Time derivative		Time normalization			Data reduction			Weight normalization		Classifier			
	No	Yes	GRF	ΔtGRF	11	101	1001	TC	TD	PCA	No	Yes	SVM	RFC	MLP	CNN
Score	18,564	22,764	22,953	18,375	10,392	15,512	15,424	14,337	6,870	20,121	20,635	20,693	11,926	12,170	10,373	6,859
% max	39.9	60.1	61.0	39.0	21.1	39.6	39.3	35.4	8.4	56.3	49.9	50.1	30.1	30.9	25.1	13.8

*(1) The total rank score is for each preprocessing step is 41,328. For GRF filtering, time derivative, and weight normalization the minimum rank score is 10,296 (0.0%) and the maximum rank score is 31,032 (100.0%). For time normalization and data reduction the minimum rank score is 4,560 (0.0%) and the maximum is 22,992 (66.7%). For the classifiers the minimum rank score is 2,556 (0%) and the maximum is 18,108 (50.0%). %max: relative rank score of ranks scaled to the interval between the minimum rank score and the maximum total rank score. (2) The rank scores for precision and recall (= accuracy) can be found in **Supplementary Tables S4, S5**.*

reducing the amount of data to the relevant characteristics is not only a cost-reducing method, but can also improve machine-learning classifications.

Weight Normalization

While weight normalization is necessary in inter-individual comparisons (Mao et al., 2008; Laroche et al., 2014), there have been no recommendations regarding intra-personal comparisons so far. The results of this analysis suggest that performing or not performing weight normalization leads to almost the same results and therefore shows no difference in prediction performance. Consequently, multiplication by a constant factor seems to play no role in the machine-learning classifications. This could be particularly interesting if different datasets are combined.

Machine-Learning Classifier

Four commonly used machine-learning classifiers (SVM, RFC, MLP, and CNN) were compared in this analysis. The RFCs achieved significantly higher mean F1-scores across all data preprocessing methods than the SVMs, MLPs, and CNNs. Compared to the other classifiers, the RFC seems to be most robust in case of a strong reduction of data (i.e., the time normalization to 11 data points or the unsupervised data reduction using the selection of time-discrete gait variables). However, the highest performance scores were achieved by SVMs followed by MLPs, RFCs, and CNNs. For gait data the SVM seems to be a powerful machine-learning classifier as often described in the literature (Figueiredo et al., 2018). The MLPs provided only mediocre prediction performances, which could be due to the fact that the total amount of data is simply too small for ANNs (Chau, 2001b; Begg and Kamruzzaman, 2005; Begg et al., 2005; Lai et al., 2008). This impression is reinforced by the even lower prediction performances of the CNNs as "deep" ANN architecture. In addition, the MLPs and CNNs required a lot of computation time for the classification, while the classification based on SVM and RFC was much more timesaving. Based on the presented results, using linear SVMs for the classification of gait data can be recommended. Furthermore, in line with recent research (Slijepcevic et al., 2020), a majority vote could possibly provide an even better classification. However, it should be noted that only a small selection of classifiers and architectures were examined in this analysis.

CONCLUSION

Based on a systematic comparison, the results provide first domain-specific recommendations for commonly used preprocessing methods prior to classifications using machine learning. However, caution is advised here, as the present findings may be limited to the classification task examined (six-session classification of intra-individual gait patterns) or even to the dataset. Furthermore, the derived recommendations are based exclusively on the prediction scores of the models. Therefore, no information can be obtained about the actual impact of the preprocessing methods and their combinations on the training process and the class representations of the

trained models. Overall, it can be concluded that preprocessing has a crucial influence on machine-learning classifications of biomechanical gait data. Nevertheless, further research on this topic is necessary to find out general implications for domain-specific standard procedures.

DATA AVAILABILITY STATEMENT

The dataset analyzed during the current study is available in the Mendely Data Repository (Burdack et al., 2020) (http://dx.doi.org/10.17632/y55wfcsrhz.1).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethical committee of the medical association Rhineland-Palatinate in Mainz (Germany). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FH, SD, and IH recorded the data. JB, FH, and WS conceived the presented idea. JB, FH, and SG performed the data analysis and designed the figures. JB and FH wrote the manuscript. JB, FH, SG, SD, IH, and WS reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00260/full#supplementary-material

REFERENCES

Alaqtash, M., Sarkodie-Gyan, T., Yu, H., Fuentes, O., Brower, R., and Abdelgawad, A. (2011a). Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2011, 453–457. doi: 10.1109/IEMBS.2011.6090063

Alaqtash, M., Yu, H., Brower, R., Abdelgawad, A., and Sarkodie-Gyan, T. (2011b). Application of wearable sensors for human gait analysis using fuzzy computational algorithm. *Eng. Appl. Artif. Intell.* 24, 1018–1025. doi: 10.1016/j.engappai.2011.04.010

Badesa, F. J., Morales, R., Garcia-Aracil, N., Sabater, J. M., Casals, A., and Zollo, L. (2014). Auto-adaptive robot-aided therapy using machine learning techniques. *Comput. Methods Programs Biomed.* 116, 123–130. doi: 10.1016/j.cmpb.2013.09.011

Begg, R. K., and Kamruzzaman, J. (2005). A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. *J. Biomech.* 38, 401–408. doi: 10.1016/j.jbiomech.2004.05.002

Begg, R. K., Palaniswami, M., and Owen, B. (2005). Support vector machines for automated gait classification. *IEEE Trans. Biomed. Eng.* 52, 828–838. doi: 10.1109/TBME.2005.845241

Bishop, C. M. (1995). *Neual Neworks for Pattern Recognition.* New York, NY: Oxford University Press.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92* (New York, NY: ACM), 144–152. doi: 10.1145/130385.130401

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Burdack, J., Horst, F., Giesselbach, S., Hassan, I., Daffner, S., and Schöllhorn, W. I. (2020). A public dataset of overground walking kinetics in healthy adult individuals on different sessions within one day. *Mendeley Data*, v1. doi: 10.17632/y55wfcsrhz.1

Challis, J. H. (1999). A procedure for the automatic determination of filter cutoff frequency of the processing of biomechanical data. *J. Appl. Biomech.* 15, 304–3017. doi: 10.1123/jab.15.3.303

Chau, T. (2001a). A review of analytical techniques for gait data. part 1: fuzzy, statistical and fractal methods. *Gait Posture* 13, 49–66. doi: 10.1016/S0966-6362(00)00094-1

Chau, T. (2001b). A review of analytical techniques for gait data. Part 2: neural network and wavelet methods. *Gait Posture* 13, 102–120. doi: 10.1016/S0966-6362(00)00095-3

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Academic Press.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Daffner, S. (2018). *Individualität und variablität der kinematischen und dynamischen gangmuster bei zwillingen* (Master's thesis). Johannes Gutenberg-University, Mainz, Germany.

Deluzio, K. J., and Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. *Gait Posture* 25, 86–93. doi: 10.1016/j.gaitpost.2006.01.007

Eskofier, B. M., Federolf, P., Kugler, P. F., and Nigg, B. M. (2013). Marker-based classification of young–elderly gait pattern differences via direct PCA feature extraction and SVMs. *Comput. Methods Biomech. Biomed. Engin.* 16, 435–442. doi: 10.1080/10255842.2011.624515

Ferber, R., Osis, S. T., Hicks, J. L., and Delp, S. L. (2016). Gait biomechanics in the era of data science. *J. Biomech.* 49, 3759–3761. doi: 10.1016/j.jbiomech.2016.10.033

Figueiredo, J., Santos, C. P., and Moreno, J. C. (2018). Automatic recognition of gait patterns in human motor disorders using machine learning : a review. *Med. Eng. Phys.* 53, 1–12. doi: 10.1016/j.medengphy.2017.12.006

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* 222, 309–368. doi: 10.1098/rsta.1922.0009

Flash, T., and Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* 5, 1688–1703. doi: 10.1523/JNEUROSCI.05-07-01688.1985

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. (2018). Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J. Biomech.* 81, 1–11. doi: 10.1016/j.jbiomech.2018.09.009

Hassan, I. (2019). *A dataset of overground walking kinetics in healthy adult individuals on different sessions within one day* (Dataset thesis). Johannes Gutenberg-University, Mainz, Germany.

Horst, F., Eekhoff, A., Newell, K. M., and Schöllhorn, W. I. (2017). Intra-individual gait patterns across different time-scales as revealed by means of a supervised learning model using kernel-based discriminant regression. *PLoS ONE* 12:e0179738. doi: 10.1371/journal.pone.0179738

Horst, F., Eekhoff, A., Newell, K. M., and Schöllhorn, W. I. (2019a). A public dataset of overground walking kinetics and lower-body kinematics in healthy adult individuals on different sessions within one day. *Mendeley Data*, v1. doi: 10.17632/b48n46bfry.1

Horst, F., Lapuschkin, S., Samek, W., Müller, K. R., and Schöllhorn, W. I. (2019b). Explaining the unique nature of individual gait patterns with deep learning. *Sci. Rep.* 9:2391. doi: 10.1038/s41598-019-38748-8

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441. doi: 10.1037/h0071325

Hsu, C., Chang, C., and Lin, C. (2003). *A Practical Guide to Support Vector Classification.* Technical report, National Taiwan University. Retrieved from: https://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf (accessed March 24, 2020).

Kaczmarczyk, K., Wit, A., Krawczyk, M., and Zaborski, J. (2009). Gait classification in post-stroke patients using artificial neural networks. *Gait Posture* 30, 207–210. doi: 10.1016/j.gaitpost.2009.04.010

Lai, D. T. H., Begg, R. K., Taylor, S., and Palaniswami, M. (2008). Detection of tripping gait patterns in the elderly using autoregressive features and support vector machines. *J. Biomech.* 41, 1762–1772. doi: 10.1016/j.jbiomech.2008.02.037

Laroche, D., Tolambiya, A., Morisset, C., Maillefert, J. F., French, R. M., Ornetti, P., et al. (2014). A classification study of kinematic gait trajectories in hip osteoarthritis. *Comput. Biol. Med.* 55, 42–48. doi: 10.1016/j.compbiomed.2014.09.012

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, M., Roan, M., and Smith, B. (2009). An application of principal component analysis for lower body kinematics between loaded and unloaded walking. *J. Biomech.* 42, 2226–2230. doi: 10.1016/j.jbiomech.2009.06.052

Mao, Y., Saito, M., Kanno, T., Wei, D., and Muroi, H. (2008). "Walking pattern analysis and SVM classification based on simulated gaits," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vancouver, BC), 5069–5072. doi: 10.1109/IEMBS.2008.4650353

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* 12, 181–201. doi: 10.1109/72.914517

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Curran Associates, Inc.), 8026–8037.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Retrieved from: http://www.jmlr.org/papers/v12/pedregosa11a.html (accessed March 24, 2020).

Phinyomark, A., Petri, G., Ibáñez-Marcelo, E., Osis, S. T., and Ferber, R. (2018). Analysis of big data in gait biomechanics: current trends and future directions. *J. Med. Biol. Eng.* 38, 244–260. doi: 10.1007/s40846-017-0297-2

Schölkopf, B., and Smola, A. J. (2002). *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: The MIT Press.

Schöllhorn, W. I. (2004). Applications of artificial neural nets in clinical biomechanics. *Clin. Biomech.* 19, 876–898. doi: 10.1016/j.clinbiomech.2004.04.005

Schreven, S., Beek, P. J., and Smeets, J. B. J. (2015). Optimising filtering parameters for a 3D motion analysis system. *J. Electromyogr. Kinesiol.* 25, 808–814. doi: 10.1016/j.jelekin.2015.06.004

Slijepcevic, D., Zeppelzauer, M., Gorgas, A. M., Schwab, C., Schuller, M., Baca, A., et al. (2017). Automatic classification of functional gait disorders. *IEEE J. Biomed. Health Inform.* 22, 1653–1661. doi: 10.1109/JBHI.2017.2785682

Slijepcevic, D., Zeppelzauer, M., Schwab, C., Raberger, A. M., Breiteneder, C., and Horsak, B. (2020). Input representations and classification strategies for automated human gait analysis. *Gait Posture* 76, 198–203. doi: 10.1016/j.gaitpost.2019.10.021

Su, F. C., and Wu, W. L. (2000). Design and testing of a genetic algorithm neural network in the assessment of gait patterns. *Med. Eng. Phys.* 22, 67–74. doi: 10.1016/S1350-4533(00)00011-4

Wannop, J. W., Worobets, J. T., and Stefanyshyn, D. J. (2012). Normalization of ground reaction forces, joint moments, and free moments in human locomotion. *J. Appl. Biomech.* 28, 665–676. doi: 10.1123/jab.28.6.665

Wolf, S., Loose, T., Schablowski, M., Döderlein, L., Rupp, R., Gerner, H. J., et al. (2006). Automated feature assessment in instrumented gait analysis. *Gait Posture* 23, 331–338. doi: 10.1016/j.gaitpost.2005.04.004

Zhang, J., Lockhart, T. E., and Soangra, R. (2014). Classifying lower extremity muscle fatigue during walking using machine learning and inertial sensors. *Ann. Biomed. Eng.* 42, 600–612. doi: 10.1007/s10439-013-0917-0

Predicting Knee Joint Instability Using a Tibio-Femoral Statistical Shape Model

Pietro Cerveri[1]*, Antonella Belfatto[1] and Alfonso Manzotti[2]

[1] *Department of Electronics, Information and Bioengineering, Polytechnic University of Milan, Milan, Italy, [2] Orthopaedic and Trauma Department, "Luigi Sacco" Hospital, ASST FBF-Sacco, Milan, Italy*

Statistical shape models (SSMs) are a well established computational technique to represent the morphological variability spread in a set of matching surfaces by means of compact descriptive quantities, traditionally called "modes of variation" (MoVs). SSMs of bony surfaces have been proposed in biomechanics and orthopedic clinics to investigate the relation between bone shape and joint biomechanics. In this work, an SSM of the tibio-femoral joint has been developed to elucidate the relation between MoVs and bone angular deformities causing knee instability. The SSM was built using 99 bony shapes (distal femur and proximal tibia surfaces obtained from segmented CT scans) of osteoarthritic patients. Hip-knee-ankle (HKA) angle, femoral varus-valgus (FVV) angle, internal-external femoral rotation (IER), tibial varus-valgus (TVV) angles, and tibial slope (TS) were available across the patient set. Discriminant analysis (DA) and logistic regression (LR) classifiers were adopted to underline specific MoVs accounting for knee instability. First, it was found that thirty-four MoVs were enough to describe 95% of the shape variability in the dataset. The most relevant MoVs were the one encoding the height of the femoral and tibial shafts (MoV #2) and the one representing variations of the axial section of the femoral shaft and its bending in the frontal plane (MoV #5). Second, using quadratic DA, the sensitivity results of the classification were very accurate, being all >0.85 (HKA: 0.96, FVV: 0.99, IER: 0.88, TVV: 1, TS: 0.87). The results of the LR classifier were mostly in agreement with DA, confirming statistical significance for MoV #2 ($p = 0.02$) in correspondence to IER and MoV #5 in correspondence to HKA ($p = 0.0001$), FVV ($p = 0.001$), and TS ($p = 0.02$). We can argue that the SSM successfully identified specific MoVs encoding ranges of alignment variability between distal femur and proximal tibia. This discloses the opportunity to use the SSM to predict potential misalignment in the knee for a new patient by processing the bone shapes, removing the need for measuring clinical landmarks as the rotation centers and mechanical axes.

Keywords: knee alignment, knee instability, femur shape, tibia shape, statistical shape model (SSM)

1. INTRODUCTION

The three-dimensional (3D) rotation of the femur with respect to the tibia, called tibio-femoral alignment, is a fundamental clinical index in knee diagnosis and surgical planning, as it can be correlated to a large extent to the kinematic instability of the joint (Laxafoss et al., 2013; Thienpont et al., 2014). This 3D rotation is represented by five main angular variables, namely the

hip-knee-ankle, femoral varus-valgus, and tibial varus-valgus angles, describing the knee stability in the frontal plane, and the internal-external femoral rotation and tibial slope, for the axial and sagittal alignments, respectively (Salenius and Vankka, 1975; Fitzpatrick et al., 2011; Schatka et al., 2018; Maillot et al., 2019). Specific bony landmarks (e.g., head center in the proximal femur, epicondyles and intercondylar fossa in the distal femur, epicondyles and frontal tuberosity in the proximal tibia, malleoli of the distal fibula, and the distal tibia) are mandatory for computing anatomical and mechanical axes and the corresponding tibio-femoral alignment in the knee (Lyras et al., 2016; Bennett et al., 2018). Clinical practice involves the manual detection of the landmarks on tomographic images or 3D reconstructed surfaces of bones and soft tissues. Both methods are time-consuming and prone to detection errors, even when performed by radiological and orthopedic experts. In order to improve landmark detection and tibio-femoral alignment computation, novel methodologies and tools, taking both semi- and fully-automatic approaches, have been proposed in the literature (De Momi et al., 2009; Cerveri et al., 2010; Subburaj et al., 2010; Kainz et al., 2015). However, such tools can fail in the case of large pathological deformations of the bony shapes. Indeed, as the degeneration progresses, the bony morphology deviates from the physiological shape, making the landmarks difficult to measure or even meaningless. In this scenario, landmark-free tools such as statistical shape models (SSMs) can represent an alternative for the evaluation of the knee joint alignment. SSMs have been extensively studied because of their ability to represent a set of matching surfaces synthetically in terms of a representative shape, namely the average surface of the set, and distinct morphological features, usually called "modes of variation" (MoVs). The magnitude of each MoV outlines the extent to which the morphological aspect it encodes is present in the set. Applications of bony surface SSMs in biomechanics and clinics have spanned anatomical and developmental studies (Li et al., 2010; Zhu and Li, 2011; Mutsvangwa et al., 2015; Baumbach et al., 2017; Wang and Shi, 2017; Zhang and Besier, 2017), shape anomaly staging (Van Haver et al., 2014; Agricola et al., 2015; Zhang et al., 2016; Cerveri et al., 2018; Chan et al., 2018), joint osteoarthritis (Neogi et al., 2013; Van Dijck et al., 2018), surgical planning and intervention (Zheng and Schumann, 2009; Cerveri et al., 2017; Mauler et al., 2017; Youn et al., 2017), and morphology-function relations (Fitzpatrick et al., 2011; Rao et al., 2013; Baka et al., 2014; Peloquin et al., 2014; Smoger et al., 2015; Hollenbeck et al., 2018; Cerveri et al., 2019b; Clouthier et al., 2019). There have, however, been few studies attempting to extensively investigate the relationship between morphological features and the degree of deformity of the tibio-femoral joint affecting the mechanical stability of the knee. This lack is probably due to the difficulty of considering the geometry of multiple bony structures and their relative position and to the complexity of building statistical models of pathological bones affected by severe deformations. In Rao et al., the authors elucidated the relationships between MoV and the relative alignment of the knee structures by means of an SSM built using magnetic resonance imaging of 20 knees (Rao et al., 2013).

Interestingly, they reported that some mechanical features of the tibia (anterior-posterior alignment and varus-valgus angle) and the femur (internal-external rotation) were encoded by specific MoVs. However, the tibio-femoral 3D misalignment was not explicitly encoded in the MoVs. Smoger et al. proposed to link the knee articular geometry and kinematics using an SSM built on 20 cadaveric specimens considered normal from a clinical point of view. Joint kinematic data of knee flexion/extension, captured by Kansas knee simulator, were used to compare experimental angular variables to the one simulated by the SSM (Smoger et al., 2015). Correlations between specific shapes in the knee and tibio-femoral alignment were reported. However, SSM parameter variations were not general enough to produce sufficient pathological alteration and bone deformations. In Clouthier et al., the authors studied the correlation between SSM parameters and the biomechanical factors of the knee using a statistical model built on 14 asymptomatic knees composed of distal femur, patella, and proximal tibia (Clouthier et al., 2019). SSMs were used to generate a number of morphological configurations of the bones, and each one was embedded into a lower-extremity musculo-skeletal model to evaluate the corresponding knee mechanics during a simulated gait cycle. The authors examined changes in knee mechanics (both bone kinematics and contact forces) as a function of the specific SSM realization. However, the SSM construction and experimental tests were performed on healthy subjects, so that SSM parameter variations did not generate extensive pathological conditions. For example, changes in the frontal plane affected the mechanical alignment by at most $\pm 3°$, which is considered the normal range for frontal stability of the knee. Based on such literature and capitalizing on our previous works (Cerveri et al., 2017, 2018, 2019a,b), in this paper, an SSM of pathological bony shapes in the knee is proposed to investigate the correlation between MoVs and the mechanical deformity of tibia and femur, assumed to induce kinematic instability. The statistical shape model of the tibia-femur bone complex was built using 99 pathological cases. The deformity degree was described in terms of 3D tibio-femoral alignment (**Figure 1**), considering the HKA (α), FVV (β), TVV (γ), IER (θ), and TS (ω) angles. For each angular variable, a clinical range from the literature, representing average physiological conditions, was selected to define the boundary between stability and instability. For each knee joint, the MoV weights were computed and their relation with each angular variable investigated. Discriminant analysis and logistic regression models (Wang, 2014) were adopted to systematically study the relations between observations (stability/instability classes as a binomial variable) and MoV weights (covariates). In the light of these premises and leveraging the main hypothesis of relationship between shape and function, the proposed work aims at linking specific MoVs in the SSM to the parameters describing the tibio-femoral alignment. This can have an impact in the biomechanical and orthopedic domains, as it opens up the opportunity to predict knee instability by analyzing the femoral and tibial morphology in terms of MoVs expressed by the SSM without the need for direct landmark identification and analysis.
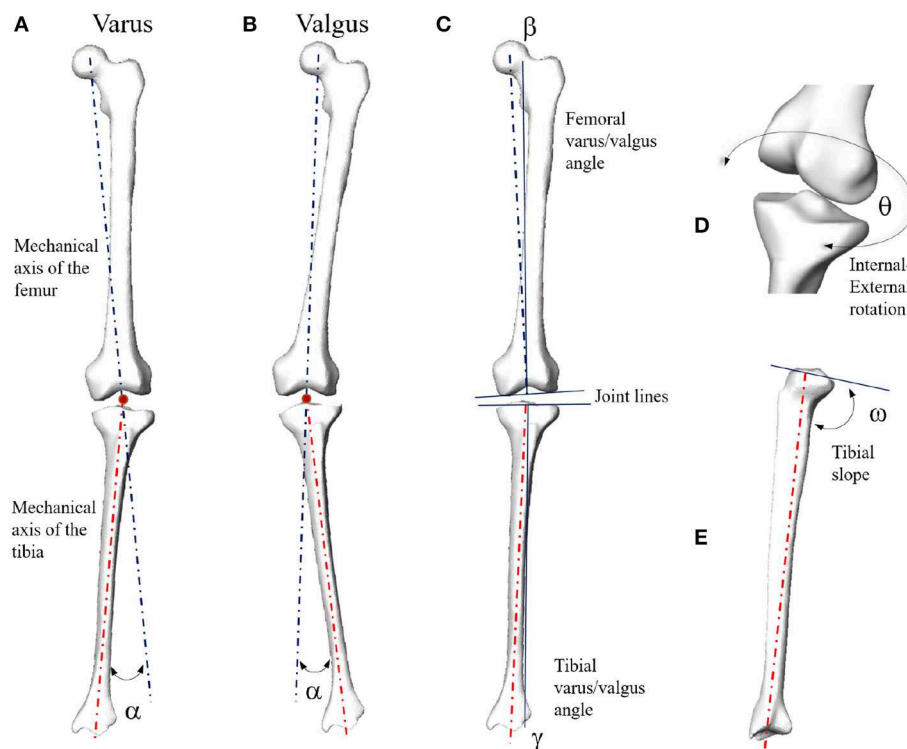
FIGURE 1 | Mechanical angular deformity (α) of the knee joint in the frontal plane (**A**: varus; **B**: valgus). Femoral (β) and tibial (γ) varus/valgus angles computed with respect to the corresponding joint lines (**C**). In case of parallel joint lines, $\beta + \gamma = \alpha$. Internal-external rotation (**D**) and tibial slope (**E**).

2. MATERIALS AND METHODS

2.1. Patient Data

Digital bony shapes of distal femur and proximal tibia were extracted from a retrospective dataset of 100 patients (70 males and 30 females) provided in anonymized form by Medacta company (Medacta International SA, Castel S. Pietro, CH), including planning CT scans (acquired in a supine position for all patients) and reconstructed bony 3D surfaces (Cerveri et al., 2017, 2018). The patients, aged 67 ± 10 years, reported localized knee pain associated with mechanical knee instability at staging time. Diagnostic imaging confirmed different degrees of cartilage defects, femoral osteophytes, and shape abnormalities, mainly at the condylar regions of the distal femur and at the tibial plateau. All patients underwent knee resurfacing or knee replacement surgery between 2014 and 2016. For surgical planning purposes, the image acquisition protocol included computed tomographic (CT) scans of the knee, hip, and ankle regions. Each CT scan consisted of about 520 slices with an image resolution of 512 $\times$ 512 pixel and a voxel size of 0.48 $\times$ 0.48 $\times$ 0.5 mm. Expert radiological operators manually performed the image segmentation of the osseous portion of the proximal and distal femur as well as of the proximal and distal tibia using Mimics software (Materialize, Belgium). For each labeled CT volume, the 3D surface meshes, composed of vertices and triangular faces, were reconstructed automatically. For all the patients, HKA, FVV, IER, TVV, and TS were computed pre-operatively,

exploiting landmarks manually detected on the surfaces. For SSM construction, distal femur and proximal tibia surfaces only were taken into account. As a function of the particular centering of the knee joint in the CT scan, the distal femur was segmented up to 2–4 cm away from the frontal notch of the trochlear region along the femur shaft. Similarly, the length of the proximal tibia shaft was variable across the set in a range of about 2–3 cm. Among the 100 cases, one was excluded from the set because of the presence of a fixation screw on the femoral shaft due to a previous intervention. All the valid surfaces underwent pre-processing by means of smoothing and sub-sampling starting from about 60,000 vertices, for both distal femur and proximal tibia samples, down to 10,000 vertices. Left surfaces were mirrored in the medio-lateral direction to obtain equivalent right surfaces for the construction of the right distal femur and proximal tibia SSM. The surface scale and the shaft lengths were not normalized. This is because, first, the normalization of a bundle of two shapes (femur and tibia) would have affected the relative size in between. Preserving the relative size of the two shapes in general increases the generality of the SSM (Pedoia et al., 2015). Second, the normalization would also affect the difference in the femur/tibial shaft lengths. The shaft length can be relevant for the bending in both sagittal and frontal planes.

2.2. Statistical Shape Model

In order to construct the SSM embedding femur and tibia shapes, the methodology extensively described in previous papers of our

group was adopted, which is based on a pair-wise matching technique (Cerveri et al., 2017, 2018, 2019a). This technique rests on the manual selection of a reference geometry for aligning all the surfaces in the training dataset and computing robust point correspondences. In the present custom implementation, first, the two reference geometries (distal femur and proximal tibia) were randomly selected within a subset of surfaces featuring only small bone deformations. Second, they were meshed and smoothed to obtain average edge lengths of 1.5 mm, resulting in triangular surfaces containing about 6,000 nodes each. This number of vertices is similar to the number of surface nodes used in previous works in the literature (Zheng and Schumann, 2009; Subburaj et al., 2010; Zhang et al., 2014). Each pair of surfaces in the overall set of 99 samples (distal femur and tibia) was rigidly registered to the reference tibio-femoral shape so that the relative position and the joint space between the two surfaces were preserved without requiring additional constraints. The deformable registration, based on a coherent point drift algorithm (Myronenko and Song, 2010), required to determine the point correspondences was, conversely, performed separately for femur and tibia to ensure consistency of the deformation field. A robust algorithm for determining one-to-one point correspondences (Cerveri et al., 2019a) was adopted in this work. The number of correspondences was determined by the number of vertices of the reference shape. After computing the mean model $\overline{m}$ from point correspondences, the covariance matrix, obtained by stacking the femur and tibia distance data from the mean model, underwent principal component analysis, providing 98 independent MoVs. Each MoV was represented by the eigenvector v_i and the corresponding eigenvalue σ_i. The percentage amount of morphological variation encoded by the jth MoV, termed explained variance (EV), was computed as:

$$EV_j = \frac{\sigma_j^2}{\sum_{i=1}^{M-1} \sigma_i^2} \quad (1)$$

where M is the number of samples in the dataset. The effect of each MoV was expressed numerically by one weight λ that modulates the corresponding eigenvalue, where a value of 0 denotes the mean shape, and negative and positive values represent the deviance from this mean in either direction. Accordingly, the SSM-based surface reconstruction, named morphing, was defined by the following equation:

$$\check{S} = \overline{m} + \sum_{i=1}^{M-1} \lambda_i \sigma_i v_i \quad (2)$$

where the reconstructed surface $\check{S}$ is obtained by summing up the mean model $\overline{m}$ with the series of all MoVs. For each surface, the weights were computed by projecting the shape pair (distal femur and proximal tibia) on the SSM by means of the scalar product (Cerveri et al., 2018). We retained enough MoVs to describe 95% of the overall shape variation, expressed by the cumulative EV, in the study population. The reconstructed surfaces were compared with the corresponding samples in the set by means of the surface distance error distribution (mean ± SD) using the Hausdorff distance.

TABLE 1 | Stability/Instability class definitions according to the thresholds for the five clinical variables.

Condition	HKA	FVV	IER	TVV	TS
Stability	28	25	84	79	70
Instability	71	74	15	20	29

2.3. Modeling Tibio-Femoral Alignment by SSM Parameters

In order to study the association between the MoVs and the condition of knee misalignment, the following normality ranges of the clinical variables were first considered: HKA: $0°\pm3$, FVV: $-6°\pm2$ (physiological valgum), IER: $\pm5°$, TVV: $\pm5°$, TS: $7°\pm4$ (Salenius and Vankka, 1975; Iranpour-Boroujeni et al., 2014; Driban et al., 2016; Schatka et al., 2018). According to each clinical variable, the 99 cases were separated into two classes, stable and unstable (**Table 1**). Note that the same instance may be considered stable according to one clinical parameter while being unstable according to another. As an example, two very different cases are depicted in **Figure 2**, the first lying within physiological ranges according to all the five angular quantities and the second featuring mechanical instability according to all but one (IER) angular quantities. We adopted two different data processing techniques, namely the discriminant analysis (DA) and the logistic regression (LR) classifiers. Both linear (LDA) and quadratic (QDA) discriminant analysis techniques were applied for the classification and the detection of low-dimensional sets of MoVs able to separate the stability from the instability condition. The accuracy (AC), sensitivity (SE), and specificity (SP) of the classification were computed for each dependent variable (clinical quantities) with respect to the explanatory variables (SSM parameters) using the leave-one-out (LOO) cross-validation technique. LOO classification based on LR was computed, and the statistical association ($p < 5\%$) between SSM parameters and the clinical quantities was determined. In order to further understand the contribution of each MoV in discriminating between stable and unstable conditions, the distributions of relevant MoVs (significant according to the previous analysis) were compared in the two conditions using a Wilcoxon signed-rank test ($p = 0.05$). Moreover, the correlations between the instability grade and each MoV were investigated. In other words, it was assumed that instability increased as the clinical parameter values drifted away from the reference physiological range and looked for a correspondence in MoV weight variations. Since both angle increases and angle decreases from the normal values relate to instability, a variable change was implemented, introducing a corrected version $\widehat{X} = |X - \overline{X}|$ of the clinical parameters by computing the absolute value of the difference between the parameter itself and its physiological average value, where X is a generic clinical parameter, $\overline{X}$ is its average value (in physiological cases), and $\widehat{X}$ is its corrected form. The correlation between the MoVs and the corrected parameters was assessed by means of the Spearman coefficient ($p = 0.05$).

3. RESULTS

3.1. Relation of MoVs to Bone Morphological Variability

Thirty-four MoVs were sufficient to describe 95% of the shape variability. Quantitatively, the reconstruction error across the 99 surfaces was 1.38 ± 0.16 mm. Qualitatively, the first ($EV_1 = 36.4\%$) MoV primarily encoded the isotropic scale of the bone complex. MoV #2 ($EV_2 = 16.3\%$) represented the size and height of the shafts of the two bones, concurrently. It also represented the shaft bending, mainly in the frontal plane (see **Table 2**). MoV #3 ($EV_3 = 9.9\%$) modeled the elongation of the femoral shaft ($\lambda_3 > 0$) and the shortening of the tibial shaft ($\lambda_3 < 0$). MoV #4 ($EV_4 = 5.6\%$) encoded the enlargement of the tibial plateau and the relative translation between the two shapes in the mediolateral direction. MoV #5 ($EV_5 = 3.6\%$) represented

FIGURE 2 | According to the five angular quantities, patient #20 featured no alignment deformation **(left)**. Patient #29, in contrast, featured mechanical instability in all but one (IER) angular quantities **(right)**.
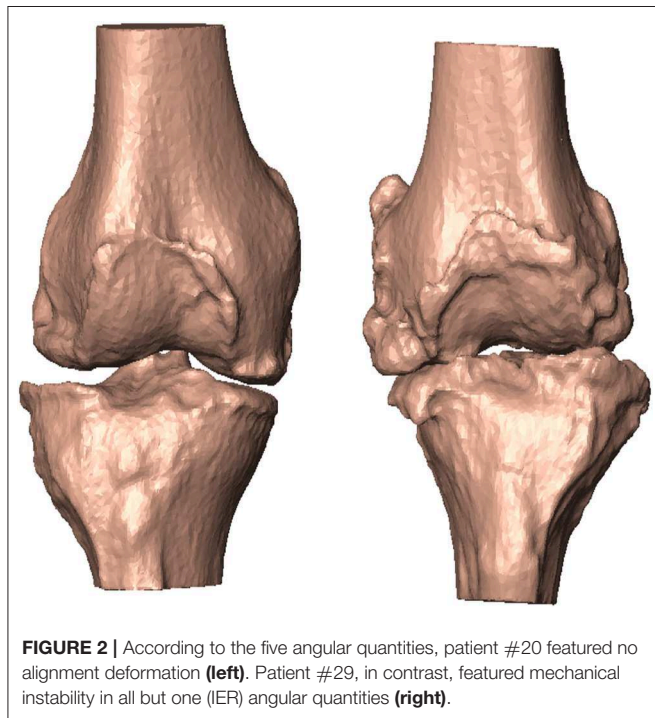
variations of the axial section of the femoral and tibial shafts and the relative bending between the two bones, in both the frontal and sagittal planes (**Table 2** and **Figure 3**). Positive values of the weight also encoded hypoplasia effects of the anterior facet of the medial condyle. MoV #6 ($EV_6 = 2.6\%$) described the concurrent modification of the anterior-posterior size of the femoral condylar region and that of the tibial plateau. Positive values of MoV #6 represented abnormal flatness in the trochlear region of the femur. This is unlike MoV #7 ($EV_7 = 2.1\%$), which modeled the tibial and femoral medio-lateral size, with positive values representing bone shrinkage. MoV #8 ($EV_8 = 1.9\%$) again represented the bending in the frontal plane between the two bones; however, the bending represented by MoV #5 was associated with a concurrent shrinkage/enlargement of the two shaft diameters, which was not encoded by MoV #8. MoV #9 ($EV_9 = 1.6\%$) modeled the medio-lateral shrinkage of the tibial plateau, with a concurrent anterio-posterior enlargement of the condylar region of the femur, up to pathological flattening. MoV #10 ($EV_10 = 1.6\%$) represented tibial inclination in both the frontal and sagittal planes. MoV #11 ($EV_11 = 1.3\%$) mainly represented tibia inclination on the sagittal plane, with a concurrent bending of the femur on the same plane. MoV #12 ($EV_12 = 1.1\%$) modeled a slight femoral bending on the frontal plane. MoV #13 ($EV_13 = 1.1\%$) modeled the relative bending between the two bones in the sagittal plane. The remaining MoVs represented small and spread morphological variations and was not straightforward to visually relate any to specific local features (see **Supplementary Materials**).

3.2. MoV Performance in Instability Modeling

3.2.1. Discriminant Analysis

Table 3 shows the classification performances (sensitivity, specificity, and accuracy) obtained using the LOO procedure for both the linear discriminant analysis and quadratic discriminant analysis (34 MoVs were employed in the task). Despite the fact that the LDA accuracy ranged between 0.7 and 0.84, the respective values of sensitivity and specificity were highly different: in two cases (IER and TS), the sensitivity was lower than 0.4 (specificity > 0.8), while, conversely, for HKA and FVV the sensitivity was as high as 0.85, while specificity reached 0.64 and 0.44, respectively. As far as QDA is concerned, higher

TABLE 2 | Morphological variability of femur (F) and tibia (T) and relative alignments mapped onto the MoVs from 2 to 9.

Bone variability	MoV #2	MoV #3	MoV #4	MoV #5	MoV #6	MoV #7	MoV #8	MoV #9
F shaft elongation	o	o						
F shaft diameter	o			o				
F shaft bending	o			o			o	
F condylar AP size					o			o
F condylar ML size						o		
T shaft elongation	o	o						
T shaft diameter	o			o				
T shaft bending	o			o			o	
T plateau AP size			o		o			
T plateau ML size						o		o

FIGURE 3 | (Upper) Morphological deviations (mm) from the mean model (frontal view) for MoVs from #1 to #5. **(Lower)** Explicit morphology in two SSM parameter sets (MoV #3 -3σ, MoV #4 +3σ) with the overlaid mean model.

TABLE 3 | Classification test, exploiting LOO cross-validation, of linear vs. quadratic discriminant analysis using all SSM parameters.

	LDA-AC	LDA-SE	LDA-SP	QDA-AC	QDA-SE	QDA-SP
HKA	0.82	0.88	0.64	0.96	0.96	1
FVV	0.74	0.85	0.44	0.99	0.99	1
IER	0.80	0.33	0.89	0.88	1	0.86
TVV	0.84	0.55	0.92	1	1	1
TS	0.70	0.37	0.82	0.87	0.79	0.91

TABLE 4 | The three most representative SSM parameters for quadratic discriminant analysis.

	MoVs	QDA-AC	QDA-SE	QDA-SP
HKA	2, 5, 13	0.83	0.88	0.71
FVV	5, 14, 17	0.79	0.95	0.32
IER	2, 9, 14	0.92	0.53	1
TVV	5, 12, 15	0.87	0.40	1
TS	3, 13, 18	0.81	0.44	0.97

0.93, 0.56), (0.84, 0.13, 0.97), (0.83, 0.75, 0.97), and (0.75, 0.31, 0.94), respectively. The statistical analysis provided significance ($p < 0.05$) in HKA for MoVs #5, #7, and #18, FVV for MoV #5, IER for MoV#2, TVV for MoVs #11, #14, #16, and #17, and TS for MoV #5 (**Table 6**). Nicely, MoVs #2 and MoV #5 were found to be largely representative of the logistic modeling, in agreement with the DA results. For these two MoVs, the box plots were reported in order to highlight the distribution differences across mechanically stable and unstable cases for each clinical parameter (**Figure 4**). As far as MoV #2 is concerned, stability and instability were significantly different in IER distributions ($p = 0.02$). As far as MoV #5 is concerned, both HKA ($p = 0.0005$) and TS ($p = 0.002$) resulted in significant differences. As far as the correlation analysis is concerned, MoV #5 showed significant correlation with $\widehat{HKA}$ (c = -0.52, $p < 10^{-7}$), $\widehat{FVV}$ (c = -0.26, $p < 0.01$), and $\widehat{TS}$ (c = -0.23, $p < 0.03$). Likewise, MoV #7 showed significant correlation with $\widehat{HKA}$ (c = -0.27, $p < 0.008$) and $\widehat{FVV}$ (c = -0.32, $p < 0.002$), while MoV #6 resulted in significant correlation only in the case of $\widehat{FVV}$ (c = 0.25, $p < 0.02$) (see **Table 7**). A scatter plot showing $\widehat{HKA}$ against MoV #5 was reported in **Figure 5**.

accuracy levels (range: 0.87–1) corresponded instead with both high sensitivity (range: 0.79–1) and high specificity (range: 0.86–1). It is worth noticing that reducing the MoVs to the three or four most relevant as shown in **Tables 4, 5**, respectively, reduces the performance, again causing specificity/sensitivity mismatches. In fact, in both cases (**Tables 3, 4**), poor accuracy was found for TVV (0.4 and 0.55) and TS (0.44 and 0.48), while FVV (0.32 and 0.60) resulted in low specificity. Nevertheless, it has to be pointed out that in **Table 5**, using four MoVs, only one value fell slightly below the threshold of 0.5 (TS sensitivity: 0.48) suggesting that, even in this reduced form, QDA was able to outperform LDA. Interestingly, considering the four-MoV-based QDA, both MoVs #2 and #5 were representative of all of the clinical measures except for TVV (MoV #5 only).

3.2.2. Logistic Regression
Classification results (AC, SE, SP) with LOO cross-validation for HKA, FVV, IER, TVV, and TS were (0.79, 0.88, 0.57), (0.83,

4. DISCUSSION AND CONCLUSIONS

4.1. Findings, Limitations, and Possible Developments
Computational approaches to study the correlation between morphological features and functional or pathological conditions of bony surfaces using SSM have been emerging in the literature, with impacts in biomechanics, especially for kinematic and dynamic analysis (Rao et al., 2013; Smoger et al., 2015; Nolte et al., 2016; Zhang et al., 2016; Hollenbeck et al., 2018; Clouthier et al., 2019), and clinics, especially for diagnostic and surgical interests

TABLE 5 | The four most representative SSM parameters for quadratic discriminant analysis.

	Param	QDA-AC	QDA-SE	QDA-SP
HKA	2, 5, 11, 17	0.88	0.93	0.75
FVV	2, 5, 10, 17	0.83	0.91	0.60
IER	2, 8, 9, 12	0.95	0.66	1
TVV	5, 12, 14, 15	0.88	0.55	0.97
TS	2, 5, 8, 11	0.84	0.48	0.99

TABLE 6 | SSM parameters that were statistically significant for the logistic regression.

	MoV/p value
HKA	5 ($p = 0.0001$), 7 ($p = 0.03$), 17 ($p = 0.02$), 18 ($p = 0.03$)
FVV	5 ($p = 0.001$), 10 ($p = 0.01$), 17 ($p = 0.01$)
IER	2 ($p = 0.02$)
TVV	11 ($p = 0.008$), 14 ($p = 0.01$), 16 ($p = 0.01$), 17 ($p = 0.03$)
TS	5 ($p = 0.02$)

(Neogi et al., 2013; Peloquin et al., 2014; Mutsvangwa et al., 2015; Cerveri et al., 2018). In particular, three studies addressed the relation between SSM parameters and knee kinematics by focusing on the link between the morphological variability of the bones and tibio-femoral alignment modifications (Rao et al., 2013; Smoger et al., 2015; Clouthier et al., 2019). The main issue of such studies was that the relationships between shape and alignment were simulated by systematically perturbing MoVs, reaching up to 95% variation with respect to the mean model. However, SSMs were computed using a very small group of asymptomatic cases. Therefore, pathological conditions were not explicitly encoded in the MoVs, leading to the simulation of mechanical axis misalignment within normality ranges. This hindered the model's ability to extrapolate non-physiological conditions of the knee. Conversely, in our work, the SSM analysis was addressed by considering a population of 99 knee cases with different morphological and mechanical anomalies at:

- the distal femur, namely condylar enlargement, osteophytes, notch narrowing, trochlear flattening;
- the proximal tibia, namely plateau enlargement, osteophytes, smoothing of the intercondylar eminence, flattening of the tibial tuberosity.

Moreover, for each case, the tibio-femoral alignment of the knee was measured not only in the frontal plane (HKA, TVV, and FVV) but also in the sagittal (TS) and axial (IER) planes. We built an SSM using the two bone sets (distal femur and proximal tibia), computed the MoV weights for each case, tested both DA and LR classifiers of stability/instability as a function of MoV weights, and analyzed the relevance of each MoV for joint instability. The major findings of the work can be summarized as:

- the computed SSM was representative of the surface set, demonstrated by the very low reconstruction error;
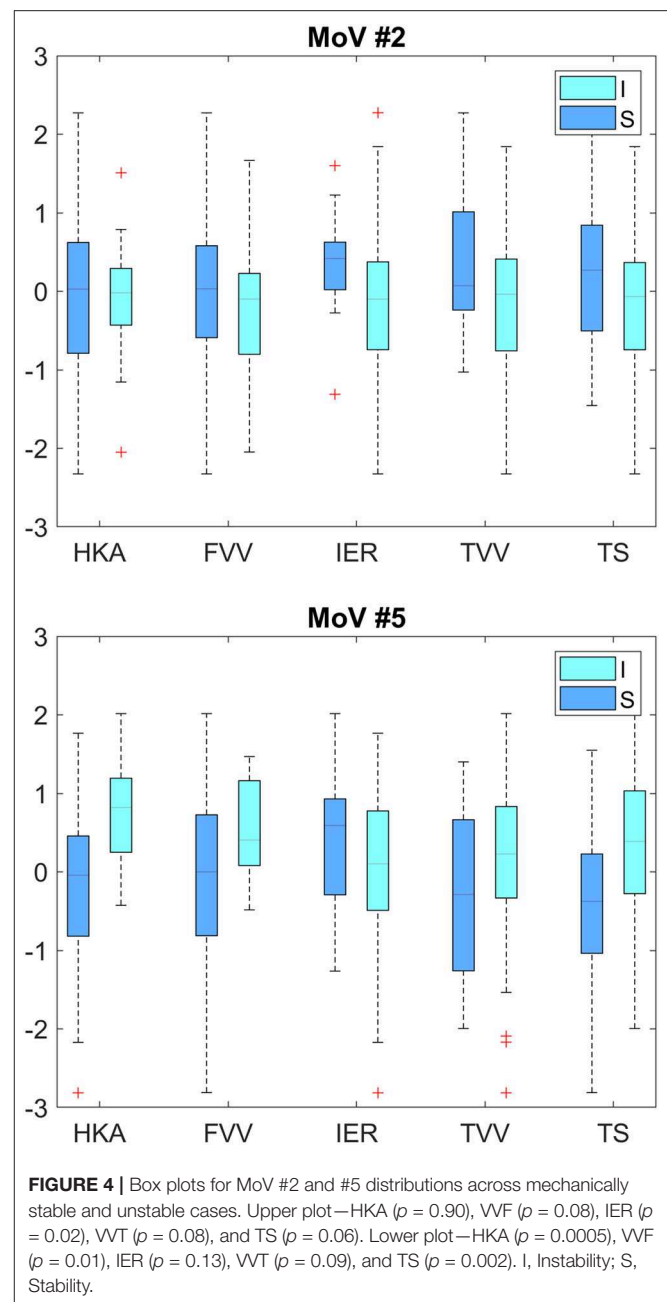
FIGURE 4 | Box plots for MoV #2 and #5 distributions across mechanically stable and unstable cases. Upper plot—HKA ($p = 0.90$), VVF ($p = 0.08$), IER ($p = 0.02$), VVT ($p = 0.08$), and TS ($p = 0.06$). Lower plot—HKA ($p = 0.0005$), VVF ($p = 0.01$), IER ($p = 0.13$), VVT ($p = 0.09$), and TS ($p = 0.002$). I, Instability; S, Stability.

- the physiological and pathological variations of the knee morphology found in the surface dataset were consistently encoded by 34 MoVs (EV > 0.95);
- together, MoV #2 (height/size of femoral/tibial shafts) and MoV #5 (femoral/tibial bending in the frontal/sagittal plane) were the most relevant MoVs, representing a total of about 20% of the variation among SSMs;
- MoV classification results were largely in agreement with morphological features determining tibio-femoral instability (compare **Table 2** with **Tables 4, 5**);
- QDA outperformed LDA in classifying unstable versus stable cases with high accuracy for all the five clinical parameters;

TABLE 7 | Correlation intensity (Spearman coefficient c) and significance (*p*-value) between relevant MoVs and clinical parameters defining knee instability.

MoV	HKA		FVV		IER		TVV		TS	
	c	p	c	p	c	p	c	p	c	p
#1	−0.14	0.17	−0.09	0.36	0.06	0.53	−0.17	0.09	0.15	0.15
#2	0.11	0.26	0.10	0.35	0.18	0.07	0.09	0.40	0.03	0.74
#3	0.14	0.18	0.02	0.84	−0.07	0.47	−0.04	0.68	0.17	0.09
#4	0.12	0.22	0.17	0.10	0.09	0.35	0.05	0.59	0.11	0.26
#5	−0.52	10^{-8}	−0.26	0.01	0.08	0.44	−0.17	0.10	−0.23	0.02
#6	0.14	0.16	0.25	0.01	0.03	0.78	0.15	0.12	0.07	0.49
#7	−0.27	0.01	−0.32	0.001	0.05	0.63	−0.11	0.27	0.08	0.43
#8	−0.11	0.28	0.01	0.92	0.16	0.12	10^{-4}	0.99	−0.10	0.32
#9	0.03	0.78	−0.05	0.65	0.19	0.05	0.01	0.90	0.07	0.47
#10	−0.05	0.59	−0.10	0.34	−0.02	0.82	−0.05	0.65	−0.02	0.83

MoVs #5 and #7 feature significant correlation results.

- despite the fact that the LR-based classification provided lower-accuracy results, statistically significant MoVs were in agreement with QDA.

Synthetically, these findings suggest that the computed SSM can be exploited for assessing whether a knee lies in a pathological condition according to the more traditional clinical parameters, namely HKA, FVV, IER, TVV, and TS, without the need for landmark selection, just fitting the SSM to the shape of interest. In more detail, the SSM decomposition showed that the first 13 were sufficient to describe 85% of the explained variance, demonstrating the SSM's ability to model large morphological variability in a very compact way. MoVs of the SSM were also related to tibio-femoral alignment and knee instability according to the five clinical parameters considered. This was confirmed by the classification performance, because four MoVs (see **Table 5**) were able to ensure more than 80% of accuracy in the quadratic discriminant analysis. Again, this makes SSM a prospective candidate tool for distinguishing stable and unstable knee conditions by analyzing the surfaces only, without the manual definition of rotation centers and mechanical axes.

An in-depth analysis of the classification performances showed that the LDA model was under-fitting. The size imbalance of the two classes (see **Table 1**) further contributed to bias the results. This was evident (see **Table 3**), for example, when considering the HKA (featuring only 28 stable cases with respect to 71 unstable cases) and IER (only 15 unstable conditions). Conversely, QDA appeared to be more robust to dataset imbalance, showing both higher sensitivity and specificity than LDA. The LR analysis highlighted a couple of MoVs relevant for discriminating between stability and instability, namely MoVs #2 and #5, representative of all the clinical measures. These two specific MoVs were found to be significant to discriminate between stability and instability. Specifically, MoV #2 mainly encoded the elongation and partially encoded the bending of the two shafts. This is in agreement with the relation with the variation of the two mechanical axes and, by consequence, with their relative inclination. This can therefore be related to the joint mechanical alignment, especially in the frontal plane. MoV #5, encoding the relative bending of the two bones in the frontal

plane (see **Table 2**), was confirmed to be related to the HKA, FVV, and TVV angles, which describe the tibio-femoral stability in the frontal plane. As confirmed by the classification results, this MoV was able to discriminate between stability and instability. In synthesis, while the effect of these two MoVs could not be predicted a priori, the morphological aspects encoded by both of them could reasonably be considered to be related to the tibio-femoral alignment. It has to be pointed out, however, that differences in the bone shaft heights encoded in MoV #2 were caused by different ranges of interest in CT scans and could not be ascribed to morphological variability. However, it is reasonable to assume that the frontal bending and lateral inclination of the shaft are morphological features relevant for the overall tibio-femoral alignment. Therefore, MoV #2 was not discarded, a choice that was justified a-posteriori, considering that it was relevant for the classification.

One shortcoming of our work is the inclusion of just the femoral and tibial surfaces, neglecting the patellar region and the cartilages. Nonetheless, this choice was motivated by previous literature findings showing that increasing the number of geometries to be included in the SSM can easily lead to difficulty in identifying specific correspondences between MoVs and morphological features. For instance, Fitzpatrick et al. used 26 healthy subjects to develop an SSM of the patello-femoral joint, reporting that the main variability of the patella articular curvature and sulcus groove was actually spread across many different MoVs (Fitzpatrick et al., 2011). As a matter of fact, our approach allowed us to identify correlations between specific MoVs and clinical parameters of the alignment between femur and tibia. In this paper, we did not investigate how a different reference shape selection would have affected the reconstruction and the classification results. The reference shape was selected randomly from a subset of surfaces little-affected by deformities. This was in agreement with the results reported in a recent paper with similar acquisition techniques, where the random selection of the SSM also had little effect (Cerveri et al., 2019a). As far as data acquisition is concerned, all the patients were lying supine during CT acquisition and images were acquired using the same protocol. As regards data processing, the images were segmented by different expert radiological

operators. Each scan was segmented by one operator, so we did not have any information about variability in segmentation. Similarly, the landmark detection and the angle computation were performed by one expert orthopedic surgeon. As the scans were all at sub-millimeter resolution, the bony segmented profiles were affected by such uncertainty, which was present in the final surface reconstruction. The surface sub-sampling lowered, on average, the surface quality by <2% (root-mean-squared-distance: <1 mm) with respect to the original surfaces. The SSM reconstruction error was, on average, lower than 1.5 mm, reasonably localized in the region affected by the largest pathological deformations. Actually, we focused on the overall SSM reconstruction ability without taking local errors into account. The analysis of the reconstruction quality in critical regions heavily affected by deformities (e.g., the presence of large osteophytes) could have provided further information about the specificity and generality of the SSM model. However, this analysis would have required a greater effort by the expert operators to manually detect and classify the regions with severe deformations, which is a very time-consuming task beyond the scope of the present work. Conversely, we aimed at relating MoV weights to angular stability determined by the five clinical indices. It is reasonable to assume that the local reconstruction errors should affect the overall knee joint alignment less, which should be mainly determined by the overall bony shape. Nevertheless, analysis of how local reconstruction errors could affect the relative 3D rotation may be carried out in a future study by means of a sensitivity analysis. Finally, the SSM model could be used to study the development of stress and strain in the knee due to applied loads as a function of surface geometry changes. This would require that a finite element description be integrated into the SSM to perform the computations, which could be used to predict the outcome of surgery, taking into account patient-specific variability. Moreover, in order to analyze the effects of the relative tibio-femoral position and orientation on gait motion patterns, the SSM could be combined with the angular trajectories reconstructed using surface markers acquired by means of an opto-electronic motion capture system. For instance, simulations could be useful for evaluating how gait cycle parameters (e.g., gait cadence, step length, etc.) would be affected. Likewise, the SSM could help to study, in knee surgical planning, how the tibio-femoral spacing would change the flexion-extension patterns of knee.

4.2. Literature Comparison

Rao et al. developed an SSM of the femur, tibia, and patella of 20 cadavers, considering different alignments obtained by using a mechanical simulator applied to the specimens (Rao et al., 2013). About 95% of the variability was captured by just 15 MoVs. Fitzpatrick et al. used 26 healthy subjects to develop an SSM model of the patello-femoral joint (Fitzpatrick et al., 2011). Similarly, 15 MoVs were sufficient to capture about 97% of the morphological variability. Fourteen asymptomatic patients scanned by MRI were used in Clouthier et al. to build an SSM of the knee that was able to represent 70% of the variability by means of 6 MoVs only (Clouthier et al., 2019).
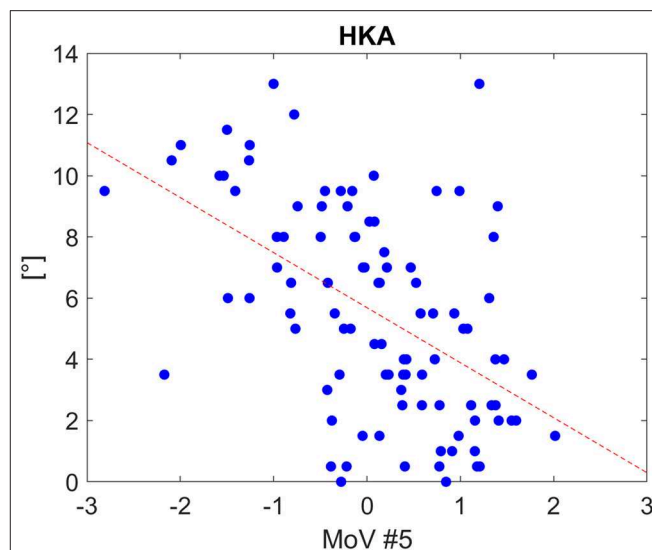
FIGURE 5 | Scatter plot for HKA distribution as a function of MoV #5 range, along with the linear fitting (Spearman correlation coefficients: −0.52 with p<1e−07). The tendency line is depicted (red dashed line).

In our work, we used a wide dataset of pathological knees featuring large anomalies at both femoral and tibial sites. As a consequence, the greater number of MoVs needed to represent most of the variability (34 MoVs accounting for 95% EV) was to be expected. This corroborates the consideration that morphological abnormalities cannot be simply extrapolated by an SSM built on normal joints. In other words, femoral and tibial deformities cannot be represented just by enlarging the weight range of the MoVs (e.g., ± 3, ± 4, ± 5 SD, etc), but, rather, there is a need to encode such information in new MoVs entirely. This is also in agreement with the limitations acknowledged in the previous literature (Fitzpatrick et al., 2011; Rao et al., 2013; Smoger et al., 2015; Clouthier et al., 2019). For instance, Hollenbeck et al. reported that a maximum range of ± 2 SD was allowed in their lumbar spine SMM in order to avoid unrealistic deformations (Hollenbeck et al., 2018). As far as the relation between MoVs and kinematics is concerned, Smoger et al. reported that their third MoV described differences in the internal-external relative rotation between femur and tibia (Smoger et al., 2015); this was nicely in agreement with our results for MoV #5. Internal−external alignment of the patellofemoral joint was described by the second mode in Rao et al. (2013). Interestingly, they also found tibial internal–external rotation and tibial varus-valgus variations encoded by the third and fourth MoVs, respectively. However, femoral alignments were not modeled by their SSM. In Pedoia et al., the authors developed distal femur and proximal tibia SSMs, avoiding the normalization of the samples (Pedoia et al., 2015). They reported that the first mode was related to the size for both SSMs, as in our case. For the femur, modes #2 and #3 were related to the relative distance between the condyles and the condylar width and height, respectively. In our model, these features were mainly encoded by MoV #6 and #7. As far as the tibia

FIGURE 6 | The application workflow.
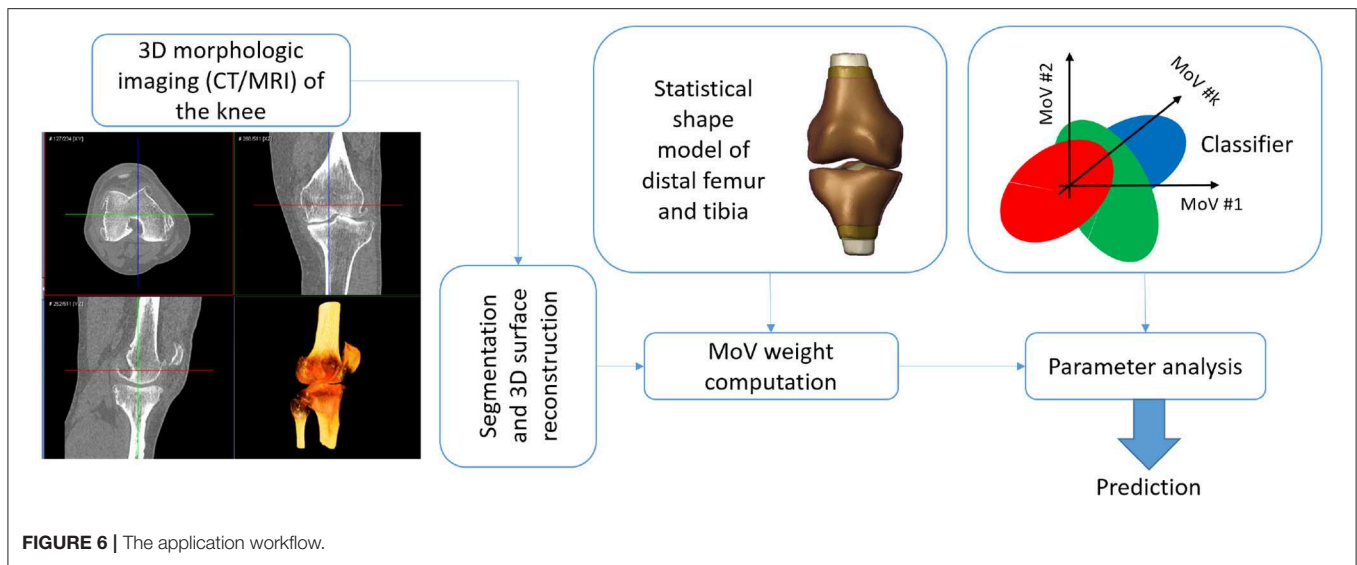
was concerned, modes #2 and #3 were related to the medial posterior curvature of the tibial plateau and the elevation of the anteromedial tibial plateau, respectively. In our model, these morphological features were spread across MoVs #4, #6, and #7. These differences were expected because we dealt with a unique SSM for the tibio-femoral joint. Our method may provide insights regarding concurrent morphological deformations in the two bones.

5. CONCLUSION

The SSM approach was proven to consistently represent both morphological anomalies and alignment deviation in the knee bones by means of few representative MoVs. In the light of such results, the proposed SSM met the objectives of providing an alternative to manual definition of bone landmarks to assess pathological conditions related to knee instability. The SSM could be exploited in a decision support system that predicts the potential instability of the joint by processing the knee scan without requiring images of other body regions (e.g., hip and ankle) and with no need for manual landmark identification. This toolbox could thus generate an automatic report with a diagnosis of stable or unstable according to each clinical variable of the five indexes considered. A potential application workflow would rest on: (1) the bone segmentation in the knee scan; (2) the surface reconstruction; (3) the weight computation by the SSM; (4) the prediction of the instability based on the five different clinical factors of alignment by means of a classifier (e.g., discriminant analysis) (**Figure 6**). Another possible exploitation of the proposed SSM approach is the simulation of the effects of different morphological conditions (achieved by varying MoV weights) on movement analysis of the knee, as suggested by Smoger et al. (2015) and Clouthier et al. (2019), studies that both proposed SSMs built on healthy subjects. An SSM including large pathological

variability, such as the one developed in this work, opens up the opportunity to study the effect of a specific misalignment of the femur and tibia on the simulated motion pattern and, consequently, the resulting load distribution affecting cartilage wear.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study involved retrospective anonymized digital data provided by Medacta International SA (CH) in accordance with the institutional ethical committee.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe. 2020.00253/full#supplementary-material

REFERENCES

Agricola, R., Leyland, K. M., Bierma-Zeinstra, S. M. A., Thomas, G. E., Emans, P. J., Spector, T. D., et al. (2015). Validation of statistical shape modelling to predict hip osteoarthritis in females: data from two prospective cohort studies (cohort hip and cohort knee and chingford). *Rheumatology* 54, 2033–2041. doi: 10.1093/rheumatology/kev232

Baka, N., Kaptein, B. L., Giphart, J. E., Staring, M., de Bruijne, M., Lelieveldt, B. P. F., et al. (2014). Evaluation of automated statistical shape model based knee kinematics from biplane fluoroscopy. *J. Biomech.* 47, 122–129. doi: 10.1016/j.jbiomech.2013.09.022

Baumbach, S. F., Binder, J., Synek, A., Mück, F. G., Chevalier, Y., Euler, E., et al. (2017). Analysis of the three-dimensional anatomical variance of the distal radius using 3D shape models. *BMC Med. Imaging* 17:23. doi: 10.1186/s12880-017-0193-9

Bennett, H. J., Weinhandl, J. T., Fleenor, K., and Zhang, S. (2018). Frontal plane tibiofemoral alignment is strongly related to compartmental knee joint contact forces and muscle control strategies during stair ascent. *J. Biomech. Eng.* 140:061011. doi: 10.1115/1.4039578

Cerveri, P., Belfatto, A., Baroni, G., and Manzotti, A. (2018). Stacked sparse autoencoder networks and statistical shape models for automatic staging of distal femur trochlear dysplasia. *Int. J. Med. Robot. Comput. Assist. Surg.* 14:e1947. doi: 10.1002/rcs.1947

Cerveri, P., Belfatto, A., and Manzotti, A. (2019a). Pair-wise vs. group-wise registration in statistical shape model construction: representation of physiological and pathological variability of bony surface morphology. *Comput. Methods Biomech. Biomed. Eng.* 22, 772–787. doi: 10.1080/10255842.2019.1592378

Cerveri, P., Belfatto, A., and Manzotti, A. (2019b). Representative 3D shape of the distal femur, modes of variation and relationship with abnormality of the trochlear region. *J. Biomech.* 94, 67–74. doi: 10.1016/j.jbiomech.2019.07.008

Cerveri, P., Marchente, M., Bartels, W., Corten, K., Simon, J.-P., and Manzotti, A. (2010). Automated method for computing the morphological and clinical parameters of the proximal femur using heuristic modeling techniques. *Ann. Biomed. Eng.* 38, 1752–1766. doi: 10.1007/s10439-010-9965-x

Cerveri, P., Sacco, C., Olgiati, G., Manzotti, A., and Baroni, G. (2017). 2D/3D reconstruction of the distal femur using statistical shape models addressing personalized surgical instruments in knee arthroplasty: a feasibility analysis. *Int. J. Med. Robot. Comput. Assist. Surg.* 13, 1–13. doi: 10.1002/rcs.1823

Chan, E. F., Farnsworth, C. L., Klisch, S. M., Hosalkar, H. S., and Sah, R. L. (2018). 3-dimensional metrics of proximal femoral shape deformities in legg-calvé-perthes disease and slipped capital femoral epiphysis. *J. Orthop. Res.* 36, 1526–1535. doi: 10.1002/jor.23791

Clouthier, A. L., Smith, C. R., Vignos, M. F., Thelen, D. G., Deluzio, K. J., and Rainbow, M. J. (2019). The effect of articular geometry features identified using statistical shape modelling on knee biomechanics. *Med. Eng. Phys.* 66, 47–55. doi: 10.1016/j.medengphy.2019.02.009

De Momi, E., Lopomo, N., Cerveri, P., Zaffagnini, S., Safran, M. R., and Ferrigno, G. (2009). *In-vitro* experimental assessment of a new robust algorithm for hip joint centre estimation. *J. Biomech.* 42, 989–995. doi: 10.1016/j.jbiomech.2009.02.031

Driban, J. B., Stout, A. C., Duryea, J., Lo, G. H., Harvey, W. F., Price, L. L., et al. (2016). Coronal tibial slope is associated with accelerated knee osteoarthritis: data from the osteoarthritis initiative. *BMC Musculoskelet. Disord.* 17:299. doi: 10.1186/s12891-016-1158-9

Fitzpatrick, C. K., Baldwin, M. A., Laz, P. J., FitzPatrick, D. P., Lerner, A. L., and Rullkoetter, P. J. (2011). Development of a statistical shape model of the patellofemoral joint for investigating relationships between shape and function. *J. Biomech.* 44, 2446–2452. doi: 10.1016/j.jbiomech.2011.06.025

Hollenbeck, J. F. M., Cain, C. M., Fattor, J. A., Rullkoetter, P. J., and Laz, P. J. (2018). Statistical shape modeling characterizes three-dimensional shape and alignment variability in the lumbar spine. *J. Biomech.* 69, 146–155. doi: 10.1016/j.jbiomech.2018.01.020

Iranpour-Boroujeni, T., Li, J., Lynch, J. A., Nevitt, M., Duryea, J., and Investigators, O. (2014). A new method to measure anatomic knee alignment for large studies of oa: data from the osteoarthritis initiative. *Osteoarthr. Cartil.* 22, 1668–1674. doi: 10.1016/j.joca.2014.06.011

Kainz, H., Carty, C. P., Modenese, L., Boyd, R. N., and Lloyd, D. G. (2015). Estimation of the hip joint centre in human motion analysis: a systematic review. *Clin. Biomech.* 30, 319–329. doi: 10.1016/j.clinbiomech.2015.02.005

Laxafoss, E., Jacobsen, S., Gosvig, K. K., and Sonne-Holm, S. (2013). The alignment of the knee joint in relationship to age and osteoarthritis: the copenhagen osteoarthritis study. *Skelet. Radiol.* 42, 531–540. doi: 10.1007/s00256-012-1509-z

Li, K., Tashman, S., Fu, F., Harner, C., and Zhang, X. (2010). Automating analyses of the distal femur articular geometry based on three-dimensional surface data. *Ann. Biomed. Eng.* 38, 2928–2936. doi: 10.1007/s10439-010-0064-9

Lyras, D. N., Loucks, C., and Greenhow, R. (2016). Analysis of the geometry of the distal femur and proximal tibia in the osteoarthritic knee: a 3D reconstruction ct scan based study of 449 cases. *Arch. Bone Joint Surg.* 4, 116–121. doi: 10.22038/abjs.2016.6436

Maillot, C., Leong, A., Harman, C., Morelli, A., Mospan, R., Cobb, J., and Riviére, C. (2019). Poor relationship between frontal tibiofemoral and trochlear anatomic parameters: implications for designing a trochlea for kinematic alignment. *Knee* 26, 106–114. doi: 10.1016/j.knee.2018.11.007

Mauler, F., Langguth, C., Schweizer, A., Vlachopoulos, L., Gass, T., Lüthi, M., et al. (2017). Prediction of normal bone anatomy for the planning of corrective osteotomies of malunited forearm bones using a three-dimensional statistical shape model. *J. Orthop. Res.* 35, 2630–2636. doi: 10.1002/jor.23576

Mutsvangwa, T., Burdin, V., Schwartz, C., and Roux, C. (2015). An automated statistical shape model developmental pipeline: application to the human scapula and humerus. *IEEE Trans. BioMed. Eng.* 62, 1098–1107. doi: 10.1109/TBME.2014.2368362

Myronenko, A., and Song, X. (2010). Point set registration: coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 2262–2275. doi: 10.1109/TPAMI.2010.46

Neogi, T., Bowes, M. A., Niu, J., De Souza, K. M., Vincent, G. R., Goggins, J., et al. (2013). Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative. *Arthr. Rheum.* 65, 2048–2058. doi: 10.1002/art.37987

Nolte, D., Tsang, C. K., Zhang, K. Y., Ding, Z., Kedgley, A. E., and Bull, A. M. J. (2016). Non-linear scaling of a musculoskeletal model of the lower limb using statistical shape models. *J. Biomech.* 49, 3576–3581. doi: 10.1016/j.jbiomech.2016.09.005

Pedoia, V., Lansdown, D. A., Zaid, M., McCulloch, C. E., Souza, R., Ma, C. B., et al. (2015). Three-dimensional mri-based statistical shape model and application to a cohort of knees with acute acl injury. *Osteoarthr. Cartil.* 23, 1695–1703. doi: 10.1016/j.joca.2015.05.027

Peloquin, J. M., Yoder, J. H., Jacobs, N. T., Moon, S. M., Wright, A. C., Vresilovic, E. J., et al. (2014). Human l3l4 intervertebral disc mean 3d shape, modes of variation, and their relationship to degeneration. *J. Biomech.* 47, 2452–2459. doi: 10.1016/j.jbiomech.2014.04.014

Rao, C., Fitzpatrick, C. K., Rullkoetter, P. J., Maletsky, L. P., Kim, R. H., and Laz, P. J. (2013). A statistical finite element model of the knee accounting for shape and alignment variability. *Med. Eng. Phys.* 35, 1450–1456. doi: 10.1016/j.medengphy.2013.03.021

Salenius, P., and Vankka, E. (1975). The development of the tibiofemoral angle in children. *J. Bone Joint Surg. Am.* 57, 259–261. doi: 10.2106/00004623-197557020-00021

Schatka, I., Weiler, A., Jung, T. M., Walter, T. C., and Gwinner, C. (2018). High tibial slope correlates with increased posterior tibial translation in healthy knees. *Knee Surg. Sports Traumatol. Arthrosc.* 26, 2697–2703. doi: 10.1007/s00167-017-4706-4

Smoger, L. M., Fitzpatrick, C. K., Clary, C. W., Cyr, A. J., Maletsky, L. P., Rullkoetter, P. J., et al. (2015). Statistical modeling to characterize relationships between knee anatomy and kinematics. *J. Orthop. Res.* 33, 1620–1630. doi: 10.1002/jor.22948

Subburaj, K., Ravi, B., and Agarwal, M. (2010). Computer-aided methods for assessing lower limb deformities in orthopaedic surgery planning. *Comput. Med. Imaging Graph.* 34, 277–288. doi: 10.1016/j.compmedimag.2009.11.003

Thienpont, E., Schwab, P.-E., Paternostre, F., and Koch, P. (2014). Rotational alignment of the distal femur: anthropometric measurements with CT-based patient-specific instruments planning show high variability of the

posterior condylar angle. *Knee Surg. Sports Traumatol. Arthrosc.* 22, 2995–3002. doi: 10.1007/s00167-014-3086-2

Van Dijck, C., Wirix-Speetjens, R., Jonkers, I., and Vander Sloten, J. (2018). Statistical shape model-based prediction of tibiofemoral cartilage. *Comput. Methods Biomech. Biomed. Eng.* 21, 1–11. doi: 10.1080/10255842.2018.1495711

Van Haver, A., De Roo, K., De Beule, M., Van Cauter, S., Audenaert, E., Claessens, T., et al. (2014). Semi-automated landmark-based 3D analysis reveals new morphometric characteristics in the trochlear dysplastic femur. *Knee Surg. Sports Traumatol. Arthrosc.* 22, 2698–2708. doi: 10.1007/s00167-013-2573-1

Wang, J., and Shi, C. (2017). Automatic construction of statistical shape models using deformable simplex meshes with vector field convolution energy. *Biomed. Eng. Online* 16:49. doi: 10.1186/s12938-017-0340-0

Wang, M. (2014). Generalized estimating equations in longitudinal data analysis: a review and recent developments. *Adv. Stat.* 2014:303728. doi: 10.1155/2014/303728

Youn, K., Park, M. S., and Lee, J. (2017). Iterative approach for 3D reconstruction of the femur from un-calibrated 2d radiographic images. *Med. Eng. Phys.* 50, 89–95. doi: 10.1016/j.medengphy.2017.08.016

Zhang, J., and Besier, T. F. (2017). Accuracy of femur reconstruction from sparse geometric data using a statistical shape model. *Comput. Methods Biomech. Biomed. Eng.* 20, 566–576. doi: 10.1080/10255842.2016.1263301

Zhang, J., Fernandez, J., Hislop-Jambrich, J., and Besier, T. F. (2016). Lower limb estimation from sparse landmarks using an articulated shape model. *J. Biomech.* 49, 3875–3881. doi: 10.1016/j.jbiomech.2016.10.021

Zhang, J., Malcolm, D., Hislop-Jambrich, J., Thomas, C. D. L., and Nielsen, P. M. (2014). An anatomical region-based statistical shape model of the human femur. *Comput. Methods Biomech. Biomed. Eng.* 2, 176–185. doi: 10.1080/21681163.2013.878668

Zheng, G., and Schumann, S. (2009). 3D reconstruction of a patient-specific surface model of the proximal femur from calibrated x-ray radiographs: a validation study. *Med. Phys.* 36, 1155–1166. doi: 10.1118/1.3089423

Zhu, Z., and Li, G. (2011). Construction of 3d human distal femoral surface models using a 3D statistical deformable model. *J. Biomech.* 44, 2362–2368. doi: 10.1016/j.jbiomech.2011.07.006

Exploring the Application of Pattern Recognition and Machine Learning for Identifying Movement Phenotypes During Deep Squat and Hurdle Step Movements

Sarah M. Remedios[1], Daniel P. Armstrong[1], Ryan B. Graham[2] and Steven L. Fischer[1]*

[1] *Occupational Biomechanics and Ergonomics Laboratory, Department of Kinesiology, University of Waterloo, Waterloo, ON, Canada,* [2] *Spine Biomechanics Laboratory, School of Human Kinetics, University of Ottawa, Ottawa, ON, Canada*

Background: Movement screens are increasingly used in sport and rehabilitation to evaluate movement competency. However, common screens are often evaluated using subjective visual detection of *a priori* prescribed discrete movement features (e.g., spine angle at maximum squat depth) and may not account for whole-body movement coordination, or associations between different discrete features.

Objective: To apply pattern recognition and machine learning techniques to identify whole-body movement pattern phenotypes during the performance of exemplar functional movement screening tasks; the deep squat and hurdle step. Additionally, we also aimed to compare how discrete kinematic measures, commonly used to score movement competency, differed between emergent groups identified via pattern recognition and machine learning.

Methods: Principal component analysis (PCA) was applied to 3-dimensional (3D) trajectory data from participant's deep squat (DS) and hurdle step performance, identifying emerging features that describe orthogonal modes of inter-trial variance in the data. A gaussian mixture model (GMM) was fit and used to cluster the principal component scores as an unsupervised machine learning approach to identify emergent movement phenotypes. Between group features were analyzed using a one-way ANOVA to determine if the objective classifications were significantly different from one another.

Results: Three clusters (i.e., phenotypes) emerged for the DS and right hurdle step (RHS) and 4 phenotypes emerged for the left hurdle step (LHS). Selected discrete points commonly used to score DS and hurdle step movements were different between emergent groups. In regard to the select discrete kinematic measures, 4 out of 5, 7 out of 7 and 4 out of 7, demonstrated a main effect ($p < 0.05$) between phenotypes for the DS, RHS, and LHS respectively.

Conclusion: Findings support that whole-body movement analysis, pattern recognition and machine learning techniques can objectively identify movement behavior phenotypes without the need to *a priori* prescribe movement features. However, we also highlight important considerations that can influence outcomes when using machine learning for this purpose.

INTRODUCTION

Movement screens are commonly used to assess an individual's quality of movement as a method to highlight poor movement patterns (McCunn et al., 2016). The quality of movement, herein termed movement competency, can be explained as an individual's ability to adopt a movement pattern that achieves the task objective, while also minimizing injury risk (Kritz et al., 2009; McGill et al., 2015). Considering the emphasis on movement and safety, sport (McCunn et al., 2016) and occupational settings (Isernhagen, 1992; Sinden et al., 2017) rely on movement screening methods to estimate performance capacity (Frost et al., 2015; Bock et al., 2016) and to reveal functional limitations that may increase risk of injury (O'Connor et al., 2011; Lisman et al., 2013). Visual assessment of body mechanics is the *de facto* method for measuring movement competency (Sinden et al., 2017), which increases the subjectivity of movement screens, thus relying on the appraisal and previous knowledge of the practitioner. As a result, it may not be surprising that inter-rater reliability issues continue to restrict the utility of movement screening approaches (i.e., Shultz et al., 2013).

In addition to inter-rater reliability challenges that affect subjective appraisal of movement competency, the current use of top-down, prescribed, discrete movement features to define "safe" or "good" movement may be inadequate. While many believe that movement competency is linked to injury risk and/or performance (where movement competency is defined using conventional *a priori* definition such as torso is parallel with the tibia when performing the deep squat), there remains little evidence supporting such connections (Gross and Battié, 2006; Mottram and Comerford, 2008; Schneiders et al., 2011; Okada et al., 2011; Parchmann and McBride, 2011). Perhaps our *a priori* criteria for subjectively evaluating movement competency are incorrect or incomplete (Bennett et al., 2017), or our clinical eye is simply not appropriately tuned to detect important and meaningful changes. As an alternative to this top-down approach, use of emerging tools in machine learning might help us to identify naturally-occurring movement phenotypes, where continued research can then explore phenotypes that are associated with positive or negative health outcomes with respect to specific task objectives.

Considering the magnitude of variability that exists in the ways individuals can complete a task, a reliance on discrete *a priori* measures, as common movement screen scoring

parameters (e.g., spine angle at maximum squat depth), instead of assessing time-series whole-body movement patterns remains as a limitation. Specifically, the use of *a priori* discrete parameters suggests that there is a single idealized pattern, which as shown by Srinivasan and Mathiassen (2012), is not necessarily optimal. Instead, it may be more beneficial to identify and screen for phenotypical patterns of movement behaviors that may differentiate and classify between those with optimal movement competency relative to those that may benefit from a targeted movement training intervention.

The Functional Movement Screen™ (FMS) (Functional Movement Systems, Chatham, VA, United States) remains a popular tool for movement screening (Bennett et al., 2017). Sinden et al. (2017) identified the FMS™ as one of the most commonly used approaches for movement screening among Kinesioligists. The FMS™ is an example of a movement screen that depends on the visual appraisal of discrete movement competency and identifies deficits and/or compensatory movement patterns in the kinetic chain (Cook et al., 2006a,b). While the FMS™ protocol includes a battery of 7 distinct movements, we focus on the Deep Squat (DS) and right and lift hurdle step (RHS; LHS) movements. Squatting is a common pattern in most athletic events (Cook et al., 2006a; Kritz et al., 2009) making it a useful movement to target first. The hurdle step movement, provides a unique contrast relative to the squat because it tests bilateral functional mobility and dynamic stability of the hips, knees, and ankles (Cook et al., 2006a). Many believe that such screening can be useful in proactive injury prevention (Kiesel et al., 2007). However, due to the lack of evidence relating the FMS™ to injury (McCunn et al., 2016), or biomechanical exposure variables in transfer tasks (Beach et al. 2014), evidence does not support that the current scoring approach is useful for injury prevention (Okada et al., 2011; Parchmann and McBride, 2011). This is not, however, to suggest that screening is not useful. Considering sound biomechanical arguments (Zazulak et al., 2008; Powers, 2010; Hewett and Myer, 2011), Beach et al. (2014), conclude that general whole-body movement screening could be used to predict likelihood of injury in physically demanding jobs if we advance beyond the current scoring approaches. Therefore, to overcome limitations associated with the subjective *a priori* driven grading criteria, data-driven methods could improve the state of movement screening (McCunn et al., 2016).

Application of pattern recognition and machine learning techniques are growing within biomechanics (Halilaj et al., 2018) and can enable data-driven methods to objectively identify movement phenotypes. As a pattern recognition tool, principal

---

Abbreviations: DS, Deep squat; GMM, Gaussian mixture model; LHS, Left hurdle step; PCA, Principal component analysis; PC, Principal component; RHS, Right hurdle step.

component analysis (PCA), allows us to identify principal movement patterns through data reduction, which explain variance within kinematic-based data sets (Troje, 2002; Wrigley et al., 2005; Brandon et al., 2013; Federolf et al., 2014; Ross et al., 2018; Armstrong et al., 2019). One strength of using PCA to determine modes of variability is that the scores can be used in downstream analysis such as in classification through cluster analysis to detect and interpret differences between subjects and/or trials (Deluzio et al., 2014, p. 319). As an example, clustering is an unsupervised machine learning method that iteratively clusters data points into groups assigning each observation to a cluster. In the biomechanical analysis of human movement data, clustering has proven useful for grouping participants with similar patterns (Sawacha et al., 2010; Bennetts et al., 2013; Gilles and Wild, 2018) and gait waveforms (Watelain et al., 2000; Toro et al., 2007; Roche et al., 2014). Previously, PCA and clustering techniques have been combined to identify and group distinct spine spatiotemporal movement strategies (Beaudette et al., 2019), which support that a combination of these methods may have utility in objectively identifying movement phenotypes in a movement screening context. However, such application of pattern recognition and clustering to identify naturally occurring movement phenotypes within the movement screening context remains a novel endeavor.

Therefore, to address issues related to the use of subjectively measured *a priori* movement competency features, the objective of this paper was to apply PCA and gaussian mixture model (GMM), as pattern recognition and machine learning techniques respectively, to objectively identify naturally occurring whole-body movement pattern phenotypes during the performance of common movement screening tasks (i.e., the deep squat and hurdle step). Secondarily, we aimed to evaluate if top-down *a priori* determined, discrete kinematic variables (typically evaluated in practice using a subjective visual appraisal), were indeed different between naturally emerging movement phenotype groups identified using unsupervised learning (i.e., bottom-up).

MATERIALS AND METHODS

Subjects

Thirty healthy participants volunteered for this study (**Table 1**). The participants were recruited from the general student body of the University of Waterloo, were older than 18 years old and did not have an injury that prevented activities of daily living in the previous 6 months. The participants completed a "Get Active Questionnaire" that indicated their physical readiness for the study. This study was approved by the University of Waterloo's Office of Research Ethics, and participants provided informed consent prior to participation.

TABLE 1 | Participant demographics.

	Age	Height (cm)	Weight (kg)
Male (n = 15)	23.6 ± 4.0	185.23 ± 6.8	87.9 ± 10.0
Female (n = 15)	23.7 ± 8.0	168.2 ± 9.8	64.3 ± 9.25

Instrumentation

Prior to coming to the lab, participants were instructed to wear tight fitting clothing. All participants were instrumented with reflective motion capture markers, including marker clusters placed over body segments and single markers positioned over anatomical landmarks (**Figure 1**). Marker clusters were used to track segment motion instead of anatomical markers to reduce soft tissue artifact (Leardini et al., 2005). Anatomical markers were used during calibration to mathematically relate the technical coordinate system of each cluster to its underlying segment specific anatomical coordinate system (Robertson et al., 2013). Motion was recorded using a 12 – camera (six, Vantage v5; six, Vero v2.2) Vicon Nexus 2.6 motion capture system (Nexus, Oxford, United Kingdom). Once participants completed a calibration trial, the following markers, bilaterally, were removed for the remainder of the study: lateral and medial epicondyles, iliac crest, anterior superior iliac spine, greater trochanter of femur, lateral and medial condyle, lateral and medial malleolus.

Protocol

Participants performed 15 movements in total for the study: 5 deep squat (DS), 5 right hurdle step (RHS), and 5 left hurdle step (LHS) movements (**Figure 2**) and were given instruction about how to perform each movement, adapted from those described in Cook et al. (2006a). Specifically, for the deep squat participants were instructed to: "stand with your feet approximately shoulder width apart, place the dowel on your head adjusting your hands
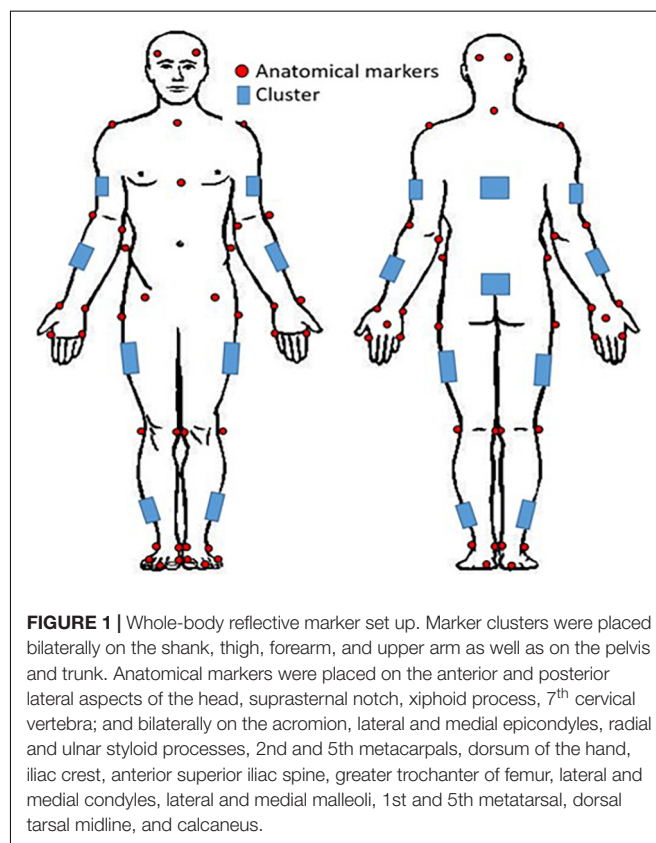
FIGURE 1 | Whole-body reflective marker set up. Marker clusters were placed bilaterally on the shank, thigh, forearm, and upper arm as well as on the pelvis and trunk. Anatomical markers were placed on the anterior and posterior lateral aspects of the head, suprasternal notch, xiphoid process, 7th cervical vertebra; and bilaterally on the acromion, lateral and medial epicondyles, radial and ulnar styloid processes, 2nd and 5th metacarpals, dorsum of the hand, iliac crest, anterior superior iliac spine, greater trochanter of femur, lateral and medial condyles, lateral and medial malleoli, 1st and 5th metatarsal, dorsal tarsal midline, and calcaneus.

FIGURE 2 | Deep squat movement and hurdle step movement as adapted from Cook et al. (2006a).

until your elbows are at 90 degrees, press the dowel overhead, straightening the elbows, the trial will begin once you descend into a deep squat position and back up keeping your heels on the floor and arms extended the entire time." For the hurdle step, instructions were: "stand facing the front of the lab with your toes touching the FMS board, place the dowel across the back of your shoulders and below your neck, with your right/left leg, step over the hurdle, touch your heel on the opposite side and bring your moving leg back to the starting position." Participants completed 5 repetitions of the DS, followed by 5 repetitions of the RHS and 5 of the LHS. Motion data were collected at 60 Hz using Vicon Nexus while participants performed the DS, RHS and LHS movements, respectively.

Data Post-processing and Conditioning

During post-processing, each trial was first labeled and gap filled in Vicon Nexus, where gaps were filled using cubic spline, pattern fill or rigid body fill functions within Vicon Nexus, where the gap filling function was dependent on the underlying gap length (Armstrong et al., 2019). Gap-filled and labeled marker trajectory data were exported to Visual 3D (C-Motion Inc., Germantown, United States). Within Visual 3D, data were filtered using a fourth order low pass butterworth filter with an effective cut off frequency of 6 Hz (Winter, 2009) to remove high frequency noise from each signal. Filtered trajectory data were then used to drive a 15 segment whole-body kinematic model, with IK constraints, where segments were defined using ISB recommended segment definitions (Wu et al., 2002, 2005), such that joint center positions (ankle, knee, hip, shoulder, elbow, wrist) and centre of mass (COM) locations (pelvis, trunk, and head) could be calculated. Joint center and COM trajectory data were combined with filtered position data from selected body landmarks (xiphoid process, suprasternal notch, 7th cervical vertebra) to provide the kinematic description of each motion.

Prior to additional data processing, start and end frames for each trial were determined (**Figure 3**). The DS "start" and "end" were defined by identifying the local maximum of the supra-sternal notch marker in the vertical direction. The "start" and "end" of the hurdle step were determined by identifying the local minimum of the lead (step-over) heel marker in the vertical direction.

To address our overarching research question, we also used the trajectory data to calculate discrete kinematic variables

commonly used in screening. **Tables 2**, **3** list the additional discrete kinematic data that were calculated and also summarizes how they were calculated to support this analysis.

To support the use of pattern recognition and machine learning, trajectory data representing the above mentioned joint centers, landmarks, and COM locations were exported to Matlab (MathWorks, Natick, United States). In Matlab, participants' trajectory data were divided by their standing height to normalize for inter-participant variance in height (Ross et al., 2018). The trajectory data were also translated such that the new origin was positioned at the center of the right (DS, LHS) or left (RHS) ankle coordinate system. This translation was necessary to eliminate variance in the trajectory data associated with each participants' relative positioning with the global coordinate system of the laboratory.

Trials were time normalized to 101 frames (100% of the trial), but in phases in order to account for the fact that participants
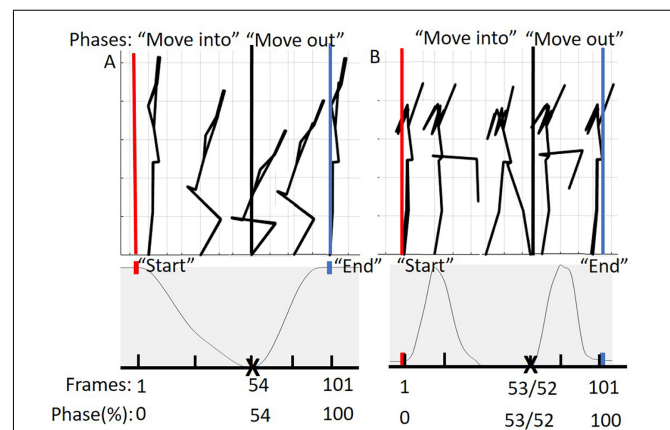
FIGURE 3 | (A) Represents the phases of the DS movement from 0 to 100%. The "move into" phase is contained by the red and black bars and in between the black and blue bars is the "move out" phase. The below graph represents tracking of the suprasternal notch marker to determine 50% of the movement (maximum squat depth) as well as define "start" and "end" points of the movement. **(B)** Represents the phases of the hurdle step movement from 0 to 100%. The "move into" phase is contained by the red and black bars and in between the black and blue bars is the "move out" phase. The graph below represents tracking of the right (LHS) or left (RHS) calcaneus marker to determine 50% of the movement (end of heel touch) as well as define "start" and "end" points of the RHS and LHS, respectively.

TABLE 2 | Calculated kinematic variables typically representative of the *a priori,* discrete DS scoring criteria.

Scoring criteria (Cook et al., 2006a)	Kinematic representation	Calculation	Meaning/Relevance
The femur is at or below horizontal	Femoral angle	The right and left femoral segment angles were calculated, averaged, and then the maximum angle was found.	Representation of squat depth
The trunk is parallel with the tibia and/or toward the vertical	Trunk- Shank angle difference Trunk angle	The maximum trunk to shank angle difference was determined as well as the maximum trunk angle elicited.	The greater the difference the more the tibia and trunk are not parallel. The greater the trunk angle, the further bent forward.
The knees remain aligned with the feet	Knee-to-ankle separation ratio (Ortiz et al., 2016)	The difference between left and right knee joint centers in global space and the difference between the left and right ankle joint centers in global space were calculated. The knee distance was divided by the ankle distance.	(>1) = knees are wider (varus) (<1) = knees fall inwards (valgus)
The dowel remains over top of the feet	Left-right hand center difference to foot center difference.	The center of the left- and right-hands in global space and left and right foot center in global space were calculated. The maximum anterior difference between the hand center and foot center was calculated.	Measure of displacement of dowel over the feet in the anterior direction.

TABLE 3 | Calculated kinematic variables typically representative of the *a priori*, discrete hurdle step scoring criteria.

Scoring criteria (Cook et al., 2006a)	Kinematic representation	Calculation	Explanation
Hips, knees, and ankles remain aligned	Hip-to-Knee difference Knee-to-ankle difference Hip-to-ankle difference Hurdling leg was the leg of interest	= (hip – knee) = (knee – ankle) = (hip-ankle) All calculations used the joint center in the y-axis (anterior plane) Peak absolute values were calculated	A difference value closer to 0, the more in-line the joint centers. The hurdle step scoring criteria #1, was determined kinematically by gathering the difference between all three joints in the anterior plane.
Little to no movement noted in lumbar spine	Lumbar Flexion – extension range Lumbar lateral flexion range Rotation range	= (maximum extension – maximum flexion) = (maximum right lateral flexion – maximum left lateral flexion) = (maximum rotation to the right – maximum rotation to the left)	To determine little movement in lumbar spine, angle ranges in all three directions were calculated. The greater the range, the more movement noted in lumbar spine.
The dowel remains parallel with the string	Right-left hand difference in the z direction (superior/inferior)	= (right hand center – left hand center)	Represented by the difference in hand displacement in the superior/inferior direction. The greater the difference of the two hands, the greater the dowel is not parallel with the string.

typically took longer to move into the required position, but less time to move out of the required position. As an example, participants for the DS generally took different lengths of time descending into maximum squat depth and returning to upright standing. To achieve our desired phase-based time-warping, first, the "move into" portion of the movement (i.e., from standing to maximum squat depth, or foot extended over hurdle) was segmented out and time normalized to 54 frames for the DS, 53 for the RHS and 52 for the LHS, respectively. Second, the "move out" portion of the movement (i.e., return to standing) was segmented out and time normalized to 47 frames for the DS, 48 for the RHS and 49 for the LHS respectively. Third, the time normalized phases were re-concatenated into a complete trial (101 frames). On average, participants tended to complete the "move into" phase of the DS, RHS, and LHS at 54, 53, and 52% of the total movement time, reinforcing the splits noted above. This process was completed to eliminate timing effects or phase shift between trials of each of the movements respectively (Moudy et al., 2018). The time normalized estimated joint centers, body landmarks, and calculated COM positions were then prepped for PCA analysis in Matlab (**Figure 3**).

Data Analysis
Feature Selection
PCA was applied to the time-series conditioned and post-processed trajectory data to identify emergent features that captured orthogonal modes of variability in the data set. Individual PCA models were developed for the DS, RHS, and LHS data, respectively, using the 'Statistics and Machine Learning' toolbox in Matlab. Described more completely in Ross et al. (2018), but briefly summarized here, we organized the time-series trajectory data into a [n, m] matrix, where, n represented the number of trials (*n* = 150, corresponding to 30 participants × 5 trials) and where m represented row vectors describing

the time-series trajectory data ($m = 5454$, corresponding to 18 trajectories × 3 axes × 101 time points). PCA was then applied as a data reduction and feature selection method to yield principal components (PCs) that capture linearly uncorrelated sources of variability within each dataset. The application of PCA in this manner, for the purpose of identifying principal movements (PCs representing linearly uncorrelated movement features) is more completely described by Troje (2002) and Ross et al. (2018). PC scores were retained, representing each observation (trial) in the principal component space. PCs that individually explained > 5% of the variance (Witte et al., 2010), were retained for classification.

Classification

A [p, q] matrix was input into a GMM, where p represents each trial's PC scores as input features ($p = 150$, corresponding to 30 participants × 5 trials) and each column of q described individuals' PC scores for those PCs that were retained. As a brief background, GMM is a model-based method where the algorithm is aimed at optimizing the fit between the data and the model to find structures (clusters) among the observations, while also assigning a measure of probability to the clustered assignment. GMM was applied to the data in Matlab using the "Statistics and Machine Learning" toolbox. To determine the optimal k (number of clusters), we used the Bayesian information criterion (BIC), where for $k = 1–10$ a GMM was fit to the dataset and the minimum BIC identified the best k. An optimal k was determined for each movement: DS, RHS and LHS, respectively. A GMM for each movement was applied to each data set respectively, running 100 repetitions to increase the likelihood of the data converging to an optimum (Beaudette et al., 2019). Following the application of the GMM to each movement dataset, centroid scores from each cluster were determined along with the clustering assignments from each individual trial, where hard clustering was performed such that each trial was assigned to only 1 phenotype. The cluster centroids therefore represent the mean movement phenotypes.

Reconstruction

Single component reconstruction was used to visualize differences in movement patterns between clusters (Brandon et al., 2013). This reconstruction was done by multiplying the loading vectors for each retained PC by the centroid scores representing each cluster and adding it to the mean loading vector (eigenvectors from the PCA models). The reconstructed patterns provided a visual representation that emphasizes differences in the underlying kinematics associated with each movement phenotype.

Statistics

Kinematic variables typically representative of the *a priori*, discrete scoring criteria (**Tables 2**, **3**) served as dependent variables in one-way ANOVA models. Cluster assignment served as the independent variable (3 levels for DS and RHS movements and 4 for LHS movement, based on the emergence of 3 and 4 clusters, respectively). An alpha value of 0.05 was used to determine significance. Where a main effect of cluster assignment emerged, *post hoc* testing, using Bonferonni corrected pairwise comparisons were used to determine significant differences in dependent measures between clusters. Partial eta squared values (η^2) were calculated for each dependent variable where, 0.01 was considered a small effect, 0.06 a medium effect and 0.14 a large effect (Cohen, 1988, p. 285–287, 383). Statistical analysis was completed using SPSS (SPSS Version 24.0, IBM Corporations, Armonk, NY, United States).

RESULTS

Feature Selection and Classification

The PCA models revealed that 4, 6, and 6 PCs each explained at least 5% of the variance in the time-series trajectory data for the DS, RHS, and LHS, respectively. Using those retained PCs, the GMM identified $k = 3$ as the optimal number for the DS and RHS and $k = 4$ for the LHS movements (**Figure 4**). For the DS, 62, 24, and 64 trials were assigned to phenotypes 1, 2, and 3, respectively, where 23 of the participants had all 5 trials classified within the same phenotype and 7 participants had trials distributed between 2 different phenotypes. For the RHS, 47, 84, and 19 trials were assigned to phenotypes 1, 2, and 3, respectively, where 21 of the participants had all 5 trials classified into the same phenotype, and 9 participants had trials distributed between different phenotypes. Lastly, for the LHS, 36, 50, 25, and 39 trials were assigned in phenotypes 1, 2, 3, and 4 respectively, where 23 participants had all 5 trials classified within the same phenotype, and 7 participants had trials distributed between different phenotypes. It is interesting to note the disproportionate clustering, where many trials were assigned to cluster 2 for the RHS, but fewer to clusters 1 or 3, as an example. It is important to note than when interpreting the data, each movement was analyzed separately, for example, we cannot claim that phenotype 1 for the RHS and phenotype 1 for the LHS are related.

Single Component Reconstruction

The results of the single component reconstructions are in **Figures 5**–**7** and in **Supplementary Material**. The purpose
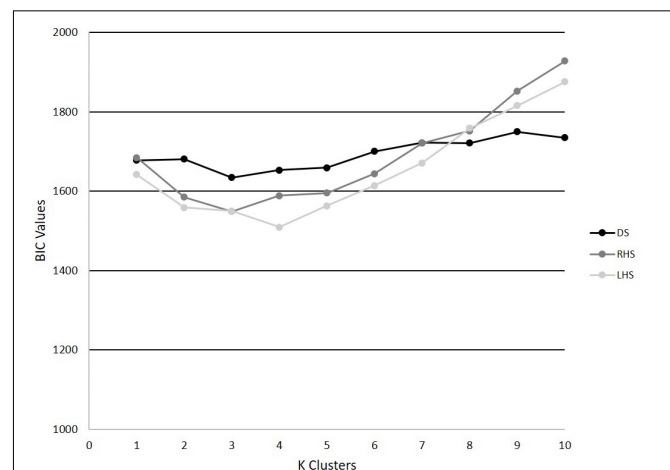
FIGURE 4 | BIC values for $k = 1–10$ for the DS, RHS, and LHS demonstrating minimum values at $k = 3$ for the DS and RHS, and $k = 4$ for the LHS.

of the single component reconstructions is to provide a visual representation of the emergent differences in movement phenotypes (cluster centroids).

Kinematic Differences Between Clusters
Deep Squat
A main effect of cluster assignment was detected for 4 of the 5 kinematic measures typically used to subjectively evaluate the DS (**Table 4** and **Figure 8**). The trunk segment angle, a measure to represent forward lean of the torso, was not different between the clusters. *Post hoc* pairwise comparisons revealed that the femoral angle (**Figure 8A**), was different between phenotypes 1 and 2 and phenotypes 1 and 3, but that phenotypes 2 and 3 were not different. Considering the trunk-shank angle difference measure (**Figure 8B**) phenotypes 1 and 2 were different, but phenotype 3 was not different from either 1 or 2. The sagittal plane dowel alignment, was different between phenotypes 1 and 3, but phenotype 2 was not different from either 1 or 3 (**Figure 8C**). Lastly, the knee-ankle separation ratio, a measure aimed to represent knee varus/valgus, showed differences between phenotypes 1 and 3 only (**Figure 8D**).

Right Hurdle Step
A main effect of cluster assignment was detected for all 7 kinematic measures commonly used to score the RHS, implying these variables soundly represent variance in the movement of the RHS (**Table 5**). *Post hoc* pairwise comparisons revealed that all three hip-knee-ankle frontal plane alignment variables were significant between phenotypes 2 and 3, and where the

ankle-hip alignment variable was different between phenotypes 1 and 3 (**Figure 9A**). Considering measures associated with lumbar spine control, *post hoc* pairwise comparisons revealed differences between phenotypes 2 and 3 in the range of motion about all three axes (**Figure 9B**). There were further differences between phenotypes 1 and 3 for both the lumbar range of motion associated with a lateral bend and rotation. Phenotypes 1 and 2 only differed for the range of lumbar movement represented by the flexion/extension axis. Lastly, consider the hands/dowel parallel to the string measure, there were significant difference between phenotypes 1 and 3, and 2 and 3 (**Figure 9C**).

Left Hurdle Step
A main effect of cluster assignment was detected for 4 of the 7 kinematic measures commonly used to score LHS, including hip-ankle alignment, lumbar flexion/extension range, lumbar lateral flexion range, and hands/dowel parallel with the string measure. No main effects were detected for hip-knee and knee-ankle alignment difference and lumbar rotation range (**Table 5**). *Post hoc* pairwise comparisons revealed that hip-ankle alignment measures were different between phenotypes 2 and 4 only (**Figure 10A**). Further, flexion-extension range were different between phenotypes 1 and 2 as well as 2 and 4. Considering lumbar movement regarding lateral flexion, phenotype 4 differed statistically from all other phenotypes (**Figure 10B**). Lastly, all phenotypes for the hands/dowel parallel with the string measure differed statistically except phenotype 4 with 1 and 3 (**Figure 10C**).
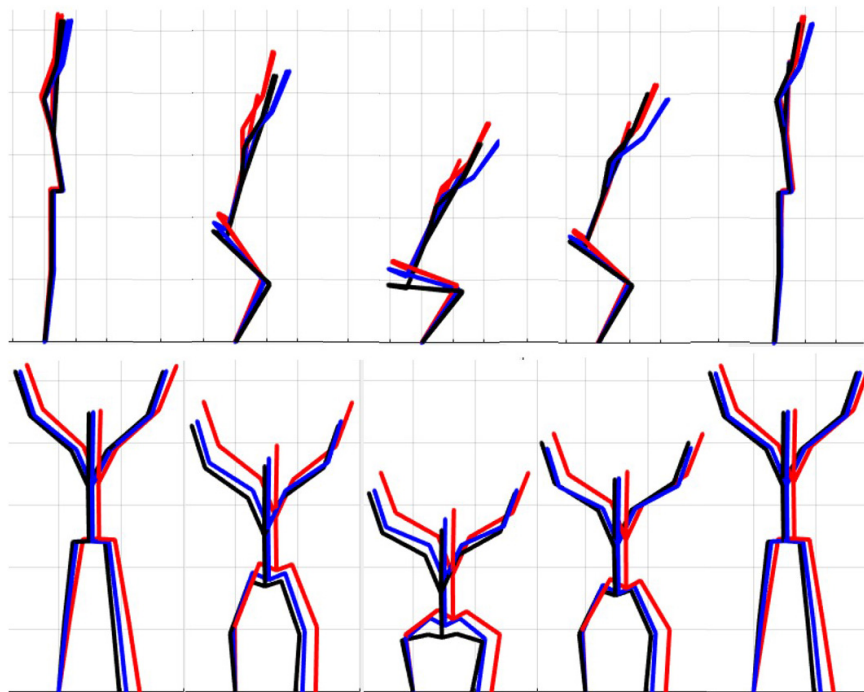
FIGURE 5 | Reconstructed movement phenotypes using the centroid PC scores from each cluster considering the deep squat movement. Black, movement phenotype 1; red, movement phenotype 2; blue, movement phenotype 3.
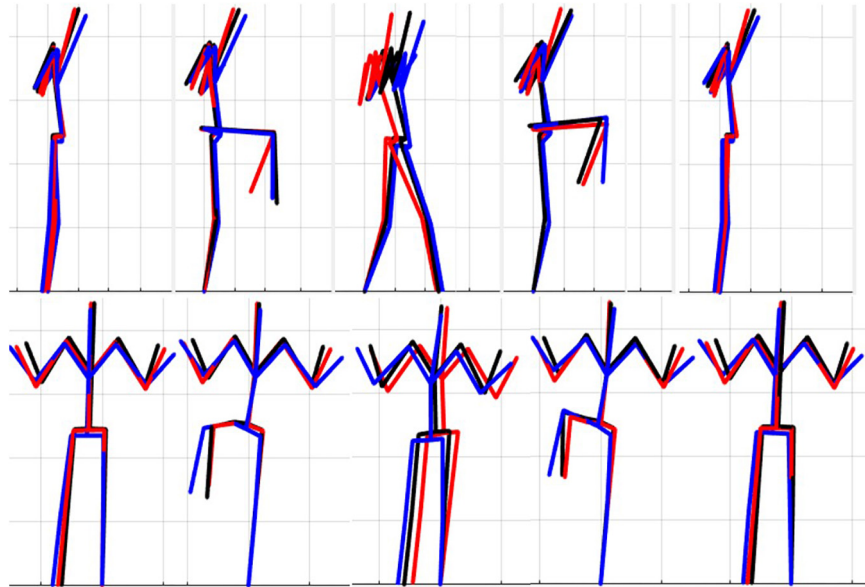
FIGURE 6 | Reconstructed movement phenotypes using the centroid PC scores from each cluster considering the right hurdle step movement phenotypes identified. Black, movement phenotype 1; red, movement phenotype 2; blue, movement phenotype 3.
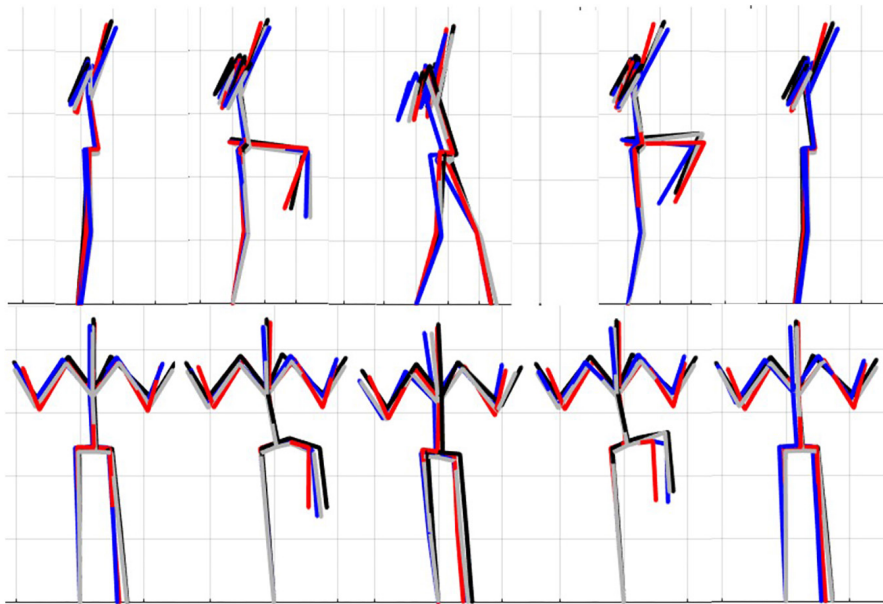
FIGURE 7 | Reconstructed movement phenotypes using the centroid PC scores from each cluster considering the left hurdle step movement phenotypes identified. Black, movement phenotype 1; red, movement phenotype 2; blue, movement phenotype 3; gray, movement phenotype 4.

DISCUSSION

The objective of this study was to identify naturally occurring whole body movement pattern phenotypes related to the performance of exemplar movement screening tasks including the DS, RHS and LHS by using pattern recognition (PCA) and machine learning techniques (GMM). Further, we aimed to contrast how kinematic measures commonly used to score these movements differed between the groups. The statistical results demonstrate that while all kinematic measures commonly used to score the RHS demonstrated a main effect, none of the kinematic measures independently were actually different between all emergent phenotypes (**Figure 9**). Instead, data demonstrate that kinematic measures commonly used score these screening movements are often different between pairs of phenotypes, but not necessarily able to independently distinguish

TABLE 4 | Level of significance (*p*-value) and effect size ($\eta^{2=}$ partial eta squared value) results for the kinematic variable representation of the DS scoring criteria.

	The thigh is at or below horizontal	The trunk remains upright and/or remains parallel with the tibia		The Dowel remains aligned over feet	The knees remain aligned with the ankle
	Thigh segment angle	Trunk segment angle	Trunk shank angle difference	Hand center – foot center difference	Knee ankle separation ratio
Deep squat	$p = 0.000^*$ $\eta^2 = 0.141$	$p = 0.140$ $\eta^2 = 0.026$	$p = 0.011^*$ $\eta^2 = 0.059$	$p = 0.000^*$ $\eta^2 = 0.143$	$p = 0.009^*$ $\eta^2 = 0.063$

*Denotes a main effect at the 0.05 level.

FIGURE 8 | Violin plot (Holger Hoffmann, 2020) demonstrating the distribution shape of each phenotype for kinematic measures commonly used to score the DS. The mean is represented by the white dotted line and median with the solid white line. **(A)** The femur is at or below horizontal; **(B)** the torso remains upright and/or is parallel with the tibia; **(C)** the dowel remains aligned over the feet; **(D)** the knees remain aligned with the feet. *The mean difference is significant at the 0.05 level.

TABLE 5 | Level of significance (*p*-value) and effect size (η^2 = partial eta squared value) results for the kinematic variable representation of the hurdle step scoring criteria.

	The hips, knees, ankle remain aligned			There is little to no movement in the lumbar spine			Dowel parallel with string
	Hip-knee	Knee-ankle	Hip-ankle	Flex/Ext range	Lat flex range	Rotation range	Hands parallel
Right hurdle step	$p = 0.003^*$	$p = 0.026^*$	$p = 0.000^*$	$p = 0.000^*$	$p = 0.000^*$	$p = 0.004^*$	$p = 0.000^*$
	$\eta^2 = 0.077$	$\eta^2 = 0.048$	$\eta^2 = 0.120$	$\eta^2 = 0.139$	$\eta^2 = 0.127$	$\eta^2 = 0.074$	$\eta^2 = 0.258$
Left hurdle step	$p = 0.238$	$p = 0.418$	$p = 0.008^*$	$p = 0.001^*$	$p = 0.000^*$	$p = 0.932$	$p = 0.000^*$
	$\eta^2 = 0.028$	$\eta^2 = 0.019$	$\eta^2 = 0.078$	$\eta^2 = 0.112$	$\eta^2 = 0.225$	$\eta^2 = 0.003$	$\eta^2 = 0.284$

Denotes a main effect at the 0.05 level.

across all groups (**Figures 8**–**10**). This suggests that, if screening continues to be performed using visual appraisal of selected kinematics markers, a hierarchical-based decision tree approach (i.e., first classify on marker A, then sub-classify within those groups based on marker B, etc.) is likely to improve the ability to truly classify and distinguish between groups. Alternatively, contrary to the common practice of analyzing movements based on *a priori* prescribed features, objective data driven analysis can identify and cluster relevant movement phenotypes while considering the system as a whole. Using data driven methods to determine common patterns of movement tasks can reduce the need to use subjective visual appraisal based on *a priori* prescribed features. However, it is also important to note that while this study has generated insights about emergent movement phenotypes during the performance of the DS and hurdle step as exemplar movement screening tasks, the pattern recognition and machine learning techniques applied in this study cannot be applied blindly.

Independent evaluation of discrete kinematics features may not adequately distinguish and separate movement phenotypes. This is particularly evident when analyzing how the phenotypes relate back to the FMS[TM] scoring criteria, as the means of some scoring criteria were significantly different between phenotypes and others were not. This demonstrates that kinematic measures commonly used to score movements likely shouldn't be considered independently, and that the interaction between multiple variables might be more revealing. For example, the results for the DS demonstrate significant differences between groups 1 and 2 for two different features (thigh angle and trunk-shank angle difference) and 1 and 3 for three different features (thigh angle, dowel alignment measure and knee-ankle separation ratio). This suggests that the thigh angle or dowel alignment measure could be used to first extract phenotype 1 DS movements (i.e., exceed an appropriate DS thigh angle or dowel alignment threshold), and also highlights a potential redundancy in the DS scoring criteria. However, there were no significant differences present between phenotypes 2 and 3 in any of the measures, which suggests that there are other features that differentiate these groups. It also reveals an important limitation regarding the use of *a priori* prescribed kinematic measure, in that a top-down *a priori* assignment of variables may not actually coincide with kinematic outcomes that do indeed differentiate between groups (Bennett et al., 2017). In this case a sub-ordinate criterion is necessary to distinguish between the movements that remain after screening out phenotype 1 movements by using

the DS thigh angle or dowel alignment measure. Highlighting the strengths of pattern recognition and machine learning, these techniques can be applied "bottom-up" to probe for other measures that might better distinguish between DS phenotype 2 and 3. Using our single component reconstructions as a starting point, the frontal view visual representation of the DS phenotypes (**Figure 5**) shows that foot width was different between phenotypes 1 and 2, and 2 and 3 and has the potential to be a sub-ordinate measure to distinguish between phenotypes. While participants are instructed to place their feet approximately shoulder width apart, some individuals may place their feet wider due to limitations (e.g., morphological). Although this is not a factor that the FMS considers, the use of a wider base may be a result of some functional or behavioral differences that clinicians can explore.

Interpreting results from the LHS and RHS also reveal important evidence underscoring limitations in the use of "top-down" discrete *a priori* measures to score screening movements and further reinforcing the utility of pattern recognition and clustering approaches as "bottom-up" strategies to identify movement phenotypes. In comparison to the FMS[TM] based kinematic measures used to score the RHS, the clustering revealed that not all features were significant between groups and may be differentially affected in the synergistic control of movement features, further reinforcing the potential of a hierarchical-based approach to screening. For example, the hip-knee, knee-ankle, and hip-ankle alignments were all statistically different between phenotypes 2 and 3, indicating redundancy in the ability of these measures to differentiate, but also in the ability of any one of these measures to be a useful initial measure. However, only the hip-ankle alignment was statistically different between phenotypes 1 and 3, suggesting that this measure could be useful as a sub-ordinate to further refine grouping assignments. Similarly, when considering the lumbar motion related criteria, range of motion about each axis was different between phenotypes 2 and 3, emphasizing redundancy. Nevertheless, the results suggest the ability to differentiate between phenotypes 1 and 2 based on lumbar flexion extension range, and groups 1 and 3 based on the lumbar lateral flexion and lumbar rotation ranges. Like the DS, the use of a bottom-up, data-driven approach has also revealed a factor that might be important, but that is not currently considered: anterior-posterior centre of mass (COM) range of motion (**Figure 6**). Phenotype 3 (RHS), as an example, elicited a larger range of motion of the COM in the anterior direction, which could not be
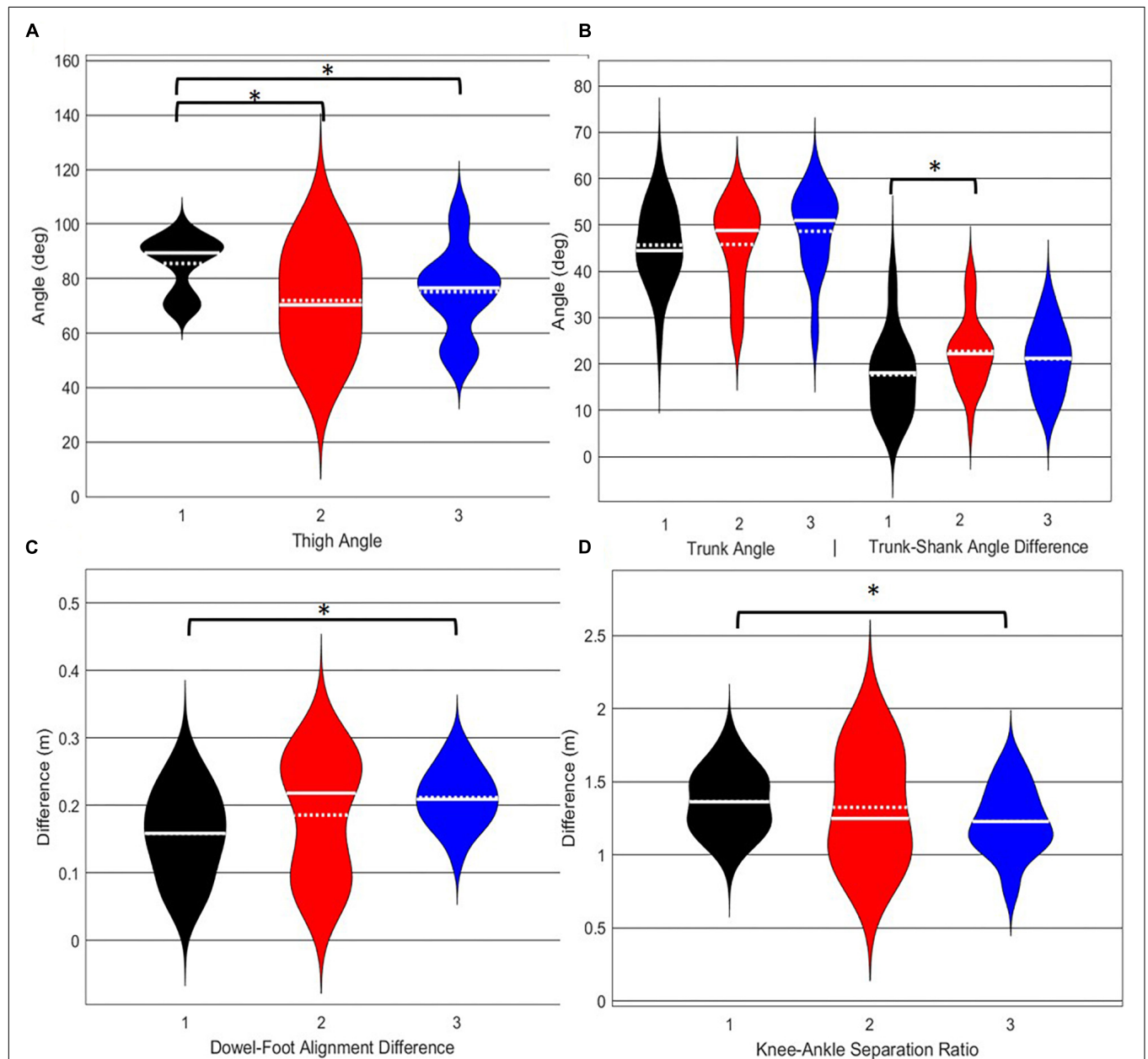
FIGURE 9 | Violin plot (Holger Hoffmann, 2020) demonstrating the distribution shape of each phenotype for kinematic measures commonly used to score the RHS. The mean is represented by the white dotted line and median with the solid white line. **(A)** Hip, knee, and ankle remain aligned; **(B)** there is little to no movement in the lumbar spine; **(C)** the hands/dowel remains parallel to the string. *The mean difference is significant at the 0.05 level.

explained by lumbar angle range of motion, suggesting a possible necessity as an additional movement assessment consideration.

For the LHS, the hip-ankle alignment was the only feature that showed differences between groups for the frontal plane hip-knee-ankle alignment measures. The results demonstrated that phenotype 4 can be differentiated from the other phenotypes based on the lateral lumbar flexion range, but would need subsequent analysis to differentiate further. Whereas, the feature separating phenotypes 1 and 2 and 2 and 4 were due to lumbar lateral flexion range, again supporting a hierarchical approach to screening in the absence of direct data-driven methods. Further,

the dowel/hands parallel to string measure elicited differences for phenotype 1 from 2 and 3, as well as phenotype 2 from 3 and 4, thus demonstrating that the kinematic representation of keeping the dowel parallel to the string of FMS^{TM} scoring criteria is a useful tool for differentiating differences in the hierarchy for the LHS. Considering that there were few differences between the frontal plane hurdling leg alignment, perhaps this is a feature that does not demonstrate as much variance as the RHS. Moreover, at this point in the analysis we are not able to explicitly state why the number of optimal clusters differs between the LHS and RHS, but speculate that there may be more variability in the LHS compared

FIGURE 10 | Violin plot (Holger Hoffmann, 2020) demonstrating the distribution shape of each phenotype for kinematic measures commonly used to score the LHS. The mean is represented by the white dotted line and median with the solid white line. **(A)** Hip, knee, and ankle remain aligned; **(B)** there is little to no movement in the lumbar spine. **(C)** The hands/dowel remains parallel to the string. *The mean difference is significant at the 0.05 level.

to the RHS possibly due to foot dominance. Unfortunately, we did not record foot dominance so we cannot further verify this speculation. Further analysis would be needed to identify the specific kinematic features that further aid to differentiate in the hypothesized hierarchical approach.

This study applied PCA and GMM to a dataset of DS, RHS, and LHS movements as performed by healthy individuals. As a result, the grouping assignments and underlying kinematic

difference will likely be different among samples, or perhaps even a larger sample, although the FMS[TM] has a target population of healthy, active individuals within the general population (Bennett et al., 2017), so our sample may be representative. However, with access to such a larger, representative dataset, this paper provides evidence to support and inform how motion capture, pattern recognition and machine learning can advance movement screening approaches. But, it is also important

discuss the assumptions and challenges that emerge when deploying this approach.

One challenge that emerged earlier on in the process of using PCA to identify principal movements was the determination of how many principal movements (PCs) to retain. In this study, PCs that individually explain >5% variance were kept and retained for analysis (Witte et al., 2010), since this method elicited the least number of PCs. However, other common strategies for PC retention include: PCs retained until a trace criterion of 90% of the total variance was retained (Deluzio and Astephen, 2007; Deluzio et al., 2014, p. 322) and PCs retained until a trace criterion of 95% of the total variance was retained (Deluzio et al., 2014, p. 322). However, in other applications, such as optimizing the prediction of a dependent variable, retaining PCs that individually explain >5% variability may not be sufficient (Richter et al., 2014). While retaining a greater number of PCs will include more of the variance within the dataset, when working with clustering, reducing the dimensionality is an important consideration, reinforcing our selection of a PC retention strategy that balance the variance explained with the number of PCs retained.

We chose a GMM as our clustering approach although other types of clustering algorithms may be considered. A GMM was chosen for its advantages of being a distribution-based model. GMM is a soft clustering method based on how probable it is that all data points in the cluster belong to the same distribution. This is contradictory to a centroid-based model with hard clustering (i.e., k-means clustering), where the notion of clustering is based on how close each data point is to the centroid and are assigned to a cluster without considering its variance. While this distinction may not be critical for this paper, it has important implications when clustering for the purpose of movement screening. Considering human variability (within and between), philosophically, it is unlikely that any one individual will absolutely cluster the same way every time. Instead, movements are likely to look more or less like a representative cluster (mean movement), where the GMM can provide an estimate of that likelihood. Such likelihood estimates may inform a hierarchical assessment approach, whereas a mover could be considered not just on the clustered assignment, but also on their likelihood weighting with regards to their assignment to each cluster. This is a concept that requires further contemplation and investigation.

Selecting the optimal number of clusters is also an important consideration, particularly when aiming to quickly screen a wide population of movers, such that they can be appropriately triaged (i.e., identify movers that require targeted training to improve). With the distribution, soft clustering-based method, GMM, the clusters can represent different ellipsoid shapes, overlap or be relatively close to one another which can skew results determined by a method such as a silhouette analysis. Silhouette analysis measures the separability of the clusters based on how close each point in one cluster is to points in the neighboring clusters (Beaudette et al., 2019). As an alternative, the BIC is a criterion for model selection among a finite set of models partly based on the likelihood function. The lower the BIC, the better the model to predict using the

data, this model avoids overfitting by penalizing models with big number of clusters (Bishop, 2006, p. 217). Although this may be interpreted as a drawback, if we want to be able to generalize our phenotypes for the purpose of rapid screening or movement-based triage, it is better to penalize large number of clusters. However, if the intent was to support a more personalized diagnoses, an alternate interpretation of the BIC may be required.

Limitations

Limitations related to the sample size, kinematic trajectories chosen to represent whole-body motion, kinematic variables chosen to represent the scoring criteria and decisions required to apply PCA and GMM likely all have some influence on the results and interpretation of these data. Specific to sample size, the sample size for this study was originally intended for a different research question. However, given that we did not know how many clusters would emerge, we were challenged upon determining the *a priori* sample size. We hope moving forward this study will assist in determining *a priori* sample size. While remaining limitations have been discussed within the main body of the paper, this method nevertheless does show that objective whole-body evaluation can identify phenotypes within a data set. With further research, this method may prove useful and promising in eliminating the subjective assessment of movement screens and improving interrater reliability, or at a minimum, informing on a hierarchy of distinguishable measures that can be used to differentiate movements. It is also important to note that future studies should consider adding their classification code to enable other researchers to use their methods. At this point we are not able to differentiate "good" versus "bad" movers, although our group continues to explore this possibility (Armstrong et al., 2019).

CONCLUSION

Overall, pattern recognition and machine learning techniques were able to objectively identify phenotypes within a group of individuals performing the DS, RHS and LHS. Further, when comparing kinematic measures commonly used to score movement between the different phenotype groups, some criteria were indeed different and others were not. In most cases, independent kinematic measures were not able to distinguish between all three/four different emergent phenotypes, and several measures overlapped in their ability to differentiate between phenotype groups. In the absence of objective, data-driven movement assessment, our results suggest that visual-based screening can likely be improved by reducing the number of measures to consider by eliminating independent measures that provide redundant information (i.e., measures that are likely coordinated in their control), and by considering measures using a hierarchical approach (i.e., screen based on measure A, then screen emergent groups based on sub-ordinate measures as necessary). Objective data analysis using whole body movement patterns gives insight into features of the DS and hurdle step that

may not be elicited through *a priori* feature selection analysis. Therefore, the results from this study provide important findings to the field that open up a number of future study directions, such as identifying which movement strategy could elicit different injury risk factors to advance injury prediction. Moving toward such objective data driven analysis may further enhance the ability to apply movement screening for the purpose of injury risk identification and mitigation.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was reviewed and approved the University of Waterloo Office of Research Ethics. Participants for this study provided informed written consent prior to participation.

AUTHOR CONTRIBUTIONS

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00364/full#supplementary-material

REFERENCES

Armstrong, D. P., Ross, G. B., Graham, R. B., and Fischer, S. L. (2019). Considering movement competency within physical employment standards. *Work* 63, 603–613. doi: 10.3233/WOR-192955

Beach, T. A. C., Frost, D. M., and Callaghan, J. P. (2014). FMSTM scores and low-back loading during lifting – Whole-body movement screening as an ergonomic tool? *Appl. Ergon.* 45, 482–489. doi: 10.1016/J.APERGO.2013.06.009

Beaudette, S. M., Zwambag, D. P., Graham, R. B., and Brown, S. H. M. (2019). Discriminating spatiotemporal movement strategies during spine flexion-extension in healthy individuals. *Spine J.* 19, 1–12. doi: 10.1016/j.spinee.2019.02.002

Bennett, H., Davison, K., Arnold, J., Slattery, F., Martin, M., and Norton, K. (2017). Multicomponent musculoskeletal movement assessment tools: a systematic review and critical appraisal of their development and applicability to professional practice. *J. Strength Condit. Res.* 31, 2903–2919. doi: 10.1519/JSC.0000000000002058

Bennetts, C. J., Owings, T. M., Erdemir, A., Botek, G., and Cavanagh, P. R. (2013). Clustering and classification of regional peak plantar pressures of diabetic feet. *J. Biomech.* 46, 19–25. doi: 10.1016/j.jbiomech.2012.09.007

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bock, C., Stierli, M., Hinton, B., and Orr, R. (2016). The functional movement screen as a predictor of police recruit occupational task performance. *J. Bodywork Mov. Ther.* 20, 310–315. doi: 10.1016/J.JBMT.2015.11.006

Brandon, S. C. E., Graham, R. B., Almosnino, S., Sadler, E. M., Stevenson, J. M., and Deluzio, K. J. (2013). Interpreting principal components in biomechanics: representative extremes and single component reconstruction. *J, Electromyogr. Kinesiol.* 23, 1304–1310. doi: 10.1016/j.jelekin.2013.09.010

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, G., Burton, L., and Hoogenboom, B. (2006a). Pre-participation screening: the use of fundamental movements as an assessment of function - part 1. *North Am. J. Sports Phys. Ther.* 1, 62–72.

Cook, G., Burton, L., and Hoogenboom, B. (2006b). Pre-participation screening: the use of fundamental movements as an assessment of function – part 2. *North Am. J. Sports Phys. Ther.* 1, 132–139.

Deluzio, K. J., and Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis. An application of principal component analysis. *Gait Posture* 25, 86–93. doi: 10.1016/j.gaitpost.2006.01.007

Deluzio, K. J., Harrison, A. J., Coffey, N., and Caldwell, G. E. (2014). *Research Methods in Biomechanics*. Champaign, IL: Human Kinetics.

Federolf, P., Reid, R., Gilgien, M., Haugen, P., and Smith, G. (2014). The application of principal component analysis to quantify technique in sports. *Scand. J. Med. Sci. Sports* 24, 491–499. doi: 10.1111/j.1600-0838.2012.01455.x

Frost, D. M., Beach, T. A. C., Mcgill, S. M., and Callaghan, J. P. (2015). The predictive value of general movement tasks in assessing occupational task performance. *Work* 52, 11–18. doi: 10.3233/WOR-141902

Gilles, M. A., and Wild, P. (2018). Grasping an object at floor-level: is movement strategy a matter of age? *Appl. Ergon.* 70, 34–43. doi: 10.1016/j.apergo.2018.02.002

Gross, D. P., and Battié, M. C. (2006). Does functional capacity evaluation predict recovery in workers' compensation claimants with upper extremity disorders? *Occup. Environ. Med.* 63, 404–410. doi: 10.1136/oem.2005.020446

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. (2018). Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J. Biomech.* 81, 1–11. doi: 10.1016/j.jbiomech.2018.09.009

Hewett, T. E., and Myer, G. D. (2011). The mechanistic connection between the trunk, hip, knee, and anterior cruciate ligament injury. *Exerc. Sport Sci. Rev.* 39, 161–166. doi: 10.1097/JES.0b013e3182297439

Holger Hoffmann, (2020). *Violin Plot*. Available at: https://www.mathworks.com/matlabcentral/fileexchange/45134-violin-plot (accessed March 25, 2020).

Isernhagen, S. J. (1992). Functional capacity evaluation: rationale, procedure, utility of the kinesiophysical approach. *J. Occup. Rehabil.* 2, 157–168. doi: 10.1007/BF01077187

Kiesel, K., Plisky, P. J., and Voight, M. L. (2007). Can serious injury in professional football be predicted by a preseason functional movement screen? *North Am. J. Sports Phys. Ther.* 2, 147–158.

Kritz, M., Cronin, J., and Hume, P. (2009). The bodyweight squat: a movement screen for the squat pattern. *Strength Condit. J.* 31, 76–85. doi: 10.1519/SSC.0b013e318195eb2f

Leardini, A., Chiari, L., Della Croce, U., and Cappozzo, A. (2005). Human movement analysis using stereophotogrammetry: part 3. Soft tissue artifact assessment and compensation. *Gait Posture* 21, 212–225. doi: 10.1016/j.gaitpost.2004.05.002

Lisman, P., O'Connor, F. G., Deuster, P. A., and Knapik, J. J. (2013). Functional movement screen and aerobic fitness predict injuries in military training. *Med. Sci. Sports Exerc.* 45, 636–643. doi: 10.1249/MSS.0b013e31827a1c4c

McCunn, R., Aus der Fünten, K., Fullagar, H. H. K., McKeown, I., and Meyer, T. (2016). Reliability and association with injury of movement screens: a critical review. *Sports Med.* 46, 763–781. doi: 10.1007/s40279-015-0453-1

McGill, S., Frost, D., Lam, T., Finlay, T., Darby, K., and Cannon, J. (2015). Can fitness and movement quality prevent back injury in elite task force police officers? A 5-year longitudinal study. *Ergonomics* 58, 1682–1689. doi: 10.1080/00140139.2015.1035760

Mottram, S., and Comerford, M. (2008). A new perspective on risk assessment. *Phys. Ther. Sport* 9, 40–51. doi: 10.1016/J.PTSP.2007.11.003

Moudy, S., Richter, C., and Strike, S. (2018). Landmark registering waveform data improves the ability to predict performance measures. *J. Biomech.* 78, 109–117. doi: 10.1016/j.jbiomech.2018.07.027

O'Connor, F. G., Deuster, P. A., Davis, J., Pappas, C. G., and Knapik, J. J. (2011). Functional movement screening: predicting injuries in officer candidates. *Med. Sci. Sports Exerc.* 43, 2224–2230. doi: 10.1249/MSS.0b013e318223522d

Okada, T., Huxel, K. C., and Nesser, T. W. (2011). Relationship between core stability, functional movement, and performance. *J. Strength Condit. Res.* 25, 252–261. doi: 10.1519/JSC.0b013e3181b22b3e

Ortiz, A., Rosario-Canales, M., Rodríguez, A., Seda, A., Figueroa, C., and Venegas-Ríos, H. L. (2016). Reliability and concurrent validity between two-dimensional and three-dimensional evaluations of knee valgus during drop jumps. *Open Access J. Sports Med.* 7, 65–73. doi: 10.2147/OAJSM.S100242

Parchmann, C. J., and McBride, J. M. (2011). Relationship between functional movement screen and athletic performance. *Strength Condit.* 25, 3378–3384. doi: 10.1519/JSC.0b013e318238e916

Powers, C. M. (2010). The influence of abnormal hip mechanics on knee injury: a biomechanical perspective. *J. Orthop. Sports Phys. Ther.* 40, 42–51. doi: 10.2519/jospt.2010.3337

Richter, C., McGuinness, K., O'Connor, N. E., and Moran, K. (2014). The variance needed to accurately describe jump height from vertical ground reaction force data. *J. Appl. Biomech.* 30, 732–736. doi: 10.1123/jab.2013-0313

Robertson, G., Caldwell, G., Hamill, J., Kamen, G., and Whittlesey, S. (2013). *Research Methods in Biomechanics, 2E.* Champaign, IL: Human Kinetics.

Roche, N., Pradon, D., Cosson, J., Robertson, J., Marchiori, C., and Zory, R. (2014). Categorization of gait patterns in adults with cerebral palsy: a clustering approach. *Gait Posture* 39, 235–240. doi: 10.1016/j.gaitpost.2013.07.110

Ross, G. B., Dowling, B. R., Troje, N. F., Fischer, S. L., and Graham, R. B. (2018). Objectively differentiating movement patterns between elite and novice Athletes. *Med. Sci. Sport Exerc.* 50, 1457–1464. doi: 10.1249/MSS.0000000000001571

Sawacha, Z., Guarneri, G., Avogaro, A., and Cobelli, C. (2010). A new classification of diabetic gait pattern based on cluster analysis of biomechanical data. *J. Diabetes Sci. Technol.* 4, 1127–1138. doi: 10.1177/193229681000400511

Schneiders, A. G., Davidsson, A., Hörman, E., and Sullivan, S. J. (2011). Functional movement screen normative values in a young, active population. *Int. J. Sports Phys. Ther.* 6, 75–82.

Shultz, R., Anderson, S. C., Matheson, G. O., Marcello, B., Besier, T. (2013). Test-retest and interrater reliability of the functional movement screen. *J. Athletic Train.* 48, 331–336. doi: 10.4085/1062-6050-48.2.11

Sinden, K. E., McGillivary, T. L., Chapman, E., and Fischer, S. L. (2017). Survey of kinesiologists' functional capacity evaluation practice in Canada. *Work* 56, 571–580. doi: 10.3233/WOR-172519

Srinivasan, D., and Mathiassen, S. E. (2012). Motor variability in occupational health and performance. *Clin. Biomech.* 27, 979–993. doi: 10.1016/j.clinbiomech.2012.08.007

Toro, B., Nester, C. J., and Farren, P. C. (2007). Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy. *Gait Posture* 25, 157–165. doi: 10.1016/j.gaitpost.2006.02.004

Troje, N. F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *J. Vis.* 2, 371–387. doi: 10.1167/2.5.2

Watelain, E., Barbier, F., Allard, P., Thevenon, A., and Angué, J. C. (2000). Gait pattern classification of healthy elderly men based on biomechanical data. *Arch. Phys. Med. Rehabil.* 81, 579–586. doi: 10.1016/s0003-9993(00)90038-8

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement.* Hoboken, NJ: John Wiley & Sons.

Witte, K., Ganter, N., Baumgart, C., and Peham, C. (2010). Applying a principal component analysis to movement coordination in sport. *Math. Comput. Model. Dyn. Syst.* 16, 477–488. doi: 10.1080/13873954.2010.507079

Wrigley, A. T., Albert, W. J., Deluzio, K. J., and Stevenson, J. M. (2005). Differentiating lifting technique between those who develop low back pain and those who do not. *Clin. Biomech.* 20, 254–263. doi: 10.1016/j.clinbiomech.2004.11.008

Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., et al. (2002). ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine. *J. Biomech.* 35, 543–548. doi: 10.1016/s0021-9290(01)00222-6

Wu, G., van der Helm, F. C. T., DirkJan Veeger, H. E. J., Makhsous, M., Van Roy, P., Anglin, C., et al. (2005). ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand. *J. Biomech.* 38, 981–992. doi: 10.1016/j.jbiomech.2004.05.042

Zazulak, B., Cholewicki, J., and Reeves, N. P. (2008). Neuromuscular control of trunk stability: clinical implications for sports injury prevention. *J. Am. Acad. Orthop. Surg.* 16, 497–505. doi: 10.5435/00124635-200808000-00011

Check for updates

A Video-Based Framework for Automatic 3D Localization of Multiple Basketball Players: A Combinatorial Optimization Approach

Lucas Antônio Monezi[1], Anderson Calderani Junior[1]*, Luciano Allegretti Mercadante[2], Leonardo Tomazeli Duarte[2] and Milton S. Misuta[1]

[1] *Faculty of Physical Education, University of Campinas, Campinas, Brazil,* [2] *School of Applied Science, University of Campinas, Limeira, Brazil*

Sports complexity must be investigated at competitions; therefore, non-invasive methods are essential. In this context, computer vision, image processing, and machine learning techniques can be useful in designing a non-invasive system for data acquisition that identifies players' positions in official basketball matches. Here, we propose and evaluate a novel video-based framework to perform automatic 3D localization of multiple basketball players. The introduced framework comprises two parts. The first stage is player detection, which aims to identify players' heads at the camera image level. This stage is based on background segmentation and on classification performed by an artificial neural network. The second stage is related to 3D reconstruction of the player positions from the images provided by the different cameras used in the acquisition. This task is tackled by formulating a constrained combinatorial optimization problem that minimizes the re-projection error while maximizing the number of detections in the formulated 3D localization problem.

Keywords: machine learning, sports, computer vision, player detection, non-invasive method, tracking

INTRODUCTION

Recent advances in sports science have been made possible due to the development of appropriate technology. For instance, computer-aided systems can be applied in several sports to obtain both high and low-level data about the performance of a player or team. In basketball, a typical example of low-level data is the position of a player on the court. Knowing the players' positions reveals important information because it can be used to compute higher-order data related to technical and physical activities as well as tactical analysis. As claimed in the sports science literature (Hopkins et al., 1999; McGarry et al., 2002), sports complexity must be analyzed at competitions, which means that non-invasive methods are preferable for acquiring data such as player position. In this respect, the fields of computer vision, image processing, and machine learning can play a role as they provide useful tools for designing non-invasive video-based systems to collect player motion data during official basketball matches. In the last two decades, researchers have made important contributions to individual and team sports analysis through the development of video-based computer-aided systems (Intille and Bobick, 1995; Iwase and Saito, 2004; Figueroa et al., 2006b; Barros et al., 2007, 2011; Gomez et al., 2014; Morais et al., 2014). In team sports, these studies particularly focus on tracking the players (Figueroa et al., 2006b; Barros et al., 2011; Morais et al., 2014) and

the ball (Stennett, 2003; Spagnolo et al., 2013). When tracking objects, one goal is to obtain the object's trajectory as a function of time, however, doing so requires a previous step: accurately detecting the object of interest. Therefore, tracking by video-based methods necessarily requires both object detection and determination of the object's location within the scene. This information can then be used to associate the identified objects with their trajectories. The desirable objects to track in team sports applications for evaluating game dynamics are the players, the referees, and the ball. As reported by Figueroa et al. (2006b); Barros et al. (2007, 2011), and Morais et al. (2014), several 2D approaches have been used in video-based applications for player and referee detection, localization, and tracking. In a 2D approach, only two spatial coordinates are taken into account, however, it is also possible to consider three spatial coordinates (a 3D approach), because 3D data provides a huge variety of information that will give coaches support in their training process, besides of the possibility in get game contextualized performance data for both physical and technical aspects. Generally, 3D methods have mainly been considered for ball tracking applications (Ohno et al., 2000; Stennett, 2003; Poliakov et al., 2010). However, the vertical component of player position is essential information in basketball analysis because the players frequently jump during the game. Player detection in basketball is not an easy task; artifacts such as player occlusions, strong shadows cast by players, and sharp reflections from the polished floor significantly affect the segmentation process (Alahi et al., 2009). The midpoint between the feet and the bottom center of a player's bounding box have both been used as reference points that determine a player's position within the image, allowing subsequent reconstruction of the player's 2D position on the court (Iwase and Saito, 2004; Figueroa et al., 2006b; Lu et al., 2009; Barros et al., 2011). However, the use of the midpoint between the feet as a reference point can lead to problems during the segmentation stage, especially when the legs have a color pattern similar to the basketball court itself. Because a low error rate in identifying players' positions is key for tracking algorithms, other reference points must be explored. One possible candidate in this respect is the player's head, which is less affected by the aforementioned artifacts. In fact, player head shape, color, and size provide more stable and invariant features than feet. Moreover, choosing head position on the court space as a reference point for locating a player and detecting heads in multiple cameras are measures that conform well to a 3D reconstruction approach. Bearing in mind the limitations and requirements discussed above, in this paper, we propose a video-based framework for automatic 3D localization of multiple basketball players. The paper is organized as follows. Section named Proposed Framework presents the two stages of our method: player detection and 3D reconstruction. Then, in Framework Performance Evaluation Section, we provide a set of numerical experiments to assess the performance of the proposed method. Discussion Section contains a talk of the results, and Conclusion Section presents the closing considerations. Approval for video data collection was obtained from the Brazilian National Basketball League and Limeira Basketball Association.

MATERIALS AND METHODS

Proposed Framework

The proposed framework, summarized in **Figure 1**, comprises two main parts. The first part addresses the identification of the players' heads at the camera image level. As will be detailed in section, player detection is conducted after image acquisition and requires image processing and machine learning procedures. The second stage of our proposal concerns the 3D reconstruction of player positions. This task, as will be discussed in section, can be addressed by formulating a combinatorial optimization problem.

Player Detection
Acquisition

Image acquisition is the first step required for player detection (**Figure 1**). To accomplish this, a dedicated capturing program was built (using Vimba SDK, OpenCV, and C/C++) to directly record and synchronize videos from multiple camera views to a computer. The video data (1,038 × 7,765 Hz) used in this work were acquired using three static industrial FireWire cameras (Allied Vision Technologies GmbH, with 6 mm lens) attached inside protected cages at the highest possible places in the gym (12 m from the ground). To extend the FireWire connection for the cameras and achieve the right locations for framing the court, a converter adapter was plugged into each camera using optical fiber (Gefen Firewire 1394 400/800 Extender). Because these cameras used aspherical lenses (C-mount, 6 mm), it was necessary to perform a distortion correction over the entire image. The correction protocol involved a chessboard (planar pattern) which was moved so the cameras could take images at different orientations. In this manner, a closed-form solution was obtained and refined for modeling the radial distortion (Zhang, 2000).

Image Processing

In order to illustrate part of the player detection process, we provide in **Supplementary Video S1**. The first step in player detection is based on segmentation, which separates parts of the image. For sports applications, these parts can be the court or playing field, the players, and balls or other gear. Some image processing tools were used to perform the segmentation of the basketball players. The basic idea was to separate static image regions from moving regions and then perform background segmentation to extract the portions of the images that showed the basketball court while keeping the parts that showed the players (**Figure 2**).

The Gaussian Mixture-based Background/Foreground Segmentation algorithm was used to do this (Zivkovic, 2004). With the background model, we can also detect shadows and re-mark any shadow pixels marked with foreground labels to background labels (Prati et al., 2003). Finally, noise suppression using image processing techniques that rely on morphological filtering (erosion followed by dilation) was applied to the foreground-mask (Figueroa et al., 2006a).

FIGURE 1 | Framework Diagram. The key tasks for player detection and 3D localization. For better understanding the structure of the method we produced a series of videos of the main process used: S1-detection, S2-1st round optimization, S3-2nd round optimization, and S4-3rd round optimization.
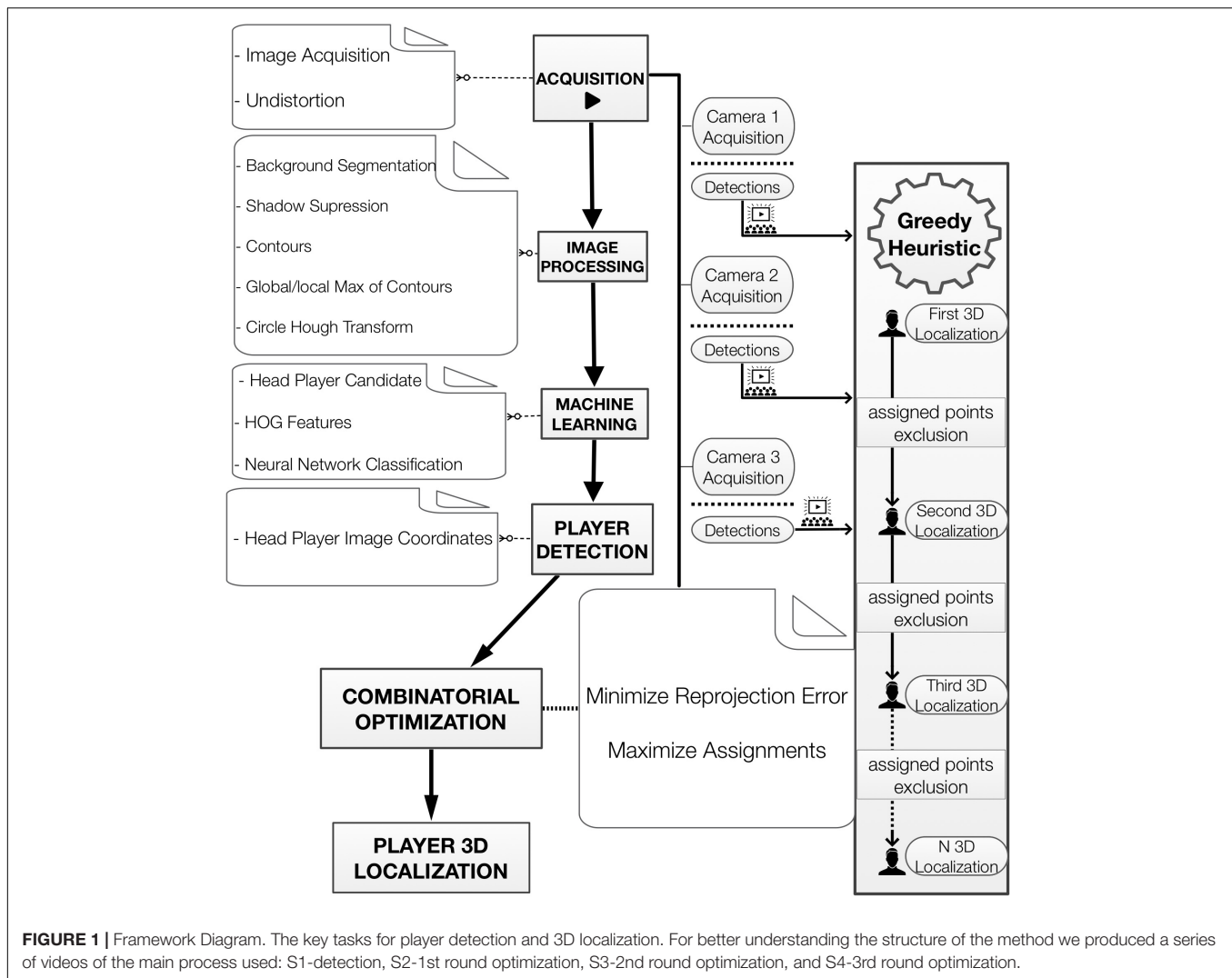
Machine Learning

The next step is to identify the players' heads. To do that, the first task is to estimate the contour of a player in the binary foreground image (Suzuki and Be, 1985). Due to the high number of players in the area, the contours found contained frequently more than one player. The highest point in a contour curve may be directly related to a player's head if the contour contains just one player (**Figure 2**, f, player from left), however, when the contour curve encloses multiple players (**Figure 2**, f, two players from right), the highest point can identify only one player's head. Therefore, to circumvent this issue, the location of the global maximum as well as of the local maxima are taken into account to search for circular patterns related to players' heads. Thus, a circle was fitted in the grayscale foreground mask near the places of maxima. Then, the Circle Hough Transform (Yuen et al., 1990) was adopted to obtain the best circle (taking head size into account) that was not too far away from the original local maxima (**Figure 3**).

A classification into "head" or "non-head" is the ultimate goal of player detection; therefore, the candidate points were classified by a multilayer perceptron neural network, which had previously

been trained. The features used were the Histogram of Oriented Gradients, HOG (Dalal and Triggs, 2005) using a fixed square region around the candidate player point (**Figures 3**, **4**).

The centers of the circles obtained are considered as the candidate points, however, these points are not necessarily heads – they could represent a raised arm (**Figure 5**, N1), the ball (**Figure 5**, N4), or any other non-head body segment. The candidate points are analyzed only if they appear inside of the interest area (a pre-determined polygon, **Figure 2**) thus, the high image variability caused by spectators and objects in areas beyond the court does not affect the detection process.

The chosen architecture was a multilayer perceptron feed-forward network with 10 hidden neurons in one hidden layer. The neural network was trained with a back-propagation algorithm (Rumelhart et al., 1986). To train and test the classifier, briefly, we selected a total of 30,009 labeled samples (**Figure 5**) and analyzed whether the HOG features and neural network classifications were suitable. The samples, collected from an official game (Game 1), were divided into three subsets: training (70%), validation (15%), and test (15%). Finally, reproducibility
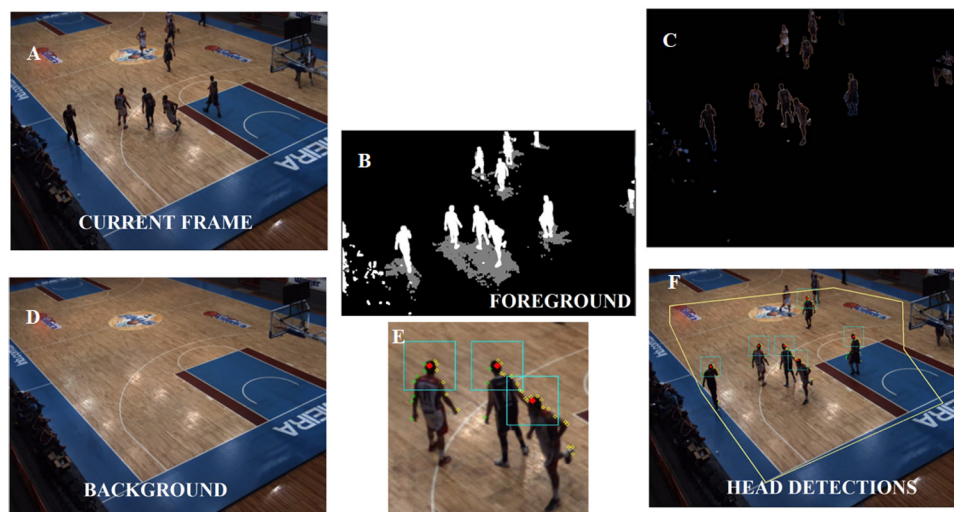
FIGURE 2 | Process inside of Player Detection stage. Player detection process over a delimited interest area **(E)** with some steps depicted. The video image **(A)** is used to model the static background image of the court **(B)**, and the foreground mask **(C)** is the result of subtracting the background **(D)**. The image in **(F)** is a closer view of the three player- head candidates.

FIGURE 3 | Head candidates' points. Heads enclosed by the Circle Hough Transform (the square around of the heads delimited region for a computation of HOG features).

FIGURE 4 | Neural network and histogram of oriented gradients representations. Architecture of neural network adopted with the input variables selected.

Reconstruction

A prerequisite task for a 3D reconstruction of a given point is to calibrate the cameras. The camera calibration aimed to estimate the parameters of each camera so we could later transform the image coordinates of the player reference point (the head in this paper) to the global coordinates associated with the court dimensions. After correcting for image distortion, a direct linear transform, DLT (Abdel-Aziz et al., 2015), was adopted to perform 3D camera calibration and player reconstruction (Qingchao et al., 1996; Rossi et al., 2015). In this calibration procedure, the intersections of the lines on the basketball playing court were chosen as reference spots; the measurements for these intersections (2D positions) were obtained from the official FIBA rules manual. The origin of the global system was defined at the intersection of one of the lateral lines

was checked using a different game (Game 2, 2,027 samples) with samples acquired in a scenario in which the players were wearing different jerseys (the visiting team) and players who had never appeared in the previous dataset of 30,009 samples. The neural network input layer had 1,764 neurons (**Figure 4**), which correspond to the values from the HOG features (Dalal and Triggs, 2005).

FIGURE 5 | Training samples examples. HOG features of positive and negative samples used as inputs for the neural network.
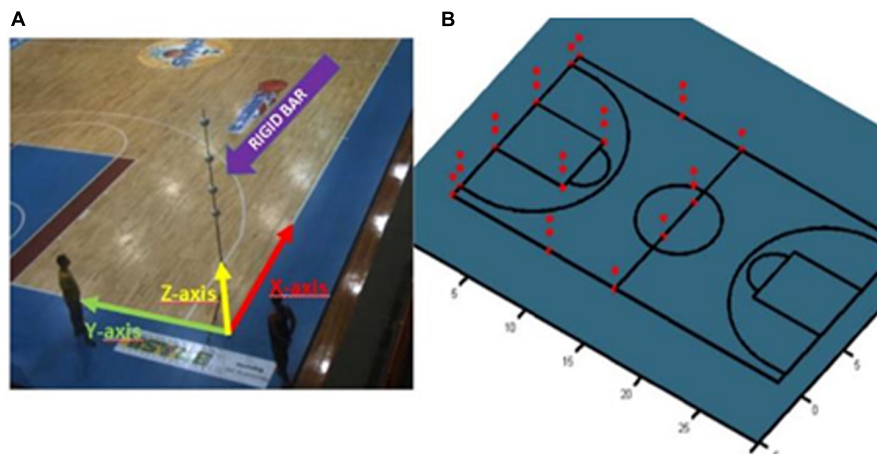
FIGURE 6 | Camera Calibration. **(A)** Rigid bar used for camera calibration, the adopted court reference system, and the calibration points (red, **B**) that appeared at least twice in the camera images.

(X-axis) with one of the bottom lines (Y-axis). Therefore, for any spot selected on the court plane, we placed a rigid vertically oriented bar (checked with a spirit level) that had demarcations along its length showing known measurements. At these demarcation points, white Styrofoam balls (with diameters of 15 cm) were fixed so they would be easy to visualize in camera images (**Figure 6**).

By measuring some points in the image with known real coordinates, it then becomes possible to solve for any point in the system using (Equations 1 and 2) to estimate the eleven DLT parameters. Eleven unknown variables require eleven or more equations, which means that a minimum of six pairs of points between the image 155 and the real measurements were required (because a point pair provides two equations) (Wood and Marshall, 1986; Abdel-Aziz et al., 2015). The parameters A1 to A11 are associated to the relation between object in space (position X, Y, and Z) and its image in the plane (position x

and y). The variable k denote to the camera index. For further details about DLT, please verify Abdel-Aziz et al. (2015) and http://www.kwon3d.com/theory/dlt/dlt.html. The absolute re-projection errors of the calibration points that appeared at least twice (36 points, **Figure 6B**) averaged 0.026 m (X-axis), 0.031 m (Y-axis), and 0.043 m (Z-axis). The DLT equations are as follows:

$$
\begin{aligned}
(\lambda_{1_k} - \lambda_{3_k} x_{p_k})\mathrm{X} &+ (\lambda_{4_k} - \lambda_{6_k} x_{p_k})\mathrm{Y} + (\lambda_{7_k} - \lambda_{9_k} x_{p_k})\mathrm{Z} \\
&+ \lambda_{10_k} - x_{p_k} = 0 \quad (1)
\end{aligned}
$$

$$
\begin{aligned}
(\lambda_{2_k} - \lambda_{3_k} y_{p_k})\mathrm{X} &+ (\lambda_{5_k} - \lambda_{6_k} y_{p_k})\mathrm{Y} + (\lambda_{8_k} - \lambda_{9_k} x_{p_k})\mathrm{Z} \\
&+ \lambda_{11_k} - y_{p_k} = 0 \quad (2)
\end{aligned}
$$

After obtaining the parameters, it was then possible to reconstruct the X, Y, and Z coordinates in court space from at least two pairs of camera image coordinates.

The goal of the proposed 3D reconstruction process is to estimate the 3D localization of all the players in the scene. The process is built upon a constrained combinatorial optimization problem. The underlying problem-to assign the points detected to a true player-can be modeled by the following assignment matrix:

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \tag{3}$$

where $A_{p,l}$ is a binary (decision) variable that takes a value of 1 when an indexed image point from player head detection p is related to a labeled player l; otherwise, it takes a value of 0. To locate a given player in court space, a reconstruction must be performed by considering only the points that represent the given player. The problem here is that we do not know the labels of the detected points; in other words, the associations between detected points and players are unknown. A possible solution to that problem would be to test all possible combinations, searching for the combination that minimizes the re-projection error. Unfortunately, that is a combinatorial optimization problem that can be extremely costly in terms of computation. Moreover, an additional complication is that the number of players in the scene is unknown. Thus, to estimate the assignment matrix A, we propose a constructive greedy solution that initially locates a single player. Having located this first player, the image points related to the head of the located player are dropped out of the next interaction. This heuristic drastically decreases the number of required calculations. The 3D localization of a new player stops when no more feasible solutions are available. If there are remaining player head image points that have not been assigned yet, the method takes a priori information into account (Z equal to the mean player height) to locate the last players on the court.

Let us detail our approach (**Table 1**) presents the notation considered herein). Basically, we seek to optimize two cost functions: (4) the minimization of the sum of the re-projection errors associated with the assigned points, which is mathematically given by the following equation:

$$\min \sum_{l=1}^{n} \sum_{p=1}^{m} \left[(x_{pk} - x_{lk})^2 (y_{pk} - y_{lk})^2 \right] A_{pl} \tag{4}$$

and, (5) the maximization of the number of the assigned points, which is given by the following:

$$\max \sum_{l=1}^{n} \sum_{p=1}^{m} A_{pl} \tag{5}$$

The rationale behind this cost function comes from the notion that greater numbers of designated image points allow for better approximations of the players' localizations. Of course, this is not the case for outlier points, which require additional constraints to prevent their designation. Note that the cost functions expressed in Equations (4) and (5) are in conflict because a larger number

TABLE 1 | Notation.

n	Number of players
m	Number of points detected as player's heads
p	Point index
l	Player label index
$x_{lk} y_{lk}$	Coordinates of player l re-projected
w	Total number of cameras
k	Camera index
$X_l Y_l Z_l$	Court-space coordinates of player l
$\lambda_{1_k}, \lambda_{2_k}, \cdots, \lambda_{11_k}$	DLT parameters of camera k

of designated players increases the re-projection error. In view of this fact, we propose to merge these cost functions into a single function, as follows:

$$\min \sum_{l=1}^{n} \left[\frac{(x_{pk} - x_{lk})^2 (y_{pk} - y_{lk})^2}{\left(\sum_{p=1}^{m} A_{pl} \right)^2} - \sum_{p=1}^{m} A_{pl} \right] A_{pl} \tag{6}$$

For the greedy solution, we solve the 1st round to obtain the location of the 1st player (**Figures 7**, **8**) and then proceed to the subsequent rounds (2nd player, 3rd player, etc.) by discarding the image points that have already been designated.

The examples can be seen in **Supplementary Videos S1–S4**; respectively, for first, second and third players localized in each round in straight for a given frame. The minimization of (6) must be conducted by considering the following set of constraints:

$$\sum_{l=1}^{n} A_{pl} \leq 1, \forall p \tag{7}$$

$$\sum_{\iota=1}^{m} A_{p\iota} \leq w, \forall l \tag{8}$$

$$\sum_{\iota=1}^{n} A_{p\iota} \geq 2, \forall l \tag{9}$$

$$A_{p\iota} \left[(x_{pk} - x_{lk})^2 (y_{pk} - y_{lk})^2 \right] \geq , \forall p \tag{10}$$

$$h_{min} \leq Z_l \leq h_{max} \tag{11}$$

Constraint (7) means that for every point p, only one player l can be assigned. Constraint (8) means that for every player, the number of points assigned must be equal to or less than the number of cameras. Constraint (9) means that for every player, the number of points assigned must be equal or greater than two (this is required for 3D reconstruction). Finally, Constraint (10) sets the maximum re-projection error in terms of pixel tolerance, and Constraint (11) imposes the head height limits. Having estimated the assignment matrix A, the reconstruction of the 3D position of each player can be obtained by solving a set of algebraic equations, as shown in **Table 2**. An example of combination tested is represented in **Figure 8**.

FIGURE 7 | The best feasible solution obtained in first player localization. In this case, the head points are correctly assigned to the player obtaining a low re-projection error in all three cameras and then a god value for the optimization function.

FIGURE 8 | An example of combination tested with re-projection errors detailed. The head detected (red asterisk) that were designed to a given player (green circles) resulting in a poor 3D reconstruction, which the value of re-projection errors (red lines between green circles and yellow plus sign) are too large, and then provide an infeasible solution due to the one constraint.

As already mentioned, due to the limited number of cameras along with the requirement of having at least two points from different views for 3D reconstruction, some detected points are not assigned as players and, thus, are not located in court space. In these cases, we used a priori information to solve a DLT with a fixed mean height. **Figure 9** describes an example of a 3D

TABLE 2 | Systems equations build upon DLT.

$p = 1$	$\left[(\lambda_{1_k} - \lambda_{3_k}x_{1_k})X + (\lambda_{4_k} - \lambda_{6_k}x_{1_k})Y + (\lambda_{7_k} - \lambda_{9_k}x_{1_k})Z + \lambda_{10_k} - x_{1_k}\right]A_{1_l} = 0$
	$\left[(\lambda_{2_k} - \lambda_{3_k}y_{1_k})X + (\lambda_{5_k} - \lambda_{6_k}y_{1_k})Y + (\lambda_{8_k} - \lambda_{9_k}y_{1_k})Z + \lambda_{11_k} - y_{1_k}\right]A_{1_l} = 0$
$p = 2$	$\left[(\lambda_{1_k} - \lambda_{3_k}x_{2_k})X + (\lambda_{4_k} - \lambda_{6_k}x_{2_k})Y + (\lambda_{7_k} - \lambda_{9_k}x_{2_k})Z + \lambda_{10_k} - x_{2_k}\right]A_{1_l} = 0$
	$\left[(\lambda_{2_k} - \lambda_{3_k}x_{1_k})X + (\lambda_{5_k} - \lambda_{6_k}y_{2_k})Y + (\lambda_{8_k} - \lambda_{9_k}y_{2_k})Z + \lambda_{11_k} - y_{2_k}\right]A_{1_l} = 0$
...	...
$p = m$	$\left[(\lambda_{1_k} - \lambda_{3_k}x_{m_k})X + (\lambda_{4_k} - \lambda_{6_k}x_{m_k})Y + (\lambda_{7_k} - \lambda_{9_k}x_{m_k})Z + \lambda_{10_k} - x_{m_k}\right]A_{m_l} = 0$
	$\left[(\lambda_{2_k} - \lambda_{3_k}x_{m_k})X + (\lambda_{5_k} - \lambda_{6_k}y_{m_k})Y + (\lambda_{8_k} - \lambda_{9_k}y_{m_k})Z + \lambda_{11_k} - y_{m_k}\right]A_{m_l} = 0$

Direct Linear Transform (DLT) equation composing a system that solve the 3D position (X; Y; Z) for a given player.

localization result with the proposed framework; the player head detection (red asterisk) and assignments (colored circles) were obtained for a given frame as an example.

The first player localized is shown by a blue circle, the second by a green circle, and the third by a red circle. These are the top optimizations. For this example, 3 players were 3D localized by optimization and 3 players was located using a priori information (the remaining points detected only in camera 1-up left) at a position near the middle of the court (the white/red diamond). **Supplementary Video S5** depicted the framework result in a short sequence movie.

RESULTS

For a better understanding of results presented below, the results section will be divided into 3 subsections, as named: Neural Network Classification Performance, Players Detection Evaluation and Player Localization Evaluation.

Neural Network Classification Performance

The performance obtained by the neural network for the 30,009 samples in Game 1 is illustrated in **Figure 10** using confusion matrices.

Each confusion matrix depicts the occurrences of true classification (\head" classified as \head" or \non-head" classified as \non-head"), false positive classification (\non-head" classified as \head"), and false negative classification (\head" classified as \non-head") for one of the subsets or for all subsets together. A classification was considered positive only if the head appears centralized in the square region where the HOG features were computed (e.g., in N3 in **Figure 5**, the head is not centralized). Finally, it is worth mentioning that the values presented in **Figure 10** represent the rates of the neural network classification task, which do not correspond to the rates at the player detection stage. **Figure 11** shows the results obtained for Game 2 (2,027 samples). We denote good performance, however, we cannot forget to state that is an unbalanced classification problem.

Note that – despite the fact that the neural network was trained using samples from Game – the results obtained for Game 2 were also satisfactory. Finally, in the **Table 3** is presented the 2D localization error's benchmark of our method with others presented in literature.

FIGURE 9 | 3D Localization of a given frame. Localization of 3 players, in meters, by optimization (the colored circles) and 3 remaining point. Player localizations represented on the basketball playing court **(D)**.

The player detection accuracy was compared against a manual measurement (ground truth) performed using the DVideo (Campinas, SP, BRAZIL) system (Figueroa et al., 2006b; Barros et al., 2007) run by an expert operator (5 years of experience). The detection rates for each camera were calculated by considering that a detection was true when the pixel distance between a detected player's head and the ground truth was less than 25 pixels. A pixel distance greater than 25 pixels was considered to be a false positive detection. A misdetection occurred when no point was found near a manual measure. The measures were performed only inside of a designated interest area (a pre-determined polygon). In addition, the player localization performance of the proposed framework was evaluated by comparing the real distance in meters between the player's position as reconstructed by the proposed framework and the expert's manual measurement in DVideo.

Player Detection Evaluation

The results of 10,164 detections were as follows: for cameras 1, 2, and 3, respectively, the true detection rates were 78.9, 68.9, and 79.8%; the false positive rates were 2, 1.2, and 5%; and the misdetection rates were 19.1, 29.9, and 19.7%. For just the true detections, the root mean squared error (RMSE) found was 6.59 pixels.

FIGURE 10 | Neural Network classification. Neural network confusion matrices (Matlab confusion matrix plot model), \head" is 1 and \non-head" is 0. True classifications are denoted by green squares (a,e) and false classifications are denoted by red squares (b,d). Blue squares indicate the overall rates (correctness and error rate) of classifications (l). Gray squares (c,f,g,h) show the conditional rates (correctness and error rate) given a pre-determined target (3rd row, g,h) or given a pre-determined output (3rd column, c,f). The performance was evaluated in Game 1 for each subset.

Player Localization Evaluation

Error computation of player localization by optimization were accounted for only the players present in at least two cameras visualization inside of interest area (2,941 samples). The RMSE of 0.16 m in plane court (axis X and Y), and RMSE of 0.18 m in space court reference (axis X, Y, and Z). For the remaining (917 samples) point issues that were not assigned (not localized by optimization), the errors are shown an RMSE of 0.30 m in plane court (axis X and Y) and 0.33 m in space court (axis X, Y, and Z). The processing time required to assign the players' locations by combinatorial optimization grows exponentially according to the number of player heads detected. The computation time required

TABLE 3 | Some works in the literature on player detection in team sports.

References	Sport	3D localization	Method	Localization error (2D)
Morais et al. (2014)	Futsal	No	AdaBoost detection + particle filter	⁓0.6 m
Needham and Boyle (2001)	Futsal	No	Background segmentation (BS)	⁓1.16 m
Pers et al. (2001)	Handball	No	Background segmentation + template matching	⁓0.32 m
Barros et al. (2011)	Handball	No	AdaBoost detection + graph	–
Alahi et al. (2009)	Basketball	No	Planar homography	–
Delannay et al. (2009)	Basketball	No	Mean-shift segmentation	–
Figueroa et al. (2006a)	Soccer	No	Background segmentation	–
Ours	Basketball	Yes	BS + neural network + combinatorial optimization	⁓0.16 m

Note that not all works report a value for the player position error.

FIGURE 11 | Classification in another game. Reproducibility of the classification by the neural network using images from an independent game.

FIGURE 12 | Processing time. Processing time (s) to solve the 3D localization of one player based on the number of total points detected in the camera images.
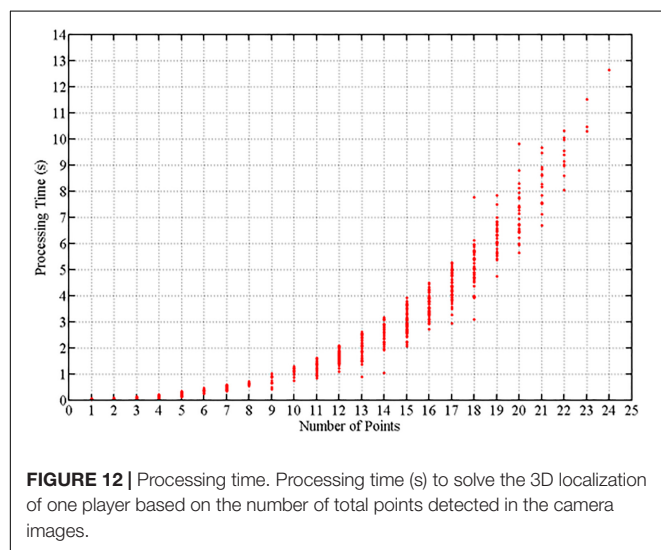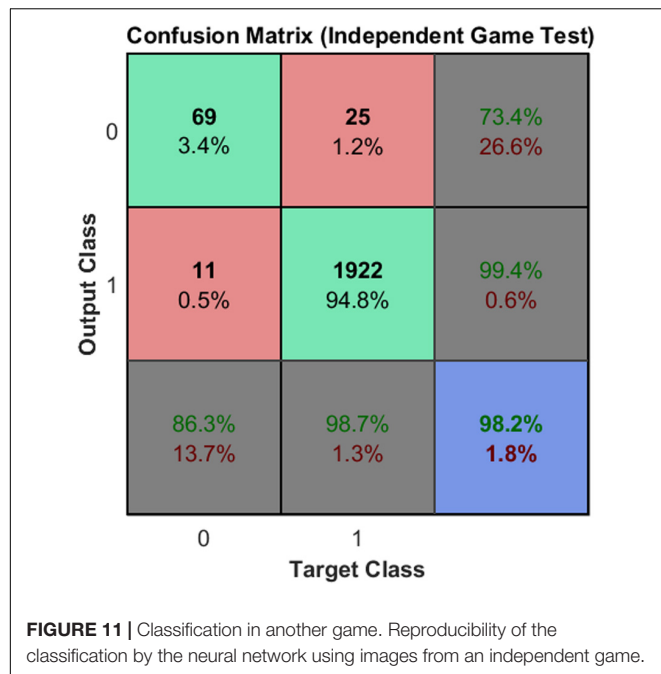
was computed for the preliminary Matlab code (not parallelized) and is depicted in **Figure 12**. This is the time required to measure the 3D localization of just one player, however, the localization of the next player in the optimization problem requires at least two fewer points.

DISCUSSION

The framework presented in the previous section and its performance are discussed in this section in more detail. The steps of player detection and localization are inherent to video-based tracking methods and play an important role in tracking. Therefore, we will discuss how tracking methods

reported both in the literature and our proposal address player detection and localization and how other methods determine the players' positions compared with our approach. Systems for data acquisition in sports must be feasible and reliable; therefore, a complete and automatic solution to measuring the players' positions on the basketball court cannot be achieved using the knowledge from only a single research field. Thus, it was necessary to integrate tools from different fields for our approach to successfully localize the players. Searching for high-interest objects in player tracking works by integrating methods from several fields. We observed that such integrations corroborate the choice of going beyond the frontiers of knowledge in any one field. Some integration examples in the literature include image processing mixed with graph representations (Figueroa et al., 2006b), AdaBoost detection mixed with a particle filter (Lu et al., 2009; Morais et al., 2014), AdaBoost detection mixed with graph representations (Barros et al., 2011), and image processing mixed with clustering (Chen et al., 2012). Mixing together techniques from image processing, machine learning, computer vision and optimization was vital for the localization of multiple basketball players on a court. Even after the technical procedures were in place, we still faced innumerable difficult tasks in basketball player localization such as our chosen approach to seek head patterns with a Circle Hough transform, or adopting neural network classification to reject non-head points – or even using optimization to select the best assignments. State-of-the-art video-based methods for detection and player tracking in team sports take 2D positions into account, however, in the context of basketball analysis, kinematic variables that consider the vertical component of player position are essential because many sort of measurements that could be get from 3D positional data, such as specific efforts made at jump actions, fatigue index by detecting jump's height variation during game, height of ball during passes, balls trajectory, rebounds height efficiency reached by players and vertical components of shots in different game contexts (free throws at specific moments of the match, differences of 2-point and 3-point shots, difference performance at open shots and tight shots, etc.). All these information could be useful for a more accurately diagnostics by coach's staff in order to improve specific trainings and get a better performance from players in court. Therefore, our principal contribution lies in considering the 3D position of a given player reference point. Instead of using the bottom center of a player's bounding box or silhouette, which attempts to represent the position of a player's feet, we chose the player's head as the reference point, and that decision plays a key role in our framework. There were two reasons behind this choice. First the goal was to perform a 3D reconstruction using a point that lies in court space ($Z_6 = 0$) and second, the solution needed to address the frequent player occlusions. Given these goals, analyzing the player's head position was more stable and more robust to occlusions and other effects from illumination across the court. Our proposed framework comprises two main parts: (i) the detection of the players' heads from the camera images and, (ii) the 3D reconstruction of the players' positions. Starting with the first part, detection, we will present the levels of performance accuracy found in the literature. Then, the second part investigates the accuracy of these methods in estimating

the players' positions. The player detection rate of our proposed framework was ~71%. In other works that concentrated on indoor team sports, we found the following detection rates: 74% in a study of handball when applied to a game other than the training game (Barros et al., 2011), 70.5% in a basketball study (Delannay et al., 2009). Still focusing on basketball, the performance reported by a detection approach that used a mixed network of planar and omnidirectional cameras achieved a recall of 0.76 and a precision of 0.72 (Alahi et al., 2009). Works on outdoor team sports have also evaluated player detection rates. Experimental results from a soccer study reported 81.50 and 78.03% detection rates by two player detection methods based on a neural network and on Viola and Jones' AdaBoost, respectively (Lehuger et al., 2007). A method for automatically tracking soccer players can locate players in 94% of video frames (Barros et al., 2007). These studies, which focused on automatic detection and tracking of outdoor team sports, used several approaches that have also been studied for indoor applications – although, for indoor sports, the camera setup (quantity, resolution, view point) and the problems faced are slightly different and include the number of players to be detected, interference from environmental features, and the spatial organization of players. Thus, outdoor studies do not allow a direct comparison with our results in detecting basketball players indoors. The median error of nearly 10 pixels in determining the players' head positions 318 seems to be appropriate because we used images with a resolution of 1,038 × 776 pixels. An average RMS error value of ~3.4 pixels was found in a work targeted toward indoor sports applications (handball and basketball) using images with a resolution of 348 × 288 pixels obtained from gym ceiling cameras (Kristan et al., 2009). The average error in determining the position of hockey players' foot positions from images was 20% of the width of the ground-truth box, however, this work did not present the error in terms of pixels (Lu et al., 2009). Proceeding with the discussion of the accuracy in determining a player's position on court, even with a limited number of cameras (three in this study), it was possible to use the proposed framework to detect and localize multiple basketball players in 3D space. The median 3D error of 0.25 m was suitable considering the players were being localized on one-half of the full basketball court (14 × 15 m). Moreover, changes in the values of the parameters for optimization constraints and adding additional cameras can decrease the errors. To understand the error results through comparisons with other studies, the average cumulative error of 0.60 m in a 2D trajectory approach presented by Morais et al. (2014) used multiple-camera methodology developed for Futsal (on a playing surface of 20 × 40 m) with the errors attenuated by Fusion AdaBoost (Viola and Jones, 2001), detection from four camera images, and player appearance models. A mean error value of 0.20 m, which is associated with the uncertainties of the position of points on the visible court and not to player position error, was reported for a handball tracking study (Barros et al., 2011). Experiments showed an RMS error in player position of 0.28 m near the optical axis and 0.36 m for the court boundary when tracking handball players using ceiling cameras (Pers et al., 2001). An automatic tracking soccer study reported a spatial resolution of 0.3 m (Barros et al., 2007). A study focused on automatically tracking the positions of indoor 5-a-side football

players (on a playing surface of 18 × 32 m) achieved an RMS of 1.16 m and a modal value below 40 cm compared with manual tracking (Needham and Boyle, 2001). In our approach, we attack the problem of localizing multiple basketball players using a video-based framework. Yet another alternative for tracking player position is to use a global positioning system (GPS), however, at present, the errors from GPS measurements are too large even outdoors (Gray et al., 2010), and indoor use is impracticable. To provide an example of GPS accuracy, ~50% of the GPS coordinates were within 2.5 m in a static position test (Mohamad et al., 2009). However, GPS systems often do not work at all in basketball gyms. In addition, the rules of many sports do not allow the players to use GPS devices. Because no temporal information was used in our proposed framework, the results could be improved by using the players' trajectories to help predict their current positions, filtering the player trajectory data to discard outlier positions. Although linking temporal information to player detection was not the goal of this study, it is an aspect that could be investigated in future studies.

CONCLUSION

A video-based framework for automatic 3D localization of multiple basketball players was described in the context of official games. Player detection was based on image processing techniques and – considering the complexity of basketball games – the classification problem presented satisfactory results. The classification procedure was essential to properly reject head candidate points (for example, to reject other body parts such as a raised arm).

A combinatorial optimization problem was solved with a greedy heuristic and provided satisfactory results in accurately determining both the number of players in a scene and their positions. Knowing the player's positions in 3D relative to the court is crucial for basketball performance analysis due to the nature of the sport. This work helps to further systems development aiming to acquire 3D player position data during competitions, and the application can be extended to other indoor team sports in which a vertical component is relevant.

DATA AVAILABILITY STATEMENT

All the datasets generated and analyzed for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

LMo, LD, and MM: writing-original draft preparation. LMo, LD, MM, AJ, and LMe: investigation, data analysis, and conceptualization. LMo, LD, MM, and LMe: funding acquisition.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe. 2020.00286/full#supplementary-material

VIDEO S1 | Player detection process. Candidate points obtained through some image processing including a background segmentation approach to be classified into head or non-head.

VIDEO S2 | First round of optimization. Test of all possible combinations for first player localization with the best feseable solution. Head points detected depicted as red asterisk, points assign to the player showed as green circles, camera reprojection of 3D reconstructed player in the current combination tested marked as yellow plus sign, and, court plane representation with 3D player position reconstructed for tested solution depicted as green plus sign.

VIDEO S3 | Second round of optimization. Test for all combinations for localization of the second player remember that head point assigned to the first player are excluded for current cominations. Head points detected depicted as red asterisk, points assign to the player showed as green circles, camera reprojection of 3D reconstructed player in the current combination tested marked as yellow plus sign, and, court plane representation with 3D player position reconstructed for tested solution depicted as green plus sign.

VIDEO S4 | Third round of optimization. Test for all combinations for localization of the second player remember that head point assigned to the first player are excluded for current cominations. Head points detected depicted as red asterisk, points assign to the player showed as green circles, camera reprojection of 3D reconstructed player in the current combination tested marked as yellow plus sign, and, court plane representation with 3D player position reconstructed for tested solution depicted as green plus sign.

VIDEO S5 | Player Localization in game. A sequence example of player localization. The player localized in 3D court spacare reprojected in each camera keeping the same color circles for the same player.

REFERENCES

Abdel-Aziz, Y. I., Karara, H. M., and Hauck, M. (2015). Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. *Photogramm. etric Eng. ineering & Remote Sensing.* 81, 103–107. doi: 10.14358/pers.81.2.103

Alahi, A., Boursier, Y., Jacques, L., and Vandergheynst, P. (2009). "Sport players detection and tracking with a mixed network of planar and omnidirectional cameras," in *Proceedings of the 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, (Como: IEEE), 1–8. doi: 10.1109/ICDSC.2009.5289406

Barros, R. M. L., Menezes, R. P., Russomanno, T. G., Misuta, M. S., Brandão, B. C., Figueroa, P. J., et al. (2011). Measuring handball players trajectories using an automatically trained boosting algorithm. *Comput. er Methods in Biomech.anics and Biomedical . Engineering.* 14, 53–63. doi: 10.1080/10255842.2010.494602

Barros, R. M. L., Misuta, M. S., Menezes, R. P., Figueroa, P. J., Moura, F. A., Cunha, S. A., et al. (2007). Analysis of the distances covered by first division brazilian soccer players obtained with an automatic tracking method. *J. ournal of Sports Sci. ence and Medicine.* 6, 233–242.

Chen, H. T., Chou, C. L., Fu, T. S., Lee, S. Y., and Lin, B. S. P. (2012). Recognizing tactic patterns in broadcast basketball video using player trajectory. *J. ournal of Visual . Communication and. Image Representation.* 23, 932–947. doi: 10.1016/j.jvcir.2012.06.003

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1:886–93 vol. 1,* (San Diego, CA: IEEE)., doi: 10.1109/CVPR.2005.177

Delannay, D., Danhier, N., and De Vleeschouwer, C. (2009). "Detection and recognition of sports(Wo)men from multiple views," in *Proceedings of the 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, (Como: IEEE), 1–7. doi: 10.1109/ICDSC.2009.5289407

Figueroa, P. J., Leite, N. J., and Barros, R. M. L. (2006a). Background recovering in outdoor image sequences: an example of soccer players segmentation. *Image and Vision . Computing.* 24, 363–374. doi: 10.1016/j.imavis.2005.12.012

Figueroa, P. J., Leite, N. J., and Barros, R. M. L. (2006b). Tracking soccer players aiming their kinematical motion analysis. *Comput. er Vision and. Image Understanding.* 101, 122–135. doi: 10.1016/j.cviu.2005.07.006

Gomez, G., López, H. P., Link, D., and Eskofier, B. (2014). Tracking of ball and players in beach volleyball videos. *PlLoS One* 9:e111730. doi: 10.1371/journal.pone.0111730

Gray, A. J., Jenkins, D., Andrews, M. H., Taaffe, D. R., and Glover, M. L. (2010). Validity and reliability of GPS for measuring distance travelled in field-based team sports. *J. ournal of Sports Sciences.* 28, 1319–1325. doi: 10.1080/02640414.2010.504783

Hopkins, W. G., Hawley, J. A., and Burke, L. M. (1999). Design and analysis of research on sport performance enhancement. *Med. icine and Science in. Sports and Exercise.* 31, 472–485. doi: 10.1097/00005768-199903000-00018

Intille, S. S., and Bobick, A. F. (1995). "Closed-world tracking," in *Proceedings of the Fifth International Conference on Computer Vision, 672–.* ICCV '95, (Washington, DC: IEEE Computer Society).

Iwase, S., and Saito, H. (2004). "Parallel tracking of all soccer players by integrating detected positions in multiple view images," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 4:751-754 Vol.4,* (Cambridge: IEEE),. doi: 10.1109/ICPR.2004.1333881

Kristan, M., Perš, J., Perše, M., and Kovaèiè, S. (2009). Closed-world tracking of multiple interacting targets for indoor-sports applications. *Comput. er Vision and. Image Understanding.* 113, 598–611. doi: 10.1016/j.cviu.2008.01.009

Lehuger, A., Duffner, S., and Garcia, C. A. (2007). *A Robust Method for Automatic Player Detection in Sport Videos.* Paris: Orange Labs.

Lu, W. L., Okuma, K., and Little, J. J. (2009). Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image and Vision . Computing.* 27, 189–205. doi: 10.1016/j.imavis.2008.02.008

McGarry, T., Anderson, D. I., Wallace, S. A., Hughes, M. D., and Franks, I. M. (2002). Sport competition as a dynamical self-organizing system. *J. ournal of Sports Sciences.* 20, 771–781. doi: 10.1080/026404102320675620

Mohamad, I., Ali, M. A. M., and Ismail, M. (2009). "Availability, reliability and accuracy of GPS signal in bandar baru bangi for the determination of vehicle position and speed," in *Proceedings of the 2009 International Conference on Space Science and Communication*, (Negeri Sembilan: IEEE), 224–229. doi: 10.1109/ICONSPACE.2009.5352632

Monezi, L. A. (2016). *Localização de Jogadores de Basquetebol e Estimativa Da Frequência Cardíaca Em Jogos Oficiais. Aplicações Baseadas Em Aprendizado de Máquina e Otimização Combinatória. Basketball Player Localization and Heart Rate Estimation in Official Games: Applications Based on Machine Learning and Combinatorial Optimization.". recurso online (49 p.). Dissertação (Mestrado).* Campinas: Universidade Estadual de Campinas, Faculdade de Ciências Aplicadas. Available online at: http://www.repositorio.unicamp.br/handle/REPOSIP/321845

Morais, E., Ferreira, A., Cunha, S. A., Barros, R. M. L., Rocha, A., and Goldenstein, S. (2014). A multiple camera methodology for automatic localization and tracking of futsal players. *Patt. ern Recognition . Letters.* 39, 21–30. doi: 10.1016/j.patrec.2013.09.007

Needham, C. J., and Boyle, R. D. (2001). "Tracking multiple sports players through occlusion, congestion and scale," in *Proceedings of the Proceedings of the British Machine Vision Conference 2001, BMVC 2001,* (Manchester), 93–102. doi: 10.5244/C.15.11

Ohno, Y., Miura, J., and Shirai, Y. (2000). "Tracking players and estimation of the 3D position of a ball in soccer games," in *Proceedings of the 15th International Conference on Pattern Recognition. ICPR-2000, 1:145–48 vol.1,* (Barcelona: IEEE)., doi: 10.1109/ICPR.2000.905293

Pers, J., Bon, M., and Kovačič, S. (2001). *Errors and Mistakes in Automated Player Tracking.* Piscataway, NJ: IEEE, 25–36.

Poliakov, A., Marraud, D., Reithler, L., and Chatain, C. (2010). Physics based 3D ball tracking for tennis videos. *Proceedings of the 2010 International Workshop on Content Based Multimedia Indexing (CBMI)* (Grenoble: IEEE), 1–6. doi: 10.1109/CBMI.2010.5529897

Prati, A., Mikic, I., Trivedi, M. M., and Cucchiara, R. (2003). Detecting moving shadows: algorithms and evaluation. *IEEE Transact. ions on Pattern . Analysis . and Machine . Intelligence.* 25, 918–923. doi: 10.1109/TPAMI.2003.1206520

Qingchao, W., GuoQing, Z., Qing, Z., and Qinghang, Z. (1996). "On DLT method for CCD camera calibration," in *Proceedings of the Third International Conference on Signal Processing (ICSP'96), 2:883–85 vol.2,* (Beijing: IEEE)., doi: 10.1109/ICSIGP.1996.566229

Rossi, M. M., Silvatti, A. P., Dias, F. A. S., and Barros, R. M. L. (2015). Improved accuracy in 3D analysis using DLT after lens distortion correction. *Comput. er Methods in Biomechanics and. Biomedical . Engineering.* 18, 993–1002. doi: 10.1080/10255842.2013.866231

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Spagnolo, P., Leo, M., Mazzeo, P. L., Nitti, M., Stella, E., and Distante, A. (2013). "Non-invasive soccer goal line technology: a real case study," in *Proceedins of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1011–18,* (Portland, OR: IEEE)., doi: 10.1109/CVPRW.2013.147

Stennett, C. (2003). "Hawk-eye tennis system," in *Proceedins of the IET Conference Proceedings, 182-185(3).* Available online at: https://digital-library.theiet.org/content/conferences/10.1049/cp_20030517.

Suzuki, S., and Be, K. (1985). Topological structural analysis of digitized binary images by border following. *Comput. er Vision, . Graph. ics, and Image Processing.* 30, 32–46. doi: 10.1016/0734-189X(85)90016-7

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1:I–I,* (Kauai, HI: IEEE)., doi: 10.1109/CVPR.2001.990517

Wood, G. A., and Marshall, R. N. (1986). The accuracy of DLT extrapolation in three-dimensional film analysis. *J. ournal of Biomechanics.* 19, 781–785. doi: 10.1016/0021-9290(86)90201-0

Yuen, H. K., Princen, J., Illingworth, J., and Kittler, J. (1990). Comparative study of hough transform methods for circle finding. *Image and Vision. Computing.* 8, 71–77. doi: 10.1016/0262-8856(90)90059-E

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions .on Pattern . Analysis and. Machine . Intelligence.* 22, 1330–1334. doi: 10.1109/34.888718

Zivkovic, Z. (2004). "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 2:28-31 Vol.2,* (Cambridge: IEEE)., doi: 10.1109/ICPR.2004.1333992

Check for updates

Lower Limb Kinematics Trajectory Prediction Using Long Short-Term Memory Neural Networks

Abdelrahman Zaroug[1]*, Daniel T. H. Lai[1,2], Kurt Mudie[3] and Rezaul Begg[1]*

[1] *Institute for Health and Sport, Victoria University, Melbourne, VIC, Australia,* [2] *College of Engineering and Science, Victoria University, Melbourne, VIC, Australia,* [3] *Defence Science and Technology Group, Melbourne, VIC, Australia*

This study determined whether the kinematics of lower limb trajectories during walking could be extrapolated using long short-term memory (LSTM) neural networks. It was hypothesised that LSTM auto encoders could reliably forecast multiple time-step trajectories of the lower limb kinematics, specifically linear acceleration (LA) and angular velocity (AV). Using 3D motion capture, lower limb position–time coordinates were sampled (100 Hz) from six male participants (age 22 ± 2 years, height 1.77 ± 0.02 m, body mass 82 ± 4 kg) who walked for 10 min at 5 km/h on a 0% gradient motor-driven treadmill. These data were fed into an LSTM model with a sliding window of four kinematic variables with 25 samples or time steps: LA and AV for thigh and shank. The LSTM was tested to forecast five samples (i.e., time steps) of the four kinematic input variables. To attain generalisation, the model was trained on a dataset of 2,665 strides from five participants and evaluated on a test set of 1 stride from a sixth participant. The LSTM model learned the lower limb kinematic trajectories using the training samples and tested for generalisation across participants. The forecasting horizon suggested higher model reliability in predicting earlier future trajectories. The mean absolute error (MAE) was evaluated on each variable across the single tested stride, and for the five-sample forecast, it obtained 0.047 m/s^2 thigh LA, 0.047 m/s^2 shank LA, 0.028 deg/s thigh AV and 0.024 deg/s shank AV. All predicted trajectories were highly correlated with the measured trajectories, with correlation coefficients greater than 0.98. The motion prediction model may have a wide range of applications, such as mitigating the risk of falls or balance loss and improving the human–machine interface for wearable assistive devices.

Keywords: LSTM, neural networks, machine learning, forecasting, gait, walking

INTRODUCTION

An increasingly useful application of machine learning (ML) is in predicting features of human actions. If it can be shown that algorithm inputs related to actual movement mechanics can predict a limb or limb segment's future trajectory, a range of apparently intractable problems in movement science could be solved. One such problem is how to anticipate movement characteristics that can

predict the risk of tripping, slipping or balance loss. Previous work has investigated balance control using wearable sensors to estimate the body's centre of mass (CoM) trajectory (Fuschillo et al., 2012). The Internet of things (IoT) has also created a new paradigm of algorithms and systems to predict and subsequently apply interventions to prevent falls (Rubenstein, 2006; Tao and Yun, 2017; Nait Aicha et al., 2018). Perhaps the most valuable motion-prediction application is in the design and control of wearable assistive devices, such as prostheses, bionics and exoskeletons, in which smart algorithms can ensure safer, more efficient integration of the assistive device with the user's natural limb and body motion (Lee et al., 2017; Rupal et al., 2017).

Previous computational methods have investigated motion trajectory prediction, using position-time inputs and their derivatives (velocity and acceleration). Lower limb trajectory prediction has been implemented in rehabilitation robotics (Duschau-Wicke et al., 2009). Using inverse dynamics, Wang et al. (2011) designed a model for foot trajectory generation using a predefined pelvic trajectory and line fitting 10 data points from a single gait cycle. Also using inverse dynamics, Ren et al. (2007) predicted all segment motions and ground reaction forces from the average forward velocity gait, double stance duration and gait cycle period. Another technique was implemented in the Lower Extremity Powered Exoskeleton (LOPES) device to emulate the trajectories from a healthy limb to the impaired limb (Vallery et al., 2008). Prediction of the lower limb joint angles future trajectory that effectively leads to foot events timing was also investigated in the works of Aertbeliën and De Schutter (2014) and Tanghe et al. (2019) using probabilistic principal component analysis (PPCA).

Recent methods implemented ML algorithms such as artificial neural networks (ANNs) to identify subject gait trajectories to recognise neurological as well as pathological gait patterns (Alaqtash et al., 2011; Horst et al., 2019). Artificial neural networks were also used to improve user intention detection in wearable assistive devices (Jung et al., 2015; Islam and Hsiao-Wecksler, 2016; Moon et al., 2019; Trigili et al., 2019). A variation of ANNs called generalised regression neural networks (GRNNs) was found to be capable of predicting lower limb joint angles (hip, knee and ankle) from the linear acceleration (LA) and angular velocity (AV) of foot and shank segments (Findlow et al., 2008), or from subject gait and anthropomorphic parameters (Luu et al., 2014). Recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which are classes of ANNs, were able to classify human motions and activities (Murad and Pyun, 2017; Han et al., 2019).

Long short-term memory (LSTM) neural networks are a subclass of RNNs, and they have proven success in modelling a wide range of sequence problems, including human activity recognition (Ordóñez and Roggen, 2016), gait diagnosis (Zhao et al., 2018), falls prediction (Nait Aicha et al., 2018) and gait event detection (Kidziński et al., 2019). Long short-term memory autoencoder is an architecture of LSTM that has been implemented in an array of applications such as language translation (Ding et al., 2018) and in forecasting of video frames (Srivastava et al., 2015), weather (Gangopadhyay et al., 2018; Reddy et al., 2018; Poornima and Pushpalatha, 2019), traffic

flow (Park et al., 2018; Wei et al., 2019) and stock prices (Li et al., 2018).

Given the potential of lower limb trajectory prediction, no previous work was found that utilised ML techniques to predict future lower limb trajectories using simulated inertial measurement data, which could have a profound impact on human movement science. Simulated measurement data such as the kinematics output from inertial measurement units (IMUs; i.e., LA and AV) offer the opportunity to transcend a predictive model outside the laboratory settings. The aim of this work was to determine whether the kinematics of lower limb trajectories during walking could be reliably extrapolated using LSTM autoencoder neural networks. It was hypothesised that an LSTM autoencoder could reliably forecast multiple time-step trajectories of the lower limb kinematics.

MATERIALS AND METHODS

Collection Protocol

Ethics approval was granted by the Department of Defence and Veterans' Affairs Human Research Ethics Committee and Victoria University Human Research Ethics Committee (Protocol 852-17). All participants signed a consent form and volunteered freely to participate. Walking data were obtained from six male participants (22 ± 2 years old, 1.77 ± 0.02 m in height, 82 ± 4 kg in mass) who walked for 10 min at 5 km/h on a 0% gradient treadmill. A set of 25 retroreflective markers were attached to each participant in the form of clusters (Findlow et al., 2008). Each cluster comprised a group of individual markers that represent a single body segment (e.g., shank). That included left and right foot (three markers), left shank (four markers), right shank (five markers), left thigh (three markers), right thigh (four markers) and pelvis (three markers). The 3D position of each cluster was tracked using a 14-camera motion analysis system (Vicon Bonita, Version 2.8.2) at 150 Hz. Virtual markers were also established to calibrate the position and orientation of the lower body skeletal system (Garofolini, 2019). Three-dimensional ground reaction force and moment data were collected from a force-plate instrumented treadmill (Advanced Mechanical Technology, Inc., Watertown, MA, United States) at 1,500 Hz.

Dataset Processing

Recorded 3D positional and force data were processed using Visual 3D (C-motion, Inc, Version 6) to obtain LA and AV. In Visual 3D (**Figure 1**), the data were firstly filtered using a low-pass digital filter with a 15-Hz cut-off frequency and normalised to mean 0 and standard deviation 1 using standard scores (z-scores), preserving the original data properties. Secondly, raw AV was obtained as the derivative of Euler/Cardan angles (C-motion, 2015), and the raw LA was generated by the double derivative of segment linear displacement using built-in pipeline commands (Hibbeler, 2007). These data (LA and AV) simulated the kinematic outputs from body-mounted IMUs widely used in wearable assistive devices, monitoring lower limb kinematics (Santhiranayagam et al., 2011; Lai et al., 2012), controlling powered actuators (Lee et al., 2017) and recognising human
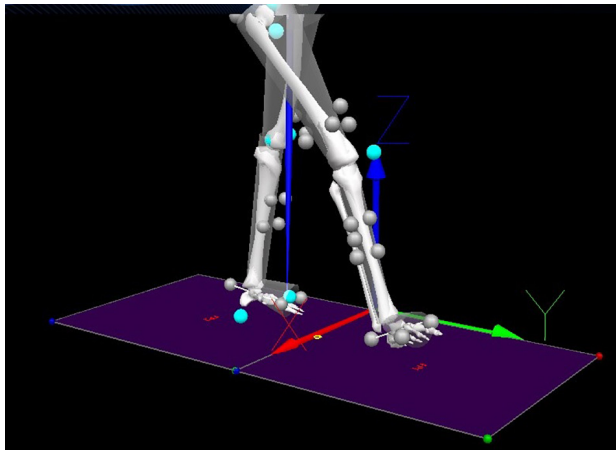
FIGURE 1 | Components (x,y,z) definition and markers setup. Grey balls are retroreflective markers. Turquoise balls are virtual markers.

actions (Van Laerhoven and Cakmakci, 2000; Jimenez-Fabian and Verlinden, 2012; Koller et al., 2016).

As shown in **Figure 1**, the main direction of movements included the translation along the Y-axis (i.e., LA) and the rotation along the X-axis (i.e., AV), which were used for LSTM prediction, resulting in four predictor variables: (i) Y_1 thigh LA, (ii) Y_2 shank LA, (iii) X_3 thigh AV and (iv) X_4 shank AV. The thigh segment was defined as the reference frame to the shank, and the shank segment was defined as the reference frame to the thigh (**Figure 2**).

Dataset Description

The data were divided into training and testing sets. The training set comprised 2,665 strides from five participants that included four kinematic feature variables (Y_1, Y_2, X_3, X_4) (N-columns) and 453,060 samples or time steps (M-rows) for each variable. To attain generalisation, a testing set was used that comprised of a single stride from the sixth participant with the four feature variables and 170 samples for each variable.

Time Series Transformation to a Supervised Learning Problem

The inputs to the LSTM were four parallel feature variables and the outputs were the successive four parallel feature variables. Prior to feeding into the LSTM model, the *MxN* training and testing datasets were transformed to a 3D dataset using a sliding window technique (Banos et al., 2014). The sliding window comprised of an input window, an output window and a sliding size. The input window consists of M samples and N features, so as the output window. The input window is the input data to the LSTM model, and the output window is the future prediction output from the LSTM model. The sliding size is how much of M samples that both the input and the output windows are sliding forward with (see **Figure 3**). The sliding size (M samples) was always equal to the output size.

Recurrent Neural Networks

While multiple layer perceptrons (MLPs) consider all inputs as independent, RNNs are designed to work with time series data (Ordóñez and Roggen, 2016). RNNs are a class of ANN architecture designed specifically to model sequence problems and exploit the temporal correlations between input data samples (Elman, 1990; Murad and Pyun, 2017). It contains feedback connections between each of its units, which enables the network to relate all the previous inputs to its outputs (**Figure 4**).

The forward pass equations from the inputs to the outputs of the RNN are given as follows.

For the hidden units:

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1} \qquad (1)$$

and differentiable activation functions are then applied:

$$b_h^t = \theta_h(a_h^t) \qquad (2)$$

The network input to output units:

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t \qquad (3)$$

where

a_h^t is the sum of inputs to unit h at time t, b_h^t is the activation of unit h at time t, θ_h is the non-linear and differentiable activation function of unit h, a_k^t is the sum of all inputs to output unit k at time t, x_i^t is the input i at time t, w_{ih} is the connection weights between input unit i and hidden unit h, $w_{h'h}$ is the connection weights between the previous hidden state h' and itself h and w_{hk} is the connection weights between the hidden state h and the output unit k. Bias was neglected for simplicity.

LSTM Networks

As the input data propagates through the standard RNN's hidden connections to the output units, it either slowly attenuates or amplifies exponentially, referred to, respectively, as vanishing or exploding gradients (Bengio et al., 1994; Hochreiter et al., 2001). The problems with this approach are that the vanishing gradient prevents the network from learning long-term dependencies and the exploding gradient leads to weights oscillation. These difficulties have been addressed using gradient norm clipping to tackle the exploding gradient and a soft constraint to deal with the vanishing gradient (Pascanu et al., 2013). The LSTM design addresses these problems by maintaining a memory cell C (**Figure 5**) that enables the network to retain information over a longer period by using an explicit gating mechanism (Hochreiter and Schmidhuber, 1997; Graves, 2012; Karpathy et al., 2015).

Each LSTM cell has an input gate, forget gate, and output gate. The input gate dictates the information used to update the memory state, and the forget gate decides which information to discard or remove from the cell. The final gate specifies the information to output based on the cell input and memory. All gates are designed such that information is exchanged from inside and outside the block (**Figure 5**). Furthermore, each
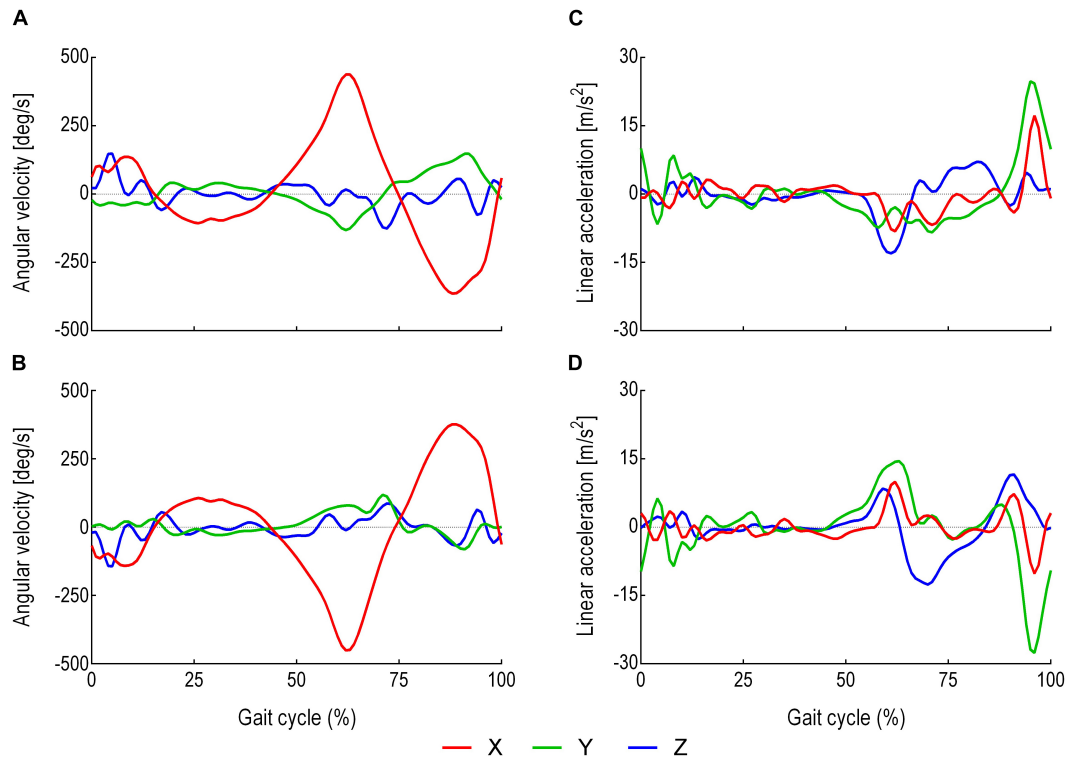
FIGURE 2 | Average thigh and shank LA and AV within a stride. A stride was defined as the interval between two successive heel strikes of the same foot (De Lisa, 1998). **(A)** Thigh three-dimensional AV (direction of the rotation around the X-axis). **(B)** Shank three-dimensional AV (direction of the rotation around the X-axis). **(C)** Thigh three-dimensional LA (direction of the progression along the Y-axis). **(D)** Shank three-dimensional LA (direction of the progression along the Y-axis). Red is the X-axis. Green is the Y-axis. Blue is the Z-axis.

memory block contains three peephole-weighted connections (dotted lines in **Figure 5**), which are the input weight $w_{c\iota}$, the output weight $w_{c\omega}$ and the memory state $w_{c\phi}$. The functions f, g and h are usually tanh or logistic sigmoid activation functions (Graves, 2012). Below are the network equations (Graves, 2012) that govern the LSTM architecture used:

Input gates:

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota}x_i^t + \sum_{h=1}^{H} w_{h\iota}b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota}s_c^{t-1} \qquad (4)$$

$$b_\iota^t = f(a_\iota^t) \qquad (5)$$

Forget gates:

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi}x_i^t + \sum_{h=1}^{H} w_{h\phi}b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi}s_c^{t-1} \qquad (6)$$

$$b_\phi^t = f(a_\phi^t) \qquad (7)$$

Cells:

$$a_c^t = \sum_{i=1}^{I} w_{ic}x_i^t + \sum_{h=1}^{H} w_{hc}b_h^{t-1} \qquad (8)$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t) \qquad (9)$$

Output gates:

$$a_c^t = \sum_{i=1}^{I} w_{iw}x_i^t + \sum_{h=1}^{H} w_{hw}b_h^{t-1} + \sum_{c=1}^{C} w_{cw}s_c^t \qquad (10)$$

$$b_w^t = f(a_w^t) \qquad (11)$$

Cell outputs:

$$b_c^t = b_w^t h(s_c^t) \qquad (12)$$

where w_{ij} is the weight of the connection from unit i to unit j; a_j^t is the network input to unit j at time t; b_j^t is the activation of unit j at time t; ι, ϕ, ω respectively stand for the input gate, the forget gate and the output gate; C is the memory cell; $w_{c\iota}$, $w_{c\phi}$, $w_{c\omega}$ are peephole weights; s_c^t is the state of cell C at time t; f is the input, output and forget gates activation function; g and h are the cell input and output activations, respectively; I is the number of inputs; H is the number of cells in the hidden layer; and index h is the cell outputs from other blocks in the hidden layer. Bias was neglected for simplicity.

Design of the LSTM Model

The implemented model was based on the autoencoder LSTM, a neural network architecture composed of an encoder and a decoder (Ding et al., 2018). The encoder encodes the input variable length vector into a fixed length feature

Shank Angular Velocity (X component)

FIGURE 3 | Sliding window illustration example using the normalised shank angular velocity X-axis component (one feature). The window in this model is 25 samples and four features and the prediction outputs are five samples of four features.

FIGURE 4 | Unfolded structure of the Recurrent Neural Network.
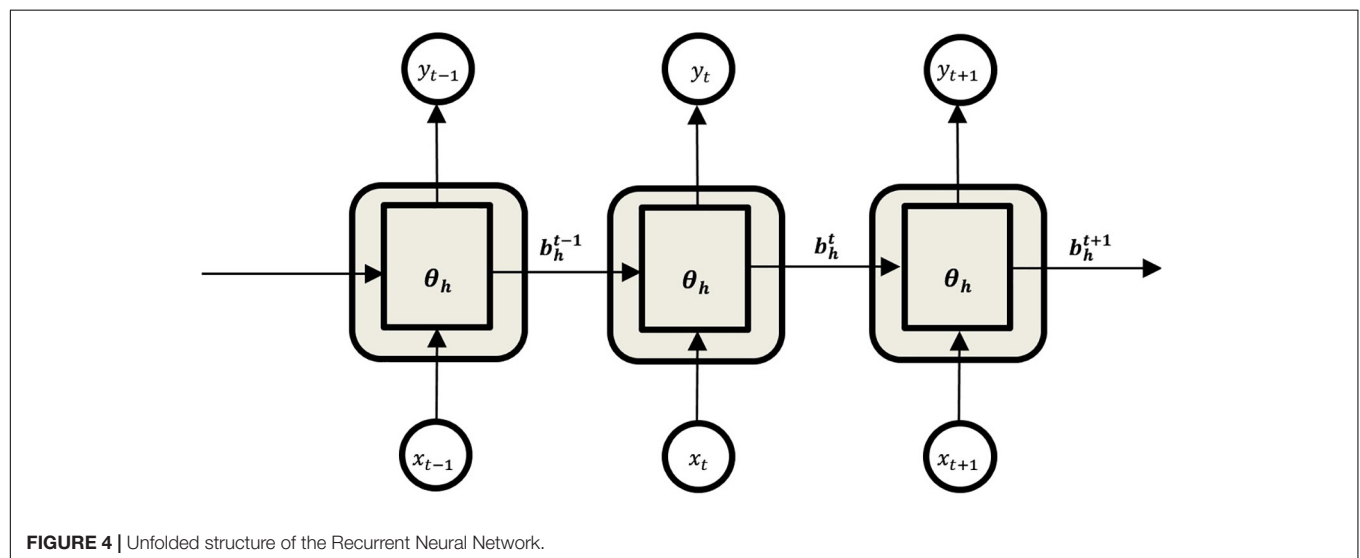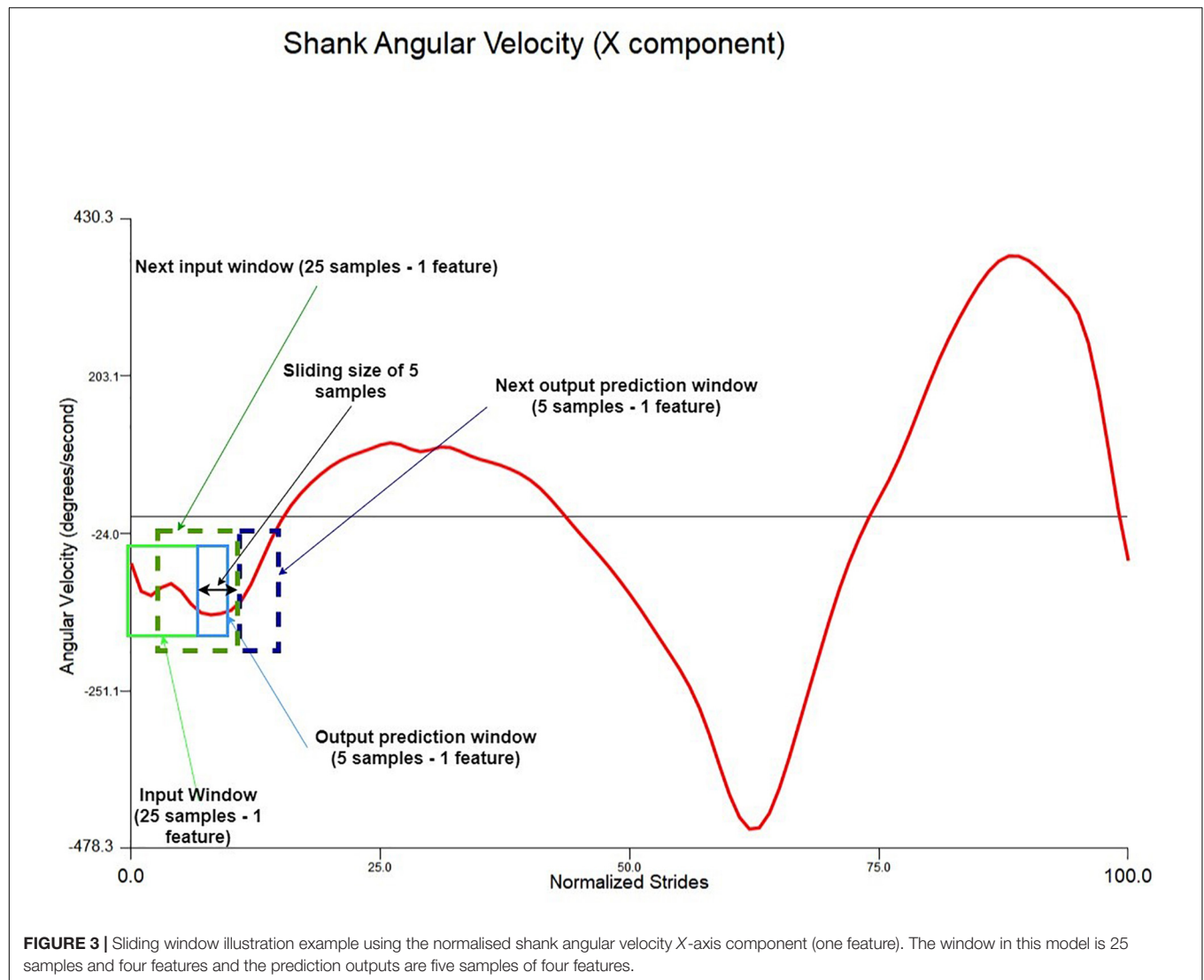
vector that captures the attributes of the variable length vector. The LSTM decoder decodes the encoded fixed length feature vector back into a variable length vector (**Figure 6**). The final layer is a fully connected dense (feedforward) mechanism for outputting predictions. The network weights and biases were updated at the end of each batch using an adaptive moment estimation (Adam) optimisation algorithm (Kingma and Ba, 2014) with mean absolute error (MAE) as an optimisation criterion. A single batch consists of 100 input/output windows. The activation for all LSTM layers was set to a rectified rectilinear unit (ReLU) activation function (Nair and Hinton, 2010). The LSTM autoencoder model was implemented in Google Colab as well as Amazon Web Services (AWS) using Python 3 (Libraries: Keras, Numpy, Pandas and Scikit learn).

Evaluation Metrics

To evaluate the network quality, three parameters were considered to calculate how closely the network predicted variable trajectories $\widehat{y_j}$ (Y_1, Y_2, X_3, X_4) were to the actual variable trajectories y_j (Y_1, Y_2, X_3, X_4) across the n samples:

1. MAE given as:

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| y_j - \widehat{y_j} \right| \qquad (13)$$

2. Mean squared error (MSE) given as:

$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \widehat{y_j} \right)^2 \qquad (14)$$
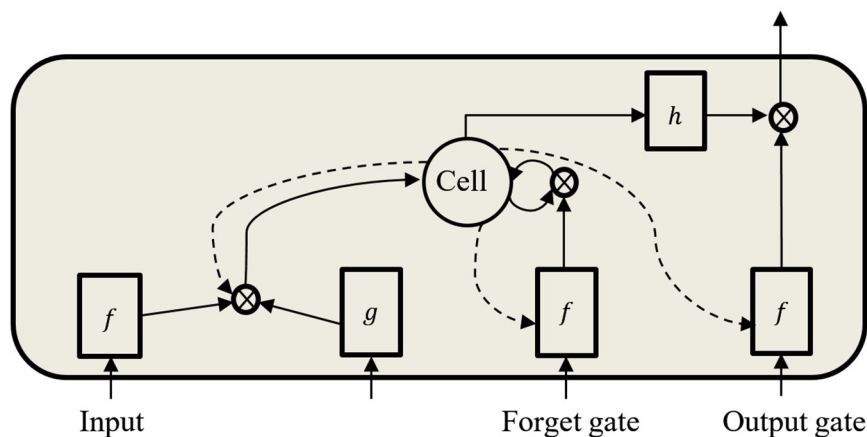
FIGURE 5 | Standard LSTM memory cell with peephole connections.
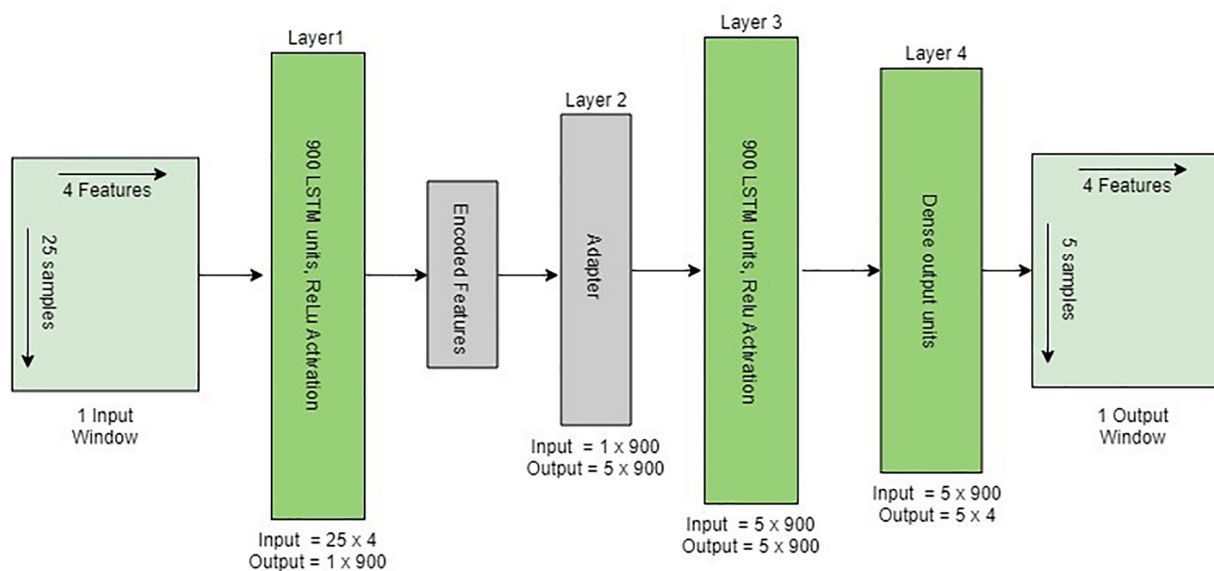
FIGURE 6 | Structure of the implemented encoder–decoder LSTM architecture given one input window. The adapter converts the 2D encoded features into 3D output to be adopted by LSTM. The last layer is a fully connected dense layer for outputting one window prediction.

3. Correlation coefficient (CC) given as:

$$P = \frac{cov(y, \hat{y})}{std\,(y) \times std(\hat{y})} \qquad (15)$$

where $std()$ is the standard deviation and $cov(y, \hat{y})$ is the covariance between variables y and $\hat{y}$.

RESULTS

Using the sparse grid search approach, the model's hyperparameters were tuned to determine the optimum model design (least MAE), including the number of epochs, batch size, layers and cells. The optimum model was then trained for 50 epochs (repetitions), and performance evaluated on the test set using MAE, MSE and the CC. The test set was a single stride that consisted of 170 samples. Initial 25 samples were used from the preceding cycle in order to start predicting the trajectories of the single stride.

Model Performance With Different Input Window Sizes

The size of the input window was varied eight times at five sample intervals (5–40 samples) to demonstrate the optimum input window size (least error). The output sliding window was fixed to five samples prediction. The model performance is shown in **Figure 7** where the impact of each input window size on the prediction of each variable is computed.

Model Performance With Five Samples Prediction

This sliding window comprised of 25 samples input and 5 samples prediction output. Results were given in two analyses: (i) predicted versus actual trajectories including the absolute error (AE) for each sample in the first output window (**Figure 8**) and for the whole gait cycle (**Figure 9**) and (ii) performance metrics (MAE, MSE and CC) for the first window of five samples (**Table 1**) and for all windows combined (**Table 2**).

Model Performance With 10 Samples Prediction

This sliding window comprised of 25 samples input, 10 samples prediction output. **Figure 10** illustrates the results as predicted versus the actual trajectories including the AE for each sample in the first output window, whereas **Figure 11** displays the results for the whole gait cycle. Performance metrics (MAE, MSE and CC) for the first window of 10 samples are presented in **Table 3** and for all windows combined in **Table 4**.

DISCUSSION

Our aim was to develop and evaluate an LSTM autoencoder model to predict the trajectories of four kinematic variables (Y_1, Y_2, X_3, X_4), simulating the output from wearable sensors (IMU). The predicted kinematic feature variables, LA and AV, for the shank and thigh were reliably predicted up to 10 samples or

FIGURE 7 | Model performance with different input window sizes. Red is MAE. Blue is MSE. **(A)** Thigh LA (Y_1). **(B)** Shank LA (Y_2). **(C)** Thigh AV (X_3). **(D)** Shank AV (X_4).
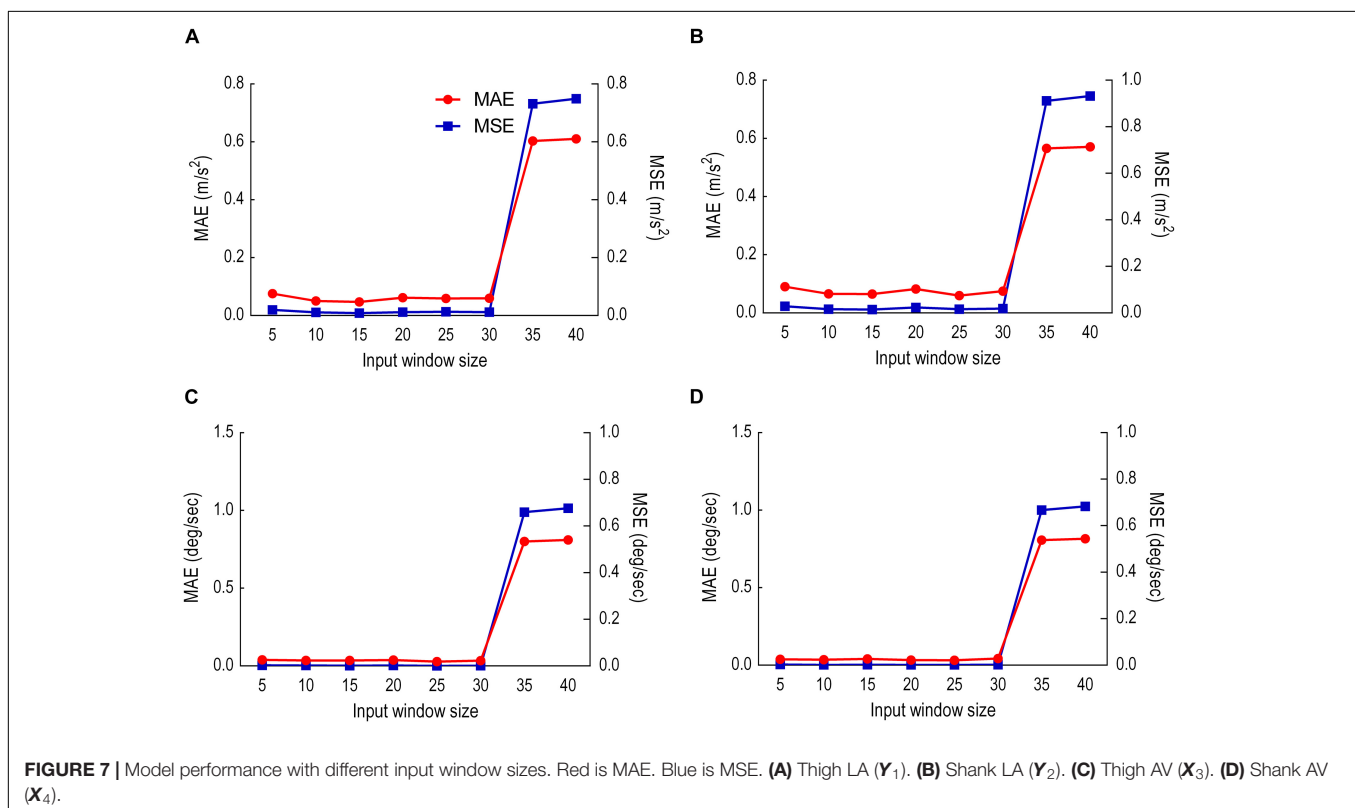
FIGURE 8 | Model performance for the first window, showing predicted trajectories (green) and actual trajectories (red). Columns represent the absolute error (AE) for the five predicted samples. **(A)** Thigh LA Y_1. **(B)** Shank LA Y_2. **(C)** Thigh AV X_3. **(D)** Shank AV X_4
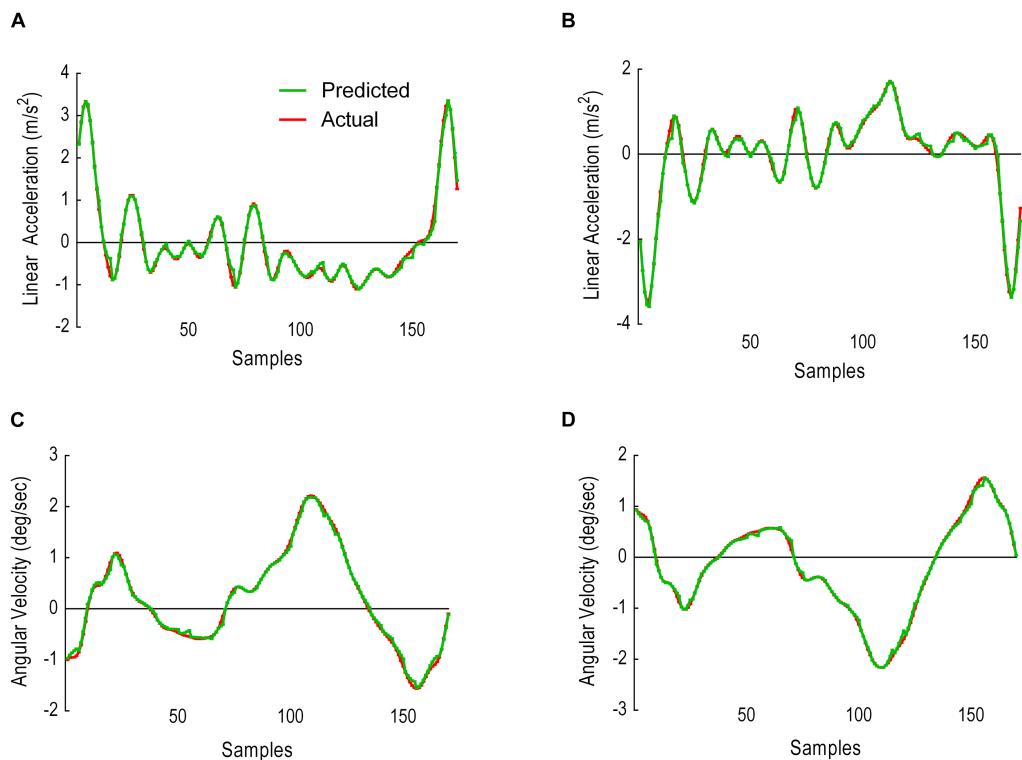
FIGURE 9 | Model performance over the entire gait cycle when five samples prediction window is used. The figure shows predicted trajectories (orange) and actual trajectories (blue). **(A)** Thigh LA Y_1. **(B)** Shank LA Y_2. **(C)** Thigh AV X_3. **(D)** Shank AV X_4.

TABLE 1 | Model performance for predicting the first five stride samples.

Feature	MAE	MSE	CC
Y_1	0.125 m/s²	0.019 m/s²	0.99
Y_2	0.133 m/s²	0.022 m/s²	0.99
X_3	0.032 deg/s	0.001 deg/s	0.98
X_4	0.033 deg/s	0.001 deg/s	0.99

TABLE 2 | Model performance for predicting the complete stride using an input window size of 25 samples and an output window size of 5 samples.

Feature	MAE	MSE	CC
Y_1	0.047 m/s²	0.006 m/s²	0.99
Y_2	0.047 m/s²	0.006 m/s²	0.99
X_3	0.028 deg/s	0.001 deg/s	0.99
X_4	0.024 deg/s	0.001deg/s	0.99

time steps, i.e., up to 60 ms in the future. A 60-ms prediction of future trajectories adds a feedforward term to an assistive device controller rather than being reactive and predominantly relying on feedback terms (i.e., sensory information; Tanghe et al., 2019). This enables the assistive device to adapt to changes in human gait, allowing smoother synchronisation with user intentions and minimising interruptions when the user changes their movement pattern (Elliott et al., 2014; Zhang et al., 2017; Ding et al., 2018; Zaroug et al., 2019). A known future trajectory might also monitor the risk of balance loss, tripping and falling, in which impending incidents can be remotely reported for early intervention (Begg and Kamruzzaman, 2006; Begg et al., 2007;

Nait Aicha et al., 2018; Hemmatpour et al., 2019; Naghavi et al., 2019). Since 60 ms falls in the range of slow (60–120 ms) and fast (10–50 ms) twitch motor units (Winter, 2009), this would enable wearable devices such as IMUs to alert (e.g., by audio/visual signal) an elderly user about an imminent risk of tripping and potentially gives them a chance to adjust their gait accordingly.

In contrast to the 1- to 2-s window for human activity recognition proposed by Banos et al. (2014), no window has previously been suggested for forecasting human movement trajectories (Banos et al., 2014). In addressing this limitation, the present project input and output sliding windows were tested to discover the optimum prediction model. The input window was varied from 5 to 40 samples, whereas the output window was fixed at 5 samples during each test. Results showed that both MAE and MSE increased after 25 samples for all variables except for the thigh LA Y_1 in which 15 samples scored lowest. Due to the majority score, 25 samples were fixed, and the output window size manipulated between 5 and 10 samples. Prediction error MAE and MSE gradually increased across the first 5 and 10 sample prediction windows, indicating better prediction early in the stride cycle. This prediction horizon suggests that an output window exceeding five samples may not be sufficiently reliable for forecasting gait trajectories. LA-predicted trajectories began to deviate earlier than AV, possibly due to the double derivative generating a noisier signal.

Across the stride cycle, an output window of 5 samples showed better model performance (lower MAE scores) than the 10-sample output window, particularly when there is less noise in the predicted signal for all variables. Predictions of five samples
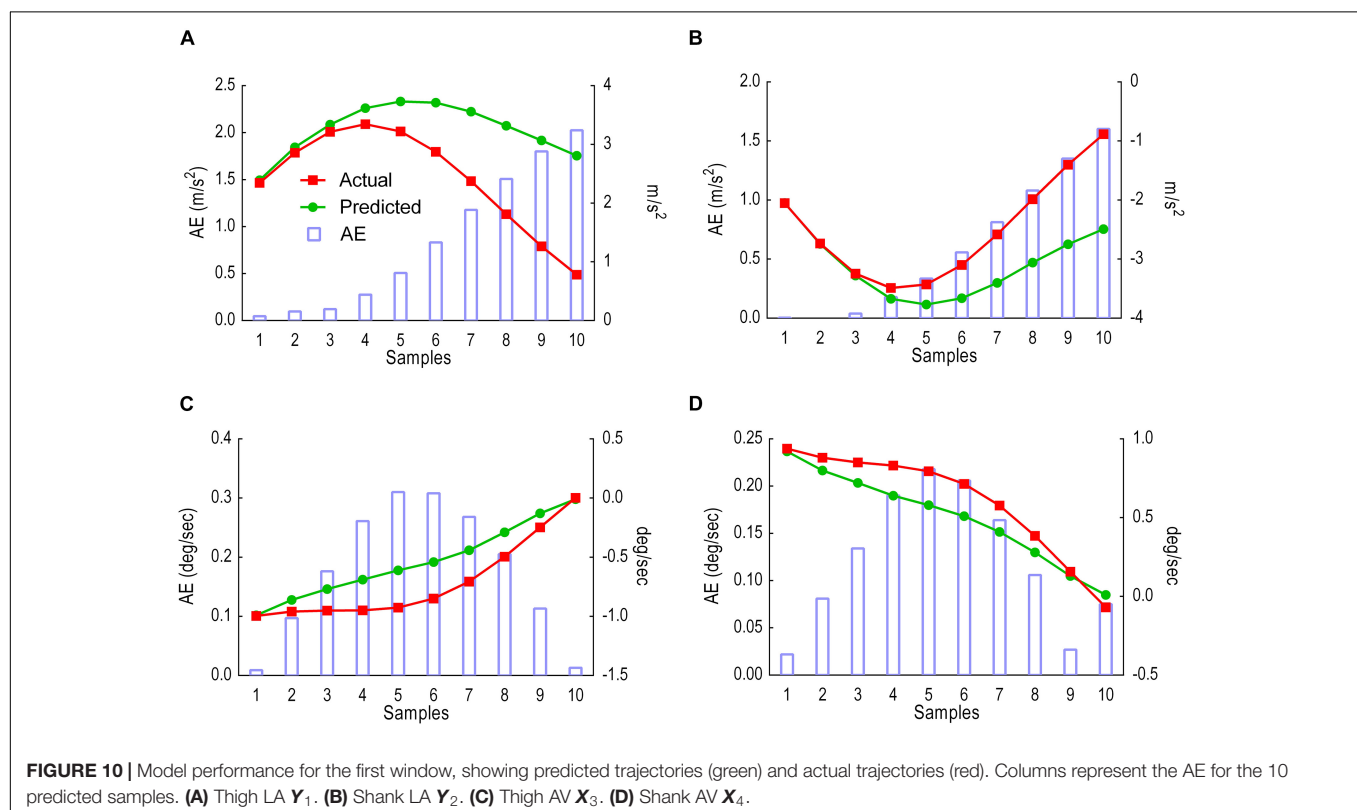
FIGURE 10 | Model performance for the first window, showing predicted trajectories (green) and actual trajectories (red). Columns represent the AE for the 10 predicted samples. **(A)** Thigh LA Y_1. **(B)** Shank LA Y_2. **(C)** Thigh AV X_3. **(D)** Shank AV X_4.

FIGURE 11 | Model performance over the entire gait cycle when 10 samples prediction window is used. The figure shows predicted trajectories (orange) and actual trajectories (blue). **(A)** Thigh LA Y_1. **(B)** Shank LA Y_2. **(C)** Thigh AV X_3. **(D)** Shank AV X_4.
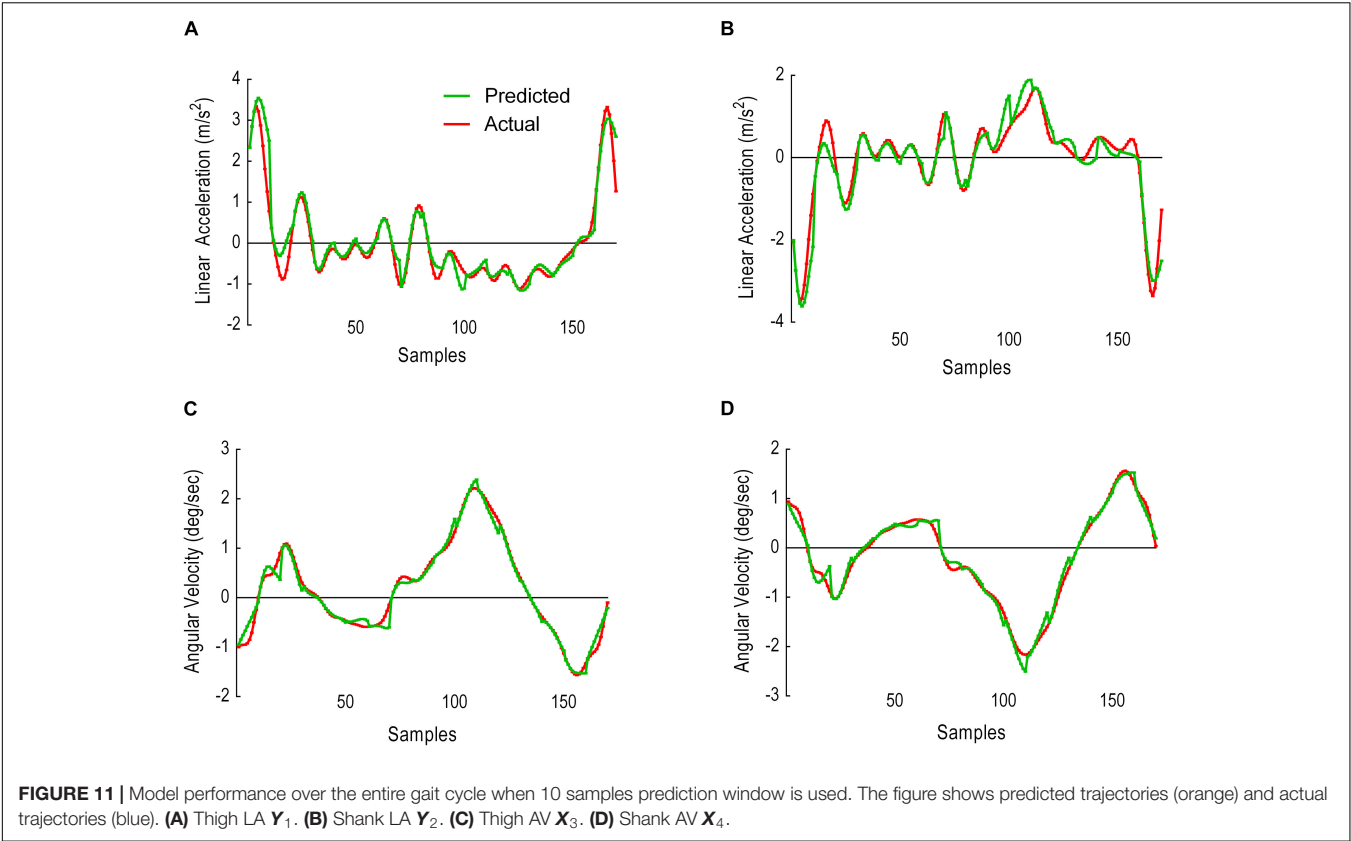
for all variables achieved high CC (0.99) and maintained below MAE 0.048 deg/s and 0.029 m/s². These result parameters are different from those of earlier work (Findlow et al., 2008; Luu et al., 2014). The difference is in the type of predicted data (lower limb joint angles of the hips, knees and ankles) and in the type of output, which was not a forecast, but rather a prediction of joint angles from the LA and AV of the lower limb segments. Nonetheless, the work presented in this paper showed higher CC values than the earlier works (Findlow et al., 2008; Luu et al., 2014) at the intersubject test. Overall, the LSTM model was able

to learn the trajectories and generalise across participants. This generalisation is invaluable to adapt algorithm performance to a wider population in assistive devices, particularly when each user responds differently to the same device (Zhang et al., 2017).
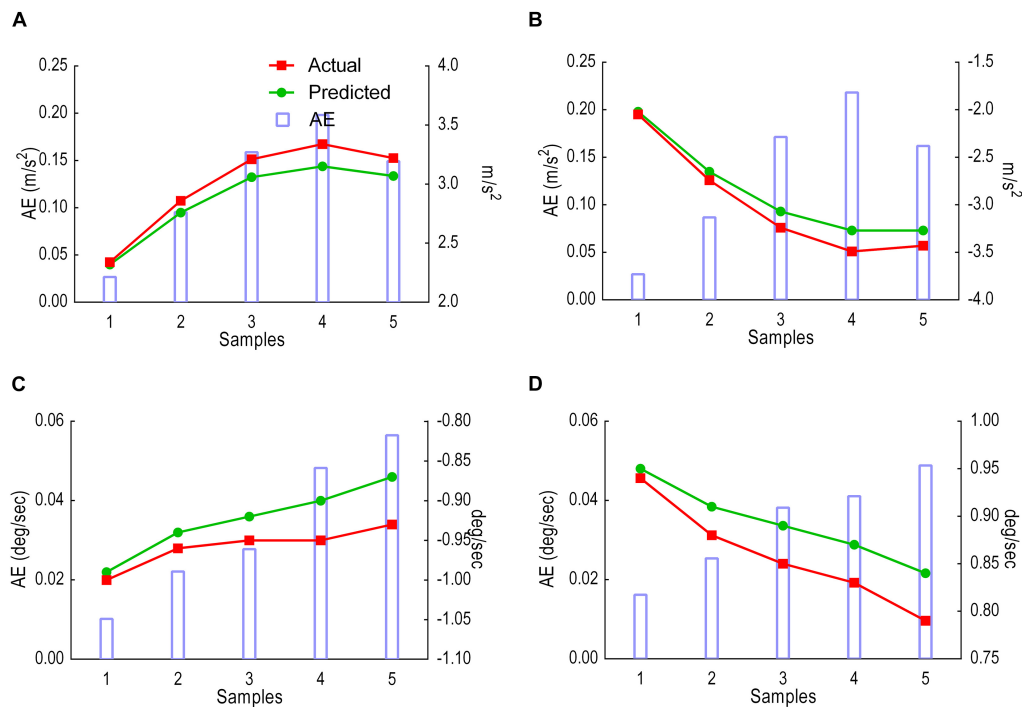
This study was limited to the walking movement with a 60-ms prediction horizon and healthy participants walking at 5 km/h. The speed was imposed to report the feasibility of whether lower limb future trajectories are predictable. In future work, the model would be developed to accommodate a higher gait variance from more participants and other populations, such as female, older adults and individuals with gait disorders walking at their preferred as well as slower and faster speeds (Winter, 1991). More participants (i.e., stride examples) would potentially improve the model performance to predict trajectories above 60 ms and also provide a more comprehensive validation set, a strategy to find the optimum number of epochs and avoid model overfitting (Graves, 2013). The LSTM autoencoder can be made flexible by automating the input/output window size depending on the detected human activity, which revamps the LSTM capacity to recognise a wider range of human action transitions, such as slow to fast walking. Although LSTM autoencoders described here were able to learn and predict future data points, further research is needed to explore other LSTM architectures, such as bi-directional LSTM (Graves and Schmidhuber, 2005). Bi-directional LSTM can be useful in forward and backward modelling of sequential data, giving further insights into sequential pattern modelling (Liu and Guo, 2019; Zhang et al., 2019).

TABLE 3 | Model performance for predicting the first 10 stride samples.

Feature	MAE	MSE	CC
Y_1	0.839 m/s²	1.206 m/s²	0.52
Y_2	0.596 m/s²	0.667 m/s²	0.75
X_3	0.176 deg/s	0.042 deg/s	0.94
X_4	0.122 deg/s	0.019 deg/s	0.96

TABLE 4 | Model performance for predicting the complete stride using an input window size of 25 samples and an output window size of 10 samples.

| Feature	MAE	MSE	CC
Y_1	0.170 m/s²	0.096 m/s²	0.96
Y_2	0.202 m/s²	0.096 m/s²	0.96
X_3	0.079 deg/s	0.015 deg/s	0.98
X_4	0.086 deg/s	0.014 deg/s	0.98

CONCLUSION

This study confirmed the possibility of predicting the future trajectories of human lower limb kinematics during steady-state walking, i.e., thigh AV, shank AV, thigh LA and shank LA. An input window of 25 samples and an output window of 5 samples were found to be the optimum sliding window sizes for future trajectories prediction in LSTM. The LSTM model prediction horizon was better able to forecast the earlier sample trajectories and was also able to learn trajectories across different participants. Further work is required to systematically investigate the effects of tuning the model's hyperparameters, including layers and cells, optimisation algorithms and learning rate. Future work could focus on automating input/output window size and using predicted kinematics to identify discrete gait cycle events such as heel strike and toe-off (Kidziński et al., 2019). Long short-term memory methods for human movement prediction have applications to balance loss, falls prevention and controlling of assistive devices.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Associate Professor Deborah Zion Chair of Victoria University Human Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AZ wrote the manuscript and coded the ML model. AZ, DL and RB contributed to research and ML model design and analysis. KM and RB designed the biomechanics experiment. KM and AZ collected and analysed the biomechanics data. All authors provided critical feedback on the manuscript and read and approved the final manuscript.

FUNDING

REFERENCES

Aertbeliën, E., and De Schutter, J. (2014). "Learning a predictive model of human gait for the control of a lower-limb exoskeleton," in *Proceedings of the 5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics* (Piscataway, NJ: IEEE). doi: 10.1109/TCYB.2020.2972582

Alaqtash, M., Sarkodie-Gyan, T., Yu, H., Fuentes, O., Brower, R., and Abdelgawad, A. (2011). "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms," in *Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Piscataway, NJ: IEEE). doi: 10.1109/IEMBS.2011.6090063

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., and Rojas, I. (2014). Window size impact in human activity recognition. *Sensors* 14, 6474–6499. doi: 10.3390/s140406474

Begg, R., Best, R., Dell'Oro, L., and Taylor, S. (2007). Minimum foot clearance during walking: strategies for the minimisation of trip-related falls. *Gait Posture* 25, 191–198. doi: 10.1016/j.gaitpost.2006.03.008

Begg, R., and Kamruzzaman, J. (2006). Neural networks for detection and classification of walking pattern changes due to ageing. *Aust. Phys. Eng. Sci. Med.* 29, 188–195. doi: 10.1007/bf03178892

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

C-motion (2015). *Joint Velocity*. Available online at: https://c-motion.com/v3dwiki/index.php/Joint_Velocity (accessed January 2020).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv [Preprint]*. doi: 10.3115/v1/D14-1179

De Lisa, J. A. M. D. (1998). *Gait Analysis in Science and Rehabilitation*. Available online at: https://ia800206.us.archive.org/6/items/gaitanalysisinsc00joel/gaitanalysisinsc00joel.pdf

Ding, Y., Kim, M., Kuindersma, S., and Walsh, C. J. (2018). Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. *Sci. Robot.* 3:eaar5438. doi: 10.1126/scirobotics.aar5438

Duschau-Wicke, A., von Zitzewitz, J., Caprez, A., Lunenburger, L., and Riener, R. (2009). Path control: a method for patient-cooperative robot-aided gait rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.* 18, 38–48. doi: 10.1109/TNSRE.2009.2033061

Elliott, G., Marecki, A., and Herr, H. (2014). Design of a clutch–spring knee exoskeleton for running. *J. Med. Dev.* 8:031002.

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211.

Findlow, A., Goulermas, J., Nester, C., Howard, D., and Kenney, L. (2008). Predicting lower limb joint kinematics using wearable motion sensors. *Gait Posture* 28, 120–126. doi: 10.1016/j.gaitpost.2007.11.001

Fuschillo, V. L., Bagalà, F., Chiari, L., and Cappello, A. (2012). Accelerometry-based prediction of movement dynamics for balance monitoring. *Med. Biol. Eng. Comput.* 50, 925–936. doi: 10.1007/s11517-012-0940-6

Gangopadhyay, T., Tan, S. Y., Huang, G., and Sarkar, S. (2018). "Temporal attention and stacked LSTMs for multivariate time series prediction," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, Montréal.

Garofolini, A. (2019). *Exploring Adaptability in Long-Distance Runners: Effect of Foot Strike Pattern on Lower Limb Neuro-Muscular-Skeletal Capacity*. Footscray VIC: Victoria University.

Graves, A. (2012). *Supervised Sequence Labelling With Recurrent Neural Networks*. Available online at: http://books.google.com/books (accessed September 12, 2019).

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1308.0850 (accessed September 12, 2019).

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

Han, B.-K., Ryu, J.-K., and Kim, S.-C. (2019). Context-Aware winter sports based on multivariate sequence learning. *Sensors* 19:3296. doi: 10.3390/s19153296

Hemmatpour, M., Ferrero, R., Montrucchio, B., and Rebaudengo, M. (2019). A review on fall prediction and prevention system for personal devices: evaluation and experimental results. *Adv. Hum. Comput. Interact.* 2019, 1–12.

Hibbeler, R. C. (2007). *Engineering Mechanics Dynamics SI Units*. Singapore: Pearson Education South Asia.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). *Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies, A Field Guide to Dynamical Recurrent Neural Networks*. Piscataway, NJ: IEEE Press.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Horst, F., Lapuschkin, S., Samek, W., Müller, K.-R., and Schöllhorn, W. I. (2019). Explaining the unique nature of individual gait patterns with deep learning. *Sci. Rep.* 9:2391. doi: 10.1038/s41598-019-38748-8

Islam, M., and Hsiao-Wecksler, E. T. (2016). Detection of gait modes using an artificial neural network during walking with a powered ankle-foot orthosis. *J. Biophys.* 2016:7984157. doi: 10.1155/2016/7984157

Jimenez-Fabian, R., and Verlinden, O. (2012). Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons. *Med. Eng. Phys.* 34, 397–408. doi: 10.1016/j.medengphy.2011.11.018

Jung, J.-Y., Heo, W., Yang, H., and Park, H. (2015). A neural network-based gait phase classification method using sensors equipped on lower limb exoskeleton robots. *Sensors* 15:27738. doi: 10.3390/s151127738

Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv [Preprint].* doi: 10.1142/9789813207813_0025

Kidziński, L., Delp, S., and Schwartz, M. (2019). Automatic real-time gait event detection in children using deep neural networks. *PLoS One* 14:e0211466. doi: 10.1371/journal.pone.0211466

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint].* Available online at: https://arxiv.org/abs/1412.6980 (accessed September 12, 2019).

Koller, J. R., Gates, D. H., Ferris, D. P., and Remy, C. D. (2016). "Body-in-the-loop"optimization of assistive robotic devices: a validation study," in *Proceedings of the 12th Conference Robotics: Science and Systems XII.* Available online at: http://www.roboticsproceedings.org/rss12/index.html

Lai, D. T., Taylor, S. B., and Begg, R. K. (2012). Prediction of foot clearance parameters as a precursor to forecasting the risk of tripping and falling. *Hum. Mov. Sci.* 31, 271–283. doi: 10.1016/j.humov.2010.07.009

Lee, G., Kim, J., Panizzolo, F., Zhou, Y., Baker, L., Galiana, I., et al. (2017). Reducing the metabolic cost of running with a tethered soft exosuit. *Sci. Robot.* 2:eaan6708.

Li, H., Shen, Y., and Zhu, Y. (2018). "Stock price prediction using attention-based multi-input lstM," *Proceedings of the Asian Conference on Machine Learning,* Nagoya.

Liu, G., and Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337, 325–338.

Luu, T. P., Low, K., Qu, X., Lim, H., and Hoon, K. (2014). An individual-specific gait pattern prediction model based on generalized regression neural networks. *Gait Posture* 39, 443–448. doi: 10.1016/j.gaitpost.2013.08.028

Moon, D.-H., Kim, D., and Hong, Y.-D. (2019). Development of a single leg knee exoskeleton and sensing knee center of rotation change for intention detection. *Sensors* 19:3960. doi: 10.3390/s19183960

Murad, A., and Pyun, J.-Y. (2017). Deep recurrent neural networks for human activity recognition. *Sensors* 17:2556. doi: 10.1038/s41467-020-15086-2

Naghavi, N., Miller, A., and Wade, E. (2019). Towards real-time prediction of freezing of gait in patients with parkinson's disease: addressing the class imbalance problem. *Sensors* 19:3898. doi: 10.3390/s19183898

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* Haifa.

Nait Aicha, A., Englebienne, G., van Schooten, K., Pijnappels, M., and Kröse, B. (2018). Deep learning to predict falls in older adults based on daily-life trunk accelerometry. *Sensors* 18:1654. doi: 10.3390/s18051654

Ordóñez, F., and Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16:115. doi: 10.3390/s16010115

Park, S. H., Kim, B., Kang, C. M., Chung, C. C., and Choi, J. W. (2018). "Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture," in *Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV)* (Piscataway, NJ: IEEE).

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *Proceedings of the International Conference on Machine Learning,* Long Beach.

Poornima, S., and Pushpalatha, M. (2019). Prediction of rainfall using intensified lstm based recurrent neural network with weighted linear units. *Atmosphere* 10:668.

Reddy, V., Yedavalli, P., Mohanty, S., and Nakhat, U. (2018). *Deep Air: Forecasting Air Pollution in Beijing, China.* Available online at:

https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/deep-air-forecasting_final.pdf (accessed September 12, 2019).

Ren, L., Jones, R. K., and Howard, D. (2007). Predictive modelling of human walking over a complete gait cycle. *J. Biomech.* 40, 1567–1574. doi: 10.1016/j.jbiomech.2006.07.017

Rubenstein, L. Z. (2006). Falls in older people: epidemiology, risk factors and strategies for prevention. *Age Ageing* 35(Suppl. 2), ii37–ii41. doi: 10.1093/ageing/afl084

Rupal, B. S., Rafique, S., Singla, A., Singla, E., Isaksson, M., and Virk, G. S. (2017). Lower-limb exoskeletons: research trends and regulatory guidelines in medical and non-medical applications. *Int. J. Adv. Robot. Syst.* 14:1729881417743554.

Santhiranayagam, B. K., Lai, D., Shilton, A., Begg, R., and Palaniswami, M. (2011). "Regression models for estimating gait parameters using inertial sensors," in *Proceedings of the 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing* (Piscataway, NJ: IEEE).

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. *Proceedings of the International Conference on Machine Learning,* Long Beach.

Tanghe, K., De Groote, F., Lefeber, D., De Schutter, J., and Aertbeliën, E. (2019). "Gait trajectory and event prediction from state estimation for exoskeletons during gait," in *Proceedings of the IEEE Transactions on Neural Systems and Rehabilitation Engineering* (Piscataway, NJ: IEEE). doi: 10.1109/TNSRE.2019.2950309

Tao, X., and Yun, Z. (2017). Fall prediction based on biomechanics equilibrium using Kinect. *Int. J. Distribut. Sensor Netw.* 13:1550147717703257.

Trigili, E., Grazi, L., Crea, S., Accogli, A., Carpaneto, J., Micera, S., et al. (2019). Detection of movement onset using EMG signals for upper-limb exoskeletons in reaching tasks. *J. Neuroeng. Rehabil.* 16:45. doi: 10.1186/s12984-019-0512-1

Vallery, H., Van Asseldonk, E. H., Buss, M., and Van Der Kooij, H. (2008). Reference trajectory generation for rehabilitation robots: complementary limb motion estimation. *IEEE Trans. Neural Syst. Rehabil. Eng.* 17, 23–30. doi: 10.1109/TNSRE.2008.2008278

Van Laerhoven, K., and Cakmakci, O. (2000). "What shall we teach our pants? Digest of papers," in *Proceedings of the Fourth International Symposium on Wearable Computers* (Piscataway, NJ: IEEE).

Wang, P., Low, K., and McGregor, A. (2011). "A subject-based motion generation model with adjustable walking pattern for a gait robotic trainer: NaTUre-gaits," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Piscataway, NJ: IEEE).

Wei, W., Wu, H., and Ma, H. (2019). An autoencoder and LSTM-based traffic flow prediction method. *Sensors* 19:2946. doi: 10.3390/s19132946

Winter, D. A. (1991). *Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological.* Waterloo, ON: University of Waterloo Press.

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement.* Hoboken, NJ: John Wiley & Sons.

Zaroug, A., Proud, J. K., Lai, D. T., Mudie, K., Billing, D., and Begg, R. (2019). "Overview of Computational Intelligence (CI) techniques for powered exoskeletons," in *Proceedings of the Computational Intelligence in Sensor Networks.* (Berlin: Springer), 353–383.

Zhang, J., Fiers, P., Witte, K. A., Jackson, R. W., Poggensee, K. L., Atkeson, C. G., et al. (2017). Human-in-the-loop optimization of exoskeleton assistance during walking. *Science* 356, 1280–1284. doi: 10.1126/science.aal5054

Zhang, Y., Yang, Z., Lan, K., Liu, X., Zhang, Z., Li, P., et al. (2019). "Sleep stage classification using bidirectional lstm in wearable multi-sensor systems," in *Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (Piscataway, NJ: IEEE).

Zhao, A., Qi, L., Dong, J., and Yu, H. (2018). Dual channel LSTM based multi-feature extraction in gait for diagnosis of Neurodegenerative diseases. *Knowledge Based Syst.* 145, 91–97.

Automatic Pose Recognition for Monitoring Dangerous Situations in Ambient-Assisted Living

Bruna Maria Vittoria Guerra, Stefano Ramat*, Giorgio Beltrami and Micaela Schmid

Laboratory of Bioengineering, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

Continuous monitoring of frail individuals for detecting dangerous situations during their daily living at home can be a powerful tool toward their inclusion in the society by allowing living independently while safely. To this goal we developed a pose recognition system tailored to disabled students living in college dorms and based on skeleton tracking through four Kinect One devices independently recording the inhabitant with different viewpoints, while preserving the individual's privacy. The system is intended to classify each data frame and provide the classification result to a further decision-making algorithm, which may trigger an alarm based on the classified pose and the location of the subject with respect to the furniture in the room. An extensive dataset was recorded on 12 individuals moving in a mockup room and undertaking four poses to be recognized: standing, sitting, lying down, and "dangerous sitting." The latter consists of the subject slumped in a chair with his/her head lying forward or backward as if unconscious. Each skeleton frame was labeled and represented using 10 discriminative features: three skeletal joint vertical coordinates and seven relative and absolute angles describing articular joint positions and body segment orientation. In order to classify the pose of the subject in each skeleton frame we built a two hidden layers multi-layer perceptron neural network with a "SoftMax" output layer, which we trained on the data from 10 of the 12 subjects (495,728 frames), with the data from the two remaining subjects representing the test set (106,802 frames). The system achieved very promising results, with an average accuracy of 83.9% (ranging 82.7 and 94.3% in each of the four classes). Our work proves the usefulness of human pose recognition based on machine learning in the field of safety monitoring in assisted living conditions.

Keywords: Ambient-Assisted Living, vision-based activity recognition, skeleton tracking, pose recognition, machine learning, geometric features

Abbreviations: AAL, Ambient-Assisted Living; ANN, artificial neural network; DL, deep learning; DT, decision tree; HMM, hidden Markov model; IoT, internet of things; KNN, K-nearest neighbor; LR, logistic regression; ML, machine learning; MLP, multi-layer perceptron; NBC, Naïve Bayes classifier; RF, random forest; SVM, support vector machine.

INTRODUCTION

The integration of *frail* people into society is a major issue in developed countries for both social and economic motivations. This inclusion should start with the environment in which these subjects live, and can be achieved by improving well-being, autonomy, care, and assistance in the home. Internet of things (IoT) and modern domotic technologies offer a plethora of solutions to implement intelligent and automated environments allowing *frail* individuals to live in autonomy and safety in place (Álvarez-García, 2013; Amiribesheli et al., 2015; Lloret et al., 2015; Debes et al., 2016; Mehr et al., 2016; Majumder et al., 2017; Guo et al., 2019). In the last years, Ambient-Assisted Living (AAL) has attracted great attention and numerous projects have proposed different networks of sensors and complex monitoring algorithms which most frequently require to shift from a low-level data collection and analysis toward high-level information integration, context processing, activity recognition and inference (Chen et al., 2012; Verrini et al., 2018).

The most commonly used sensors for AAL are wearable and environmental sensors (Delahoz and Labrador, 2014; Pannurat et al., 2014; Mehr et al., 2016; Torti et al., 2019). The first category includes radio frequency identification tags, accelerometers, gyroscopes, and more generally inertial sensors which can be embedded in devices such as smartphones and smartwatches. The main advantages of wearable sensors are to be particularly light and non-intrusive, yet they have the important drawback of being dependent on rechargeable batteries and of requiring correct body positioning to maximize the signal quality and reduce noise.

The second category, environmental sensors, commonly refers to cameras able to monitor an inhabitant behavior and environment changes (vision-based activity recognition) (Chen et al., 2012). Using properly located cameras, the inhabitant can be recorded while free to perform the normal actions of daily life without limitations and without having to be in anyway involved, e.g., having to remember to wear a device or to charge it. The cameras used for AAL purposes are commonly depth cameras, such as Asus Xtion (Taipei, Taiwan), Intel RealSense (Santa Clara, United States), Orbbec Astra (Troy, United States) and Microsoft Kinect (Redmond, United States) (Ben Hadj Mohamed et al., 2013; Han et al., 2013; Gasparrini et al., 2014; Mastorakis and Makris, 2014; Pannurat et al., 2014; Visutarrom et al., 2014, 2015; Damaševičius et al., 2016; Calin and Coroiu, 2018). Thanks to many approaches based on RGB sequences, depth images or their combination, these sensors are able to provide detailed information about 3D human motion (Wang et al., 2014; Kim et al., 2017). Moreover, real time algorithms can estimate the body skeleton, which allows to describe human poses with a lower dimensionality than RGB/RGB-D-based representations while being intrinsically anonymous, thereby respecting the privacy of the subject.

To infer what an individual is doing, and which pose he/she assumes, the data collected from both wearable sensors and cameras are commonly processed using data mining, machine learning (ML), and deep learning (DL) algorithms. Machine learning focuses on teaching computers how to learn from experience, without the need to be programed for specific

tasks. This makes ML particularly suitable to analyze data coming from smart house sensors in order to recognize falls or to detect a dangerous situation during daily life activities. Machine learning algorithms such as Naïve Bayes classifiers (NBC), K-nearest neighbor (KNN), support vector machines (SVM), hidden Markov models (HMM), and artificial neural networks (ANN), random forest (RF), decision tree (DT), and logistic regression (LR) (Begg and Hassan, 2006; Crandall and Cook, 2010; Hussein et al., 2014; Visutarrom et al., 2014; Wang et al., 2014; Amiribesheli et al., 2015; Jalal et al., 2015) are the most popular algorithms used in sensor- and vision-based activity recognition. K-nearest neighbor is widely used in real-life scenarios since it is non-parametric, meaning that it does not make any assumptions about the underlying distribution of the data. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, therefore it turns out slow with large numbers of training examples. A similar weakness affects the SVM algorithm, which nevertheless is considered relatively memory efficient. Achieving the best classification results, for any given problem, requires setting several key parameters that need to be chosen correctly (Bishop, 2006). Artificial neural networks, such as multi-layer perceptron (MLP) algorithm, can be applied to many smart home problems, ranging from activity classification, to novelty and anomaly recognition, to activity prediction (Begg and Hassan, 2006; Hussein et al., 2014). Patsadu et al. (2012) compared four ML algorithms (MLP, SVM, DT, and NBC) training the models on a dataset of 7200 frames and testing them on further 3600 frames to identify three different human poses: standing, sitting, and lying down. The poses were performed by a subject positioned in front of the camera and each videoframe was encoded as a row of 20 body-joints positions that were used as features for ML algorithms. The best classifier was found to be the MLP network (100% of accuracy vs. 99.75% of SVM, 93.19% of DT and 81.94% of NBC). Visutarrom et al. (2014) went deeper into this topic comparing six different ML classifiers and two different sets of features (geometric vs. skeletal joints features). Four poses (standing, sitting, sitting on floor, and lying down) of a subject watching television in front of the Kinect device were classified. They compared MLP, DT, NBC, RF, LR, and SVM by training and testing the six models on a dataset of geometric features and found that DT, RF, and MLP algorithms performed best (accuracy about 97.9%), followed by the SVM (accuracy about 97.5%). Altogether various ML algorithms have been successfully applied to pose recognition, yet all these approaches suffer from various limitations that do not prove their usefulness in the context of identifying dangerous situations in ecological conditions of assisted living. Indeed, all algorithms were applied to recordings performed by subjects statically facing the camera, i.e., the ideal conditions for skeleton tracking systems, which are nonetheless unlikely to occur while monitored subjects perform their daily living activities at home. Furthermore, their performance has been tested in recognizing upright standing or poses typically assumed immediately after a fall, e.g., lying down or sitting on the floor, yet omitting more general dangerous situations such as recognizing that a person has fainted while sitting on a chair.

In an automated monitoring system for AAL, the accuracy of event recognition is vital. False negatives are unacceptable as they would imply the lack of intervention in a dangerous condition. Recognition accuracy is strongly dependent not only on the model algorithm, but also on the type and number of attributes that make up the database used to train the network. In vision-based action recognition, the common approach is to extract image features from video data and to issue a corresponding action class label (Poppe, 2010; Babiker et al., 2018). Nevertheless, when skeleton representation of the human body is used, the most privileged discriminative features are the raw data coming from the skeletal tracking (joint spatial coordinates) (Patsadu et al., 2012; Youness and Abdelhak, 2016) or some indices expressing geometric relations between certain body points, such as: the vertical distance from hip joint to room floor (Visutarrom et al., 2014, 2015), the distance between the right toe and the plane spanned by the left ankle, the left hip and the foot for a fixed pose (Müller et al., 2005) the distance between two joints, two body segments, or a joint and a body segment (Yang and Tian, 2014), the relative angle between two segments within the body kinematic chain (Müller et al., 2005) and finally, the size of the 3D bounding box enclosing the body skeleton (Bevilacqua et al., 2014). Geometric features are synthetic in the sense that they express a single geometric aspect making them particularly robust to spatial variations that are not correlated with the aspect of interest (Müller et al., 2005). In order to identify the best attribute set to classify, off- and on-line, standing, bending, lying, and sitting poses, Le et al. (2013) compared the results of a ML algorithm trained and tested with four different sets of features. They trained an SVM with a radial basis function kernel on off-line data referred to a subject in front of the camera, using 7, 9, and 17 joint angles with and without scaling, and absolute joint coordinates without scaling. In off-line, optimal Kinect acquisition configuration, very good results were obtained with the absolute coordinates without scaling. They then tested the algorithm also using on-line data of a subject at a different distance from and at different angles with the Kinect camera. In the latter, more realistic scenarios, the angles were found to represent more relevant features for posture representation.

In this paper we focus on the problem of skeleton-based human pose recognition for the detection of dangerous situations. This work is part of a broader project (TheDALUS, The Disable Assisting Living for University Students), aimed to promote the inclusion of disabled students in a university environment (a room in college dorms) guaranteeing them safety and autonomy. This is done using a net of four Kinect One devices, whose data are collated and processed to identify both voluntary requests for help and dangerous situations (i.e., the subject has fainted or slipped from the wheeling chair, etc.) to trigger an alarm toward third parties, when needed. During daily activities a subject assumes a set of poses that can be very similar to those assumed during dangerous situations. Our approach is based upon the consideration that to distinguish these two different scenarios knowledge of the location domain is fundamental (the spatial position of the room inhabitant, objects and room furniture position and the relative position of the inhabitant with respect to the objects and the room furniture). Indeed, a normal pose could become a dangerous one when it takes place in relatively specific locations of the room. For example, the lying down pose is a daily living pose if it occurs on the bed. Conversely, it takes the form of a possible alarm condition if it occurs on the floor. In this context, an accurate body pose pattern recognition model must be defined first, and, in a later processing stage, the identified poses can be joined with the knowledge of the location domain. This implicit relationship between body poses and related spatial context provides the heuristics to infer the occurrence of a dangerous scenario, thereby broadening the scope of current approaches of ML in human pose recognition to the field of monitoring safety in assisted living conditions. The aim of this study is to implement the first step of this analysis procedure by using a large amount of skeleton tracking data referred to real scenarios, in which a more extensive camera coverage of the room is obtained by using four Kinect One devices. As such, here we are interested in classifying each acquired skeleton frame provided by the device in a set of predefined poses (standing, sitting, lying down and "dangerous sitting"). To this goal a three layers MLP network was trained and tested using a custom-built data set of robust and discriminative kinematic features computed based on skeleton data.

MATERIALS AND METHODS

Experimental Set-Up
In order to minimize the invasiveness of the monitoring system, a main requirement in a 24-h surveillance of daily activities setting, we decided to avoid any wearable sensor. On the other hand, considering the constraints raised by the privacy of the students inhabiting the rooms, video recording and video surveillance systems did not represent a viable option. We therefore chose a motion sensing system based on skeletal tracking (Booranrom et al., 2014; Du et al., 2015; Gasparrini et al., 2015; Visutarrom et al., 2015; Liu et al., 2018). The current implementation is based on Microsoft's Kinect One motion sensing system, yet it is easily portable to any skeletal tracking device that can provide the 3D coordinates of the chosen set of skeletal joints. The Kinect One motion sensing system can detect a human body and voice signal using an RGB camera (1920 × 1080 pixels), a depth sensor (512 × 424 pixels) and an array of four microphones (48 kHz). The depth sensor is composed of an IR emitter and an IR camera and provides depth measurements based on the Time of Flight principle (Pagliari and Pinto, 2015; Sarbolandi et al., 2015; Corti et al., 2016). Acquisitions can be carried out with a framerate up to 30 Hz and require a computer with an USB 3.0 interface for data transfer. The ideal distance of an object from the sensor is 0.8–3.5 m, with a maximum range of 0.5–4.5 m. The angle of vision is 60° vertically and 70° horizontally (Sell and O'Connor, 2014; Fankhauser et al., 2015; Pagliari and Pinto, 2015). Microsoft released also a Software Development Kit (SDK), used for skeletal estimation. It is capable of tracking 25 joints for up to six users simultaneously (Microsoft, 2019).

Experimental acquisitions were performed in a prototype room, mimicking that of the university college dorms (same dimensions and similar furniture) that was set up in the laboratory. In this setting, we decided to record each experimental trial using four Kinect One devices (K1, K2, K3, and K4 in **Figure 1**). Two are positioned to sense the whole room (K1 and K4), while the remaining two are placed to specifically acquire two areas of the room, such as the bed (K2) and the desk (K3), which were especially relevant to our aim. This decision was made after several careful eye-inspections of the different shots obtained with different camera configurations. Each arrangement was different for number, position and orientation of the devices. The goal was to ensure recording of the entire room minimizing possible blind spots. The data of the four Kinect One were acquired at the same time but processed separately. A custom-made C#-based tool with GUI was developed using VisualStudio 2017 to control the Kinect One acquisitions.

Acquisitions

We decided to focus our acquisitions around the three most frequent and recurrent poses assumed by a person in a room during daily activities (Datasets – Advise, 2019; Fall detection Dataset, 2019; Fall detection testing dataset, 2019; Weblet Importer, 2019): standing, sitting, and lying down. In addition to the listed poses, we added one further pose, labeled "dangerous sitting," which grouped all situations of malaise or fainting resulting in a seated person slumped or lying backward. This allowed us to perform a first distinction, prior to establishing a relationship between the subject location and the room furniture, between routine activities and alarm situations. Experimental protocols were designed to simulate the actions and poses performed during the daily life of a general disabled student, not necessarily having motor disabilities.

In order to build a dataset suitable for training a neural network to discern the four poses we performed a set of experimental acquisitions on a group of 12 normal subjects (7 females and 5 males; age ranging 25 and 60 years old; height ranging 1.55 and 1.90 m). All subjects gave written informed consent in accordance with the Declaration of Helsinki. The four Kinect One devices installed in the room acquired simultaneously the movement of the subject. The acquisitions were structured as four separate sessions performed on the same day for a total of about 13 min:

- subject starts to walk from standing position in front of K1 (**Figure 1**), then grabs a chair near the desk, placing it in front of the camera, and finally sits on it. While sitting, the subject first moves the head backward and then leans the trunk forward, while simultaneously pitching the head as an unconscious person. The subject then returns to the normal sitting position and finally gets up and brings the chair back to its original location. Each pose was maintained for 10 s. The sequence was then repeated in front of the other cameras (K2, K3, K4 in **Figure 1**);

- subject starts sitting on the bed, then lies down on the back, turns on the right side, then returns on the back and turns to the other side;
- subject starts lying on the ground on the back, then turns on the left side;
- subject starts sitting on the bed, then lies down. The action is repeated three times.

The sequence of poses in each acquisition was timed by the operator running the acquisitions.

Data Pre-processing

Using custom developed software based on the Kinect's SDK we computed the spatial coordinates (x, y, z) of the standardized 25 skeletal joints (Microsoft, 2019). Based on considerations relative to the reliability of the detected joints and to the aim of this study, we decided to reduce the number of skeletal joints from 25 (**Figure 2A**) to 16. An additional joint labeled Hc was instead added as the midpoint between the two hips joints (**Figure 2B**).

The 17 selected joints were (**Figure 2B**): head (1), shoulders segment's mid-point (C7), acromion (3–4), elbow (5–6), wrist (7–8), iliac crest (9–10), knee (11–12), malleolus (13–14), foot (15–16); hips segment's mid-point (Hc). In order to identify the position of the subject in the room, the coordinates of the 17 joints were roto-translated to obtain data referred to an absolute reference system (X, Y, Z) located in one corner of the mock-up room (**Figure 1**). The absolute position in space of each body joint, described by the corresponding X, Y, Z triplet, isn't the most convenient description for classifying human poses, since: (1) coordinates depend on the relative location of the individual in the room, while the same posture can be taken in different locations within the room; (2) the joint coordinates of two subjects having the same pose in the same room location have different values depending on the size of the subject's body; and (3) posture is independent of where it occurs in space while it is defined by the geometrical relationship between the different body segments. The latter can instead be efficiently captured by articular angles, so that we chose to compute the following 16 articular angles defined between two consecutive body segments measured in the plane defined by the segments themselves: head–shoulder axis (μ_1, μ_2), head–trunk (ξ), shoulder axis–trunk (τ_1, τ_2), shoulder axis–arm (η_1, η_2), arm–forearm (θ_1, θ_2), trunk–iliac crest axis (δ_1, δ_2), iliac crest axis–thigh (γ_1, γ_2), thigh–leg (β_1, β_2), and leg–foot (α_1, α_2) (**Figure 2B**). Based on the same line of reasoning we further computed the roll and pitch angles of the head and trunk and labeled them as follows: A_pitch (trunk pitch), A_roll (trunk roll), B_pitch (head pitch), and B_roll (head roll). All angles were normalized dividing them by 180°. We further considered the vertical coordinates (Z) of the skeletal joints as they are significant for distinguishing the lying down from the standing pose. On the other hand these are not so discriminative for discerning between sitting and "dangerous sitting" poses, which are more easily identified through joints angles' values. The joints' Z coordinates werethen scaled on the height of each subject.

FIGURE 1 | Kinect one positions in the prototype room (K1, K2, K3, and K4), reconstructed with a CAD software (SketchUp). Two different fields of view.

FIGURE 2 | (A) 25-joints skeleton computed by Kinect One; **(B)** reduced skeleton used in this analysis with 17 joints for the left and right hemi-body. The relative angles computed in the plane defined by the two body segments are depicted, for illustrative purpose, only for the frontal plane and labeled with Greek letters. For clarity, except for τ_1, only the angles of the right hemi-body are shown.

During the acquisition process we noted that sometimes Kinect One was not able to recognize the subject. For example, transient exits of the subject from the camera sight (**Figure 1**) could cause temporary non-identifications of all skeletal joints, and the same may occur when the subject assumes a "dangerous sitting" pose while not facing the camera. This could generate temporal holes between data frames (missing data). For these frames we decided to assign the value "999" to all the selected parameters in order to maintain consistency among the data of the four Kinect One systems. All the pre-processing algorithms were implemented using MATLAB.

Database

Once all joints' Z-coordinates, the relative angles and the chosen pitch and roll angles were obtained, i.e., a total of 37 (17 vertical joints coordinates, 16 relative angles, 4 absolute angles) features describing the skeleton in each frame. We then applied a ReliefF (Urbanowicz et al., 2018) algorithm for feature selection

(MATLAB) and selected a subset of ten attributes: A_pitch, A_roll, B_pitch, B_roll, ξ, μ_2, δ_2, Z_1, Z_C7, Z_Hc (see **Figure 2B** for the last six attributes).

Using a custom-made LabView (National Instruments, Inc.) software, angles and joints position traces were then visually inspected together with a graphical visualization of the reconstructed skeleton to label each frame with one of the following four poses:

- Class 1: standing pose;
- Class 2: sitting pose;
- Class 3: lying down pose;
- Class 4: "dangerous sitting" pose.

Using the same software we also identified the frames corresponding to the transition from a pose to another and removed them from the dataset. The data from the four Kinect One systems were collated to build the final database

FIGURE 3 | Data acquisition and processing with the proposed MLP network model having two hidden layers and a "SoftMax" output layer. The considered features are 10 kinematic parameters computed from skeletal tracking. The output classification corresponds to the one of the four classes (Class 1, Class 2, Class 3, Class 4).
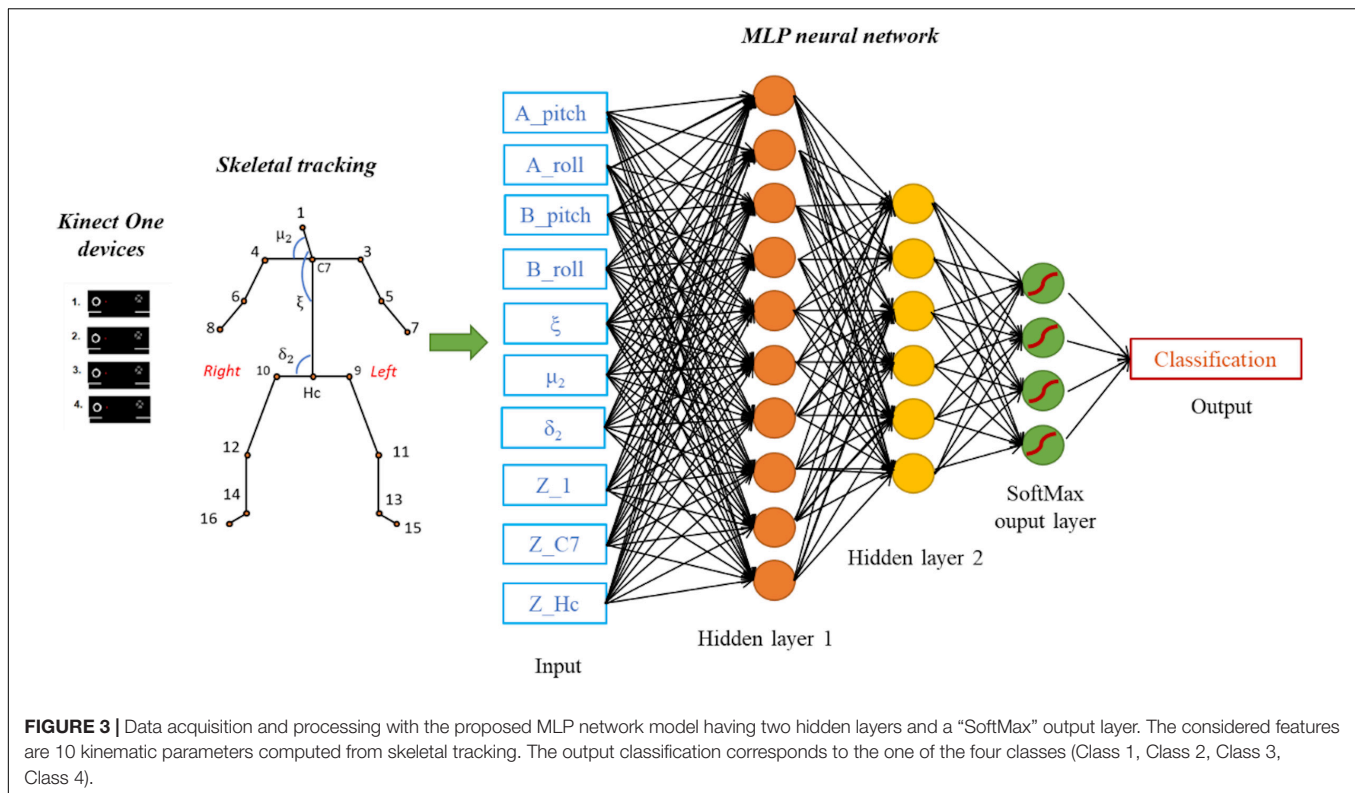
composed by 602,530 frames. Among these, 145,196 frames belonged to Class 1, 233,593 to Class 2, 86,786 to Class 3, and 136,955 to Class 4.

A training set was eventually built using the data from 10 of the 12 subjects (database of 495,728 frames). The test set was built using the data of the 2 remaining subjects (database of 106,802 frames).

Neural Network

The aim of this work was not to detect dynamic situations, such as the falling of the subject in order to prevent it, but rather to identify the subject lying on the floor immediately after the fall in order to activate an alarm and intervene with first aid actions. Therefore, in the current implementation we wanted to identify a subject pose at any one time, leaving the decision-making process about alarm triggering to a downstream algorithm having access to more data (e.g., subject's position in the room). The pose classification problem is therefore seen as a static mapping problem. For this reason, among a range of possible ML algorithms, we have selected an MLP Neural Network to classify predefined human poses. The network was implemented in MATLAB using the Neural Network Toolbox. We designed a network consisting of three fully connected layers of neurons, plus an input layer connected to the 10 features describing each frame in the database (**Figure 3**). The first hidden layer has a number of neurons equal to the number of attributes in the database (10), each with a hyperbolic tangent transfer function and a bias. The second hidden layer has a structure similar to the first one, but contains a smaller number of neural

units (6). The output layer is instead composed by a number of neurons equal to the number of target classes (4) and their transfer function is the "SoftMax" function producing, for each input element, the probabilities of belonging to each considered class. The MLP network was trained using the Levenberg-Marquardt backpropagation algorithm, first with a k-fold cross validation ($k = 10$), and then using the whole training set. The learning process was performed over a maximum of 1000 epochs, i.e., 1000 iterations on the training set.

Statistical Analysis

MLP network was trained and tested 50 times to study its classification robustness. Total accuracy (mean over the four classes), class accuracy, F-score, sensitivity, and specificity were calculated for each network simulation. These parameters rely upon the concept of True Positive (TP, a pose correctly classified as pertaining to the considered class), True Negative (TN, a pose which is correctly classified as pertaining to a class different from the one considered), False Positive (FP, a pose that is wrongly classified as pertaining to the considered class), and False Negative (FN a pose that is wrongly classified as not pertaining to the class considered).

Accuracy

Accuracy is a metric parameter for evaluating classification models. In general, for binary classification, accuracy can be calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Sensitivity

The Sensitivity (also called Recall) is a metric parameter that measures the proportion of genuinely positive samples that are currently identified as such. It is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity

The Specificity is the proportion of genuinely negative samples that are currently identified as such. It is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F-Score

F-score is an overall measure model's accuracy that combines precision and recall. Precision is the number of positive results divided by the number of all positive results returned by a classifier. Recall, instead, is the ratio between TP and the number of all samples that should have been identified as positive, which corresponds to the sensitivity parameter.

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

$$\text{Precision} = \frac{TP}{TP + FP}$$

For each of five parameters considered, the mean value over the 50 network simulations was then computed. This average operation was done only after verifying that the results listed above were normally distributed. Since the number of samples was 50, i.e., the number of network simulations, we decided to use the Shapiro–Wilk test as a hypothesis test (Hanusz and Tarasińska, 2015). The null hypothesis of this test is that the population is normally distributed. For each test performed the p-value was greater than the chosen alpha level, therefore the null hypothesis that the data came from a normally distributed population cannot be rejected (IBM SPSS Statistics, IBM). Therefore, in the result section, for each of the five parameters, the mean and the standard deviation are considered.

Confusion Matrix

Confusion matrix is a specific table summarizing the results of the classifier used to visualize the performance of a machine learning algorithm. The rows of the matrix represent the classifications predicted by the MLP network while the columns represent the instances actually belonging to each class.

In the present study we computed a confusion matrix for each of the 50 network simulations. Then, we computed a mean confusion matrix in which the number of frames reported in each cell is the mean, over the 50 confusion matrices, of the frames pertaining to that cell.

ROC Curve

ROC curve graph shows the performance of a classification model. True positive rate (sensitivity) is plotted against the false positive rate (1-specificity) at different classification thresholds.
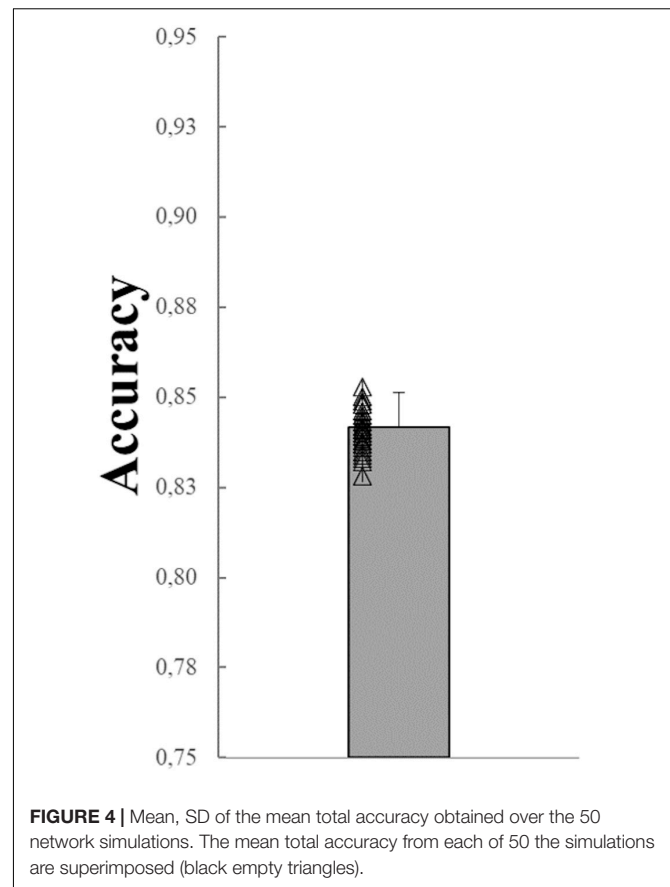
FIGURE 4 | Mean, SD of the mean total accuracy obtained over the 50 network simulations. The mean total accuracy from each of 50 the simulations are superimposed (black empty triangles).

The area under the ROC curve (AUC) gives an index of the performance of the classifier. Higher values of AUC correspond to a good prediction of the model.

In the present study we computed, for each class, the ROC curve graph for each of the 50 network simulations. Then, to obtain a mean ROC curve, we averaged the ROC curve of the 50 simulations as the mean true positive rate for each value of the false positive rate considered on the abscissa.

RESULTS

In **Figure 4** the mean value (Mean), the corresponding standard deviation (SD) and the distribution of the 50 mean total accuracy values, each corresponding to one of the 50 network simulations (0.839 ± 0.0073) are shown. Values range 0.852–0.820.

Figure 5 shows mean values, SD, and the distributions of the accuracy, F-score, sensitivity, and specificity of each of the four classes. All four variables represented in **Figure 5** show a similar trend. Class 3, which corresponds to the lying pose, and Class 1, which corresponds to the standing pose, represent the classes best identified by the net. The network, on the other hand, classified Class 2 (sitting pose) and, especially, Class 4 ("dangerous sitting" pose) with more difficulty for each of the four variables calculated (**Figure 5**).
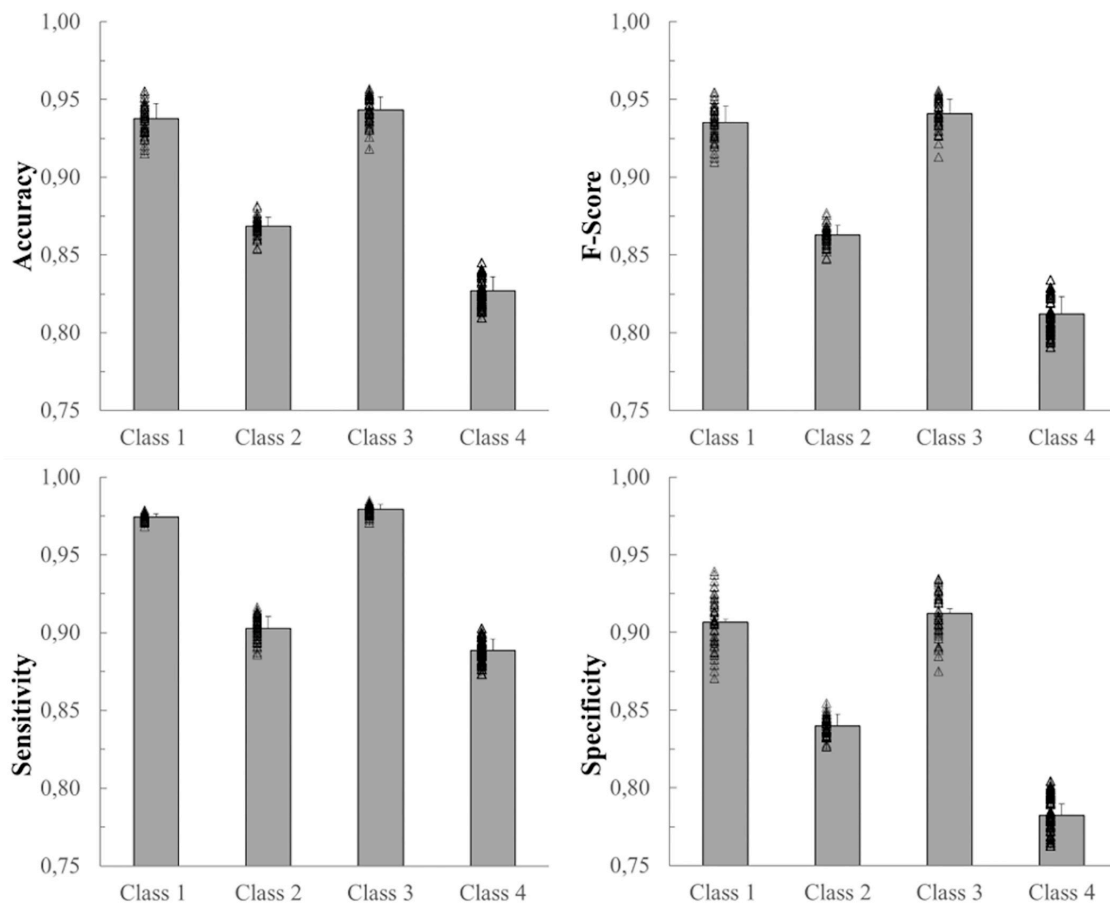
FIGURE 5 | Mean, SD and individual results (black empty triangles) of the 50 network simulations. From top left: accuracy, F-score, sensitivity, and specificity for each of the 4 classes.

Figure 6 shows, for each class, the ROC curves calculated on the 50 network simulations. The average ROC curve has been calculated for each of the four classes, in order to observe the learning behavior of the network during its 50 simulations. The average ROC curves confirm the observations made previously, i.e., that Class 1 and Class 3 are better identified by the neural network than Class 2 and Class 4. The same results are confirmed also by computing the AUC values for the average ROC curves of the four classes (97.2 for Class1, 92.1 for Class2, 98.5 for Class3 and 89.2 for Class4). **Figure 6** also shows the greater variability of the ROC curves relative to Class 4, compared to those obtained with Class 1, Class 2, and Class 3.

Figure 7 shows the mean confusion matrix computed over the entire set of 50 network simulations performed. It summarizes the average values of the False Positives (FP), False Negatives (FN), True Negatives (TN) and True Positives (TP) for each class.

DISCUSSION

In order to grant the safety of disabled students living in automated rooms of university dorms while allowing for their independency, their privacy and freedom of movement, we developed a 24/7 monitoring system being able to raise an alarm, either upon request of the student, or automatically when a danger situation is identified. The approach implemented here was based on instrumenting the room with four skeleton-tracking Kinect One devices providing the data for identifying dangerous situations.

In this work we presented a pose recognition system processing the skeleton information provided by the Kinect One devices using a static neural network that classifies the data relative to each frame in one of four classes corresponding to the four poses considered. Three of these (standing, sitting, and lying) represent the most common poses taken by a subject while living in a room, while the fourth ("dangerous sitting") represents a potential danger situation in which the subject is sitting on a chair with the head forward or backward, that might need an external intervention.

We decided to train and test an MLP model with two hidden layers and a "SoftMax" output layer, in order to classify the four poses described before. After the selection of the attributes and the construction of the database, the MLP neural network was trained and tested 50 times in order to provide data for a statistically reliable description of its performances.

FIGURE 6 | Set of 50 ROC curves calculated for each of the four classes (gray traces). The average ROC curve (black trace) was also calculated and superimposed.
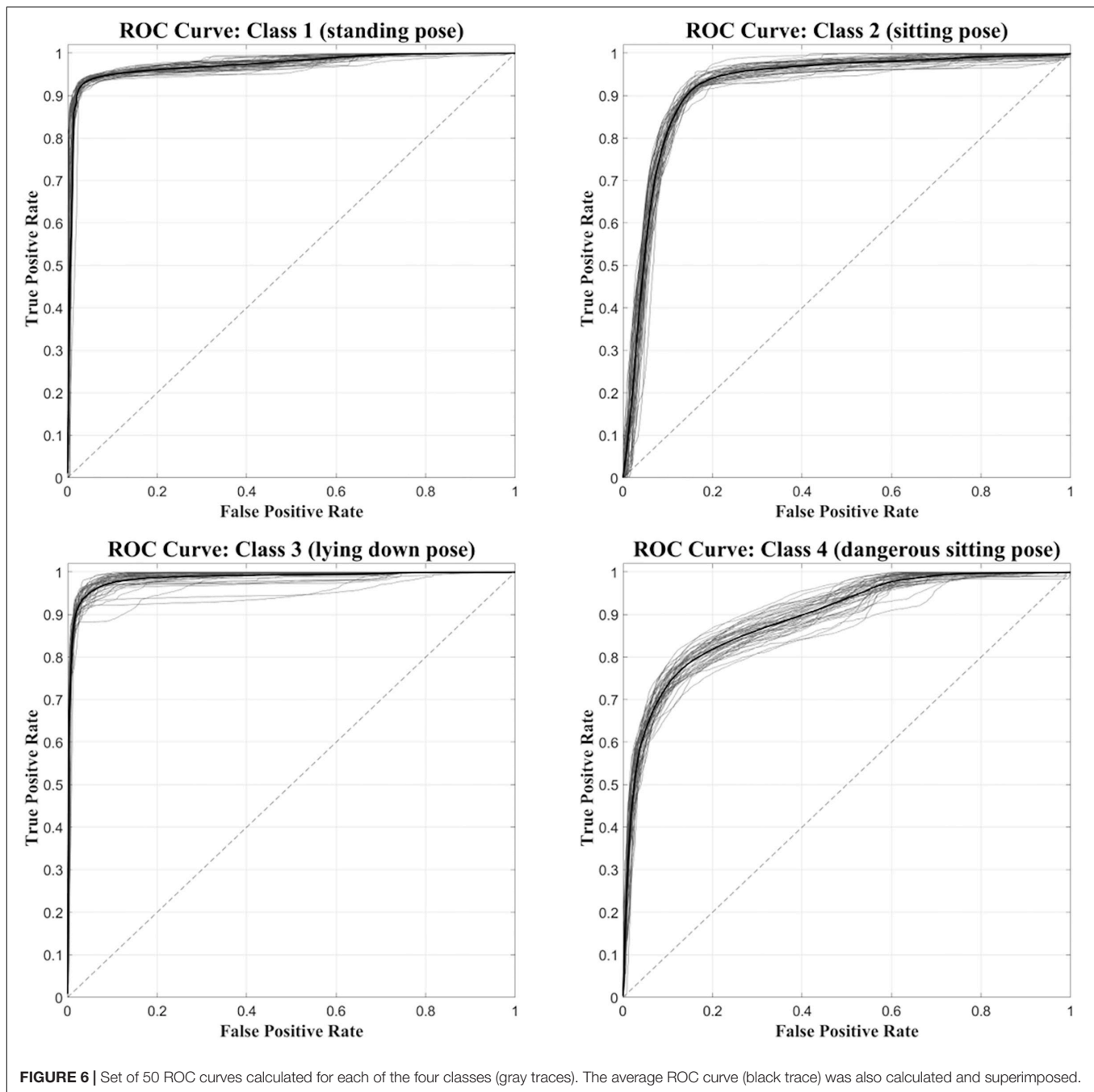
Previous studies have faced similar problems using ML algorithms with good results, although on smaller datasets and asking the subject to maintain the planned poses while facing the camera, i.e., a very favorable condition for the Kinect acquisition, yet unlikely in our project scenario (Patsadu et al., 2012; Visutarrom et al., 2014). Our study considered a less constrained dataset in which 12 subjects were recorded in the defined poses both statically (e.g., lying down) and while moving over the entire room area (e.g., the subjects were walking when assuming the standing pose) for a total of 495728 frames for training and 106802 frames for testing. As a result, our data was more variable in terms of how each subject interpreted the requested poses, and noisier for the different views recorded by each of the four Kinect One devices, which were necessarily frequently sub-optimal.

In spite of these limitations, nonetheless, required to mimic real life conditions, the proposed MLP classifier achieved good results with a total average accuracy of 83.9%. A more detailed inspection of the results relative to the four classes shows that Class 3 and Class 1 are better recognized than the remaining two classes, with average accuracies around 94% (94.3 and 93.8%, respectively). On the other hand, Class 2 and Class 4, both

FIGURE 7 | Mean confusion matrix obtained from the 50 network simulations. It represents the False Positive (FP), False Negative (FN), True Negative (TN) and True Positive (TP) values computed for each of the four classes.
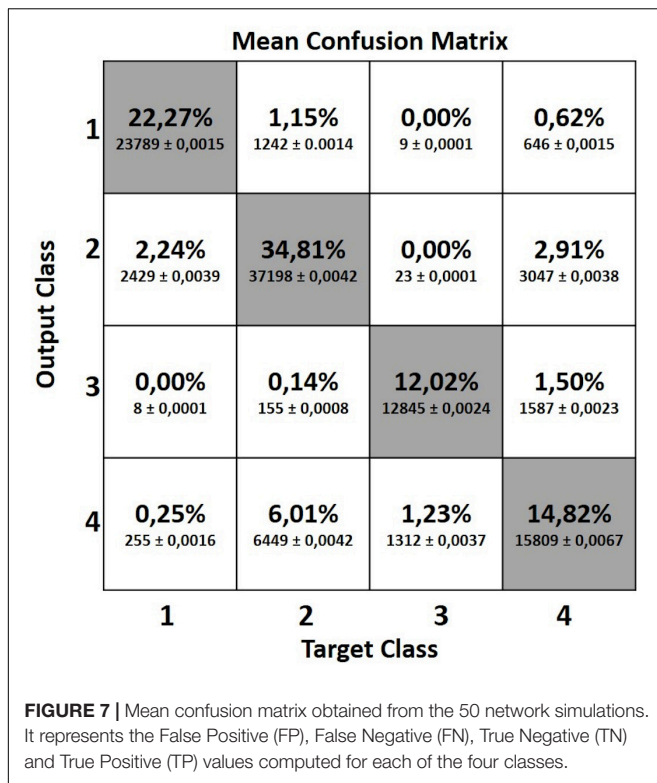
regarding sitting positions yet differing mostly in terms of trunk and head pitch angles, were less accurately recognized (86.9 and 82.7%, respectively), with frames being incorrectly assigned to the two classes (see accuracy values in **Figure 5**). These lower accuracy values are mainly due to the misclassification errors between Class 2 and Class 4 and vice-versa. Indeed the 6.01% of frames labeled as Class 2 were identified as Class 4 and the 2.91% of Class 4 data were classified as Class 2 (see the mean confusion matrix in **Figure 7**). At least two plausible reasons can be considered as contributing to this misclassification error in recognizing these two poses. First and foremost, during sitting some articular joints are covered by other body parts, thereby requiring the Kinect One system to reconstruct the positions of the hidden joints and making the resulting data very noisy. Second, despite the careful choice of features as powerful descriptors of body poses while being independent from the physical characteristics of the subjects who participated to the study, the distinction between two kinematically very similar poses is very difficult. The number of features that can help the classifier to distinguish between them is reduced. Only the upper body features may be discriminative and probably, even among these, the normalized vertical positions of the head and cervical vertebrae (Z_1, Z_{C7}), i.e., the most discriminative joint-related features for the identification of Class 1, 2 and 3, sometimes take comparable values between Class 2 and 4 due to the subjects' individual interpretation of the description of the "dangerous sitting" pose.

Another relatively important misclassification error was between Class 1 and Class 2 and vice-versa (2.24% Class 1

identified as Class 2 and 1.15% Class 2 identified as Class 1). For the identification of these two poses, the vertical position of the joints (Z_1, Z_{C7} and Z_{Hc}) should be more informative for the MLP network. Nevertheless, in our study this was not so evident probably because some of the data calculated by Kinect One devices are particularly noisy, especially when the subject is not exactly in front of the camera (Rougier et al., 2011; Li et al., 2019). The relative angles and the head and trunk absolute angles do not weight as much in the distinction between the two classes since they assume comparable values. Conversely, lower misclassification error was found for the standing pose (Class 1) and the "dangerous sitting" pose (Class 4) and vice-versa (0.25% Class 1 identified as Class 4 and 0.62% Class 4 identified as Class 1, respectively). In this case, the relative and absolute angles of head and trunk features in the database are more discriminant.

The lowest misclassification error, almost equal to zero, was that between the identification of standing (Class 1) and lying down (Class 3) poses and vice-versa, where the vertical position of the joints is very discriminative.

Considering the assumptions made so far in order to explain the misclassification errors we can hypothesize that an appropriate pre-processing of the data could significantly reduce the number of misclassifications. A classification model requires a reliable and valid dataset to efficiently generate the decision-making rules. To reduce classification errors, the quality of the data provided to the classifier is important during both the training and the usage phases, so that data pre-processing techniques removing anatomically implausible body reconstructions resulting in longer than real limbs or in impossible articular angles may be needed.

CONCLUSION

We have proposed an implementation of a pose classification system for monitoring frail individuals in their daily living facilities. Kinect One devices, recording an inhabitant moving in a room in real scenarios, provided skeleton data frames. These data were processed to compute a set of features that make up the database for training and testing a three layers MLP neural network for inhabitant pose recognition (standing, lying, sitting and "dangerous sitting"). We built a database with a large amount of data (over 600,000 frames) in which each pose was described by a set of geometric and vertical joint position features. Despite the data were quite noisy as they were acquired with the subject not necessarily facing the camera, the proposed MLP network achieves a good mean total accuracy of 83.9%.

This work is, to our knowledge, the first attempt to classify human poses based on skeleton tracking data acquired in an ecological daily living scenario, with no constraints on the relative position of the subject with respect to the recording devices, and with an extensive dataset comprising sitting and "dangerous sitting" classes.

This work has been designed for a room tailored to disabled students, but it can be extended to all those categories of individuals living in community environments, such as the

elderly, where safety, accessibility and autonomy can be a restriction to participation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The experimental protocol was conducted in accordance with the Declaration of Helsinki. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Álvarez-García, J. A. (2013). Evaluating human activity recognition systems for AAL environments. *Commun. Comput. Inf. Sci.* 362, 131–136. doi: 10.1007/978-3-642-37419-7_10

Amiribesheli, M., Benmansour, A., and Bouchachia, A. (2015). A review of smart homes in healthcare. *J. Ambient Intell. Humaniz. Comput.* 6, 495–517. doi: 10.1007/s12652-015-0270-272

Babiker, M., Khalifa, O. O., Htike, K. K., Hassan, A., and Zaharadeen, M. (2018). "Proceedings of the 2017 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2017," in *Automated Daily Human Activity Recognition for Video Surveillance Using Neural Network*, Putrajaya. doi: 10.1109/ICSIMA.2017.8312024

Begg, R., and Hassan, R. (2006). "Artificial neural networks in smart homes," in *Designing Smart Homes. Lecture Notes in Computer Science*, eds J. C. Augusto and C. D. Nugent (Berlin: Springer), 146–164.

Ben Hadj Mohamed, A., Val, T., Andrieux, L., and Kachouri, A. (2013). "Assisting people with disabilities through Kinect sensors into a smart house," in *Proceedings of the 2013 International Conference on Computer Medical Applications, ICCMA*, Sousse.

Bevilacqua, V., Nuzzolese, N., Barone, D., Pantaleo, M., Suma, M., D'Ambruoso, D., et al. (2014). "Fall detection in indoor environment with kinect sensor," in *Proceedings 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, Alberobello, 319–324. doi: 10.1109/INISTA.2014.6873638

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer.

Booranrom, Y., Watanapa, B., and Mongkolnam, P. (2014). "Smart bedroom for elderly using kinect," in *Proceedings 2014 International Computer Science and Engineering Conference (ICSEC)*, Khon Kaen, 427–432. doi: 10.1109/ICSEC.2014.6978235

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., and Yu, Z. (2012). Sensor-based activity recognition. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 42, 790–808. doi: 10.1109/TSMCC.2012.2198883

Corti, A., Giancola, S., Mainetti, G., and Sala, R. (2016). A metrological characterization of the Kinect V2 time-of-flight camera. *Rob. Auton. Syst.* 75, 584–594. doi: 10.1016/j.robot.2015.09.024

Crandall, A. S., and Cook, D. J. (2010). Using a hidden markov model for resident identification. *Proc. 2010 6th Int. Conf. Intell. Environ. IE* 2010, 74–79. doi: 10.1109/IE.2010.21

Calin, A. D., and Coroiu, A. (2018). "Interchangeability of kinect and orbbec sensors for gesture recognition," in *Proceedings of the IEEE 14th International Conference on Intelligent Computer*, Cluj-Napoca, 309–315. doi: 10.1109/ICCP.2018.8516586

Damaševičius, R., Vasiljevas, M., Šalkevičius, J., and Woźniak, M. (2016). Human activity recognition in AAL environments using random projections. *Comput. Math. Methods Med.* 2016:4073584. doi: 10.1155/2016/4073584

Datasets – Advise (2019). Available online at: https://iiw.kuleuven.be/onderzoek/advise/datasets#High Quality Fall Simulation Data (accessed October 18, 2019).

Debes, C., Merentitis, A., Sukhanov, S., Niessen, M., Frangiadakis, N., and Bauer, A. (2016). Monitoring activities of daily living in smart homes: understanding human behavior. *IEEE Signal. Process. Mag.* 33, 81–94. doi: 10.1109/MSP.2015.2503881

Delahoz, Y. S., and Labrador, M. A. (2014). Survey on fall detection and fall prevention using wearable and external sensors. *Sensors* 14, 19806–19842. doi: 10.3390/s141019806

Du, Y., Wang, W., and Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceeding 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 1110–1118. doi: 10.1109/CVPR.2015.7298714

Fall detection Dataset (2019). *Fall Detection Dataset - Le2i - Laboratoire Electronique, Informatique et Image*. Available online at: http://le2i.cnrs.fr/Fall-detection-Dataset?lang=fr (accessed October 18, 2019).

Fall detection testing dataset (2019). *Fall Detection Testing Dataset | MEBIOMEC*. Available online at: https://mebiomec.ai2.upv.es/filedepot_folder/fall-detection-testing-dataset (accessed October 18, 2019).

Fankhauser, P., Bloesch, M., Rodriguez, D., Kaestner, R., Hutter, M., and Siegwart, R. (2015). "Kinect v2 for mobile robot navigation: evaluation and modeling," in *Proceedins of the 2015 International Conference on Advanced Robotics (ICAR)*, Istanbul, 388–394. doi: 10.1109/ICAR.2015.7251485

Gasparrini, S., Cippitelli, E., Spinsante, S., and Gambi, E. (2014). A depth-based fall detection system using a Kinect® sensor. *Sensors* 14, 2756–2775. doi: 10.3390/s140202756

Gasparrini, S., Cippitelli, E., Spinsante, S., and Gambi, E. (2015). Depth cameras in AAL environments: technology and real-world applications. *Gamification Concepts Methodol. Tools Appl.* 2–4, 1056–1075. doi: 10.4018/978-1-4666-8200-9.ch053

Guo, X., Shen, Z., Zhang, Y., and Wu, T. (2019). Review on the application of artificial intelligence in smart homes. *Smart Cities* 2, 402–420. doi: 10.3390/smartcities2030025

Han, J., Shao, L., Xu, D., and Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: a review. *IEEE Trans. Cybern.* 43, 1318–1334. doi: 10.1109/TCYB.2013.2265378

Hanusz, Z., and Tarasińska, J. (2015). Normalization of the kolmogorov–smirnov and shapiro–wilk tests of normality. *Biometr. Lett.* 52, 85–93. doi: 10.1515/bile-2015-2018

Hussein, A., Adda, M., Atieh, M., and Fahs, W. (2014). Smart home design for disabled people based on neural networks. *Proc. Comput. Sci.* 37, 117–126. doi: 10.1016/j.procs.2014.08.020

Jalal, A., Kamal, S., and Kim, D. (2015). "Shape and motion features approach for activity tracking and recognition from kinect video camera," in *Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, Gwangiu, 445–450. doi: 10.1109/WAINA.2015.38

Kim, Y., Baek, S., and Bae, B. C. (2017). Motion capture of the human body using multiple depth sensors. *ETRI J.* 39, 181–190. doi: 10.4218/etrij.17.2816.0045

Le, T. L., Nguyen, M. Q., and Nguyen, T. T. M. (2013). "Human posture recognition using human skeleton provided by Kinect," in *Proceedings of the 2013 International Conference on Computing, Management and Telecommunications (ComManTel)*, Ho Chi Minh City, 340–345. doi: 10.1109/ComManTel.2013.6482417

Li, B., Han, C., and Bai, B. (2019). Hybrid approach for human posture recognition using anthropometry and BP neural network based on Kinect V2. *Eurasip J. Image Video Process* 2019:8. doi: 10.1186/s13640-018-0393-394

Liu, J., Shahroudy, A., Xu, D., Kot, A. C., and Wang, G. (2018). Skeleton-Based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 3007–3021. doi: 10.1109/TPAMI.2017.2771306

Lloret, J., Canovas, A., Sendra, S., and Parra, L. (2015). A smart communication architecture for ambient assisted living. *IEEE Commun. Mag.* 53, 26–33. doi: 10.1109/MCOM.2015.7010512

Majumder, S., Aghayi, E., Noferesti, M., Memarzadeh-Tehran, H., Mondal, T., Pang, Z., et al. (2017). Smart homes for elderly healthcare–Recent advances and research challenges. *Sensors* 17:2496. doi: 10.3390/s17112496

Mastorakis, G., and Makris, D. (2014). Fall detection system using Kinect's infrared sensor. *J. Real Time Image Process.* 9, 635–646. doi: 10.1007/s11554-012-0246-249

Mehr, H. D., Polat, H., and Cetin, A. (2016). "Resident activity recognition in smart homes by using artificial neural networks," in *Proceedings of the 2016 4th International Istanbul Smart Grid Congress and Fair (ICSG)*, Istanbul. doi: 10.1109/SGCF.2016.7492428

Microsoft (2019). *Kinect for Windows SDK 2.0 from Official Microsoft Download Center*. Available online at: https://www.microsoft.com/en-us/download/details.aspx?id=44561 (accessed October 15, 2019).

Müller, M., Röder, T., and Clausen, M. (2005). Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.* 1:677. doi: 10.1145/1186822.1073247

Pagliari, D., and Pinto, L. (2015). Calibration of Kinect for Xbox One and comparison between the two generations of microsoft sensors. *Sensors* 15, 27569–27589. doi: 10.3390/s151127569

Pannurat, N., Thiemjarus, S., and Nantajeewarawat, E. (2014). Automatic fall monitoring: a review. *Sensors* 14, 12900–12936. doi: 10.3390/s140712900

Patsadu, O., Nukoolkit, C., and Watanapa, B. (2012). "Human gesture recognition using Kinect camera," in *Proceeding of the 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, Bangkok. doi: 10.1109/JCSSE.2012.6261920

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 976–990. doi: 10.1016/j.imavis.2009.11.014

Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M., and Meunier, J. (2011). "Fall detection from depth map video sequences," in *Toward Useful Services for Elderly and People with Disabilities. ICOST 2011. Lecture Notes in Computer Science*, eds B. Abdulrazak, S. Giroux, B. Bouchard, H. Pigot, and M. Mokhtari (Berlin: Springer), doi: 10.1007/978-3-642-21535-3_16

Sarbolandi, H., Lefloch, D., and Kolb, A. (2015). Kinect range sensing: structured-light versus time-of-flight kinect. *Comput. Vis. Image Underst.* 139, 1–20. doi: 10.1016/j.cviu.2015.05.006

Sell, J., and O'Connor, P. (2014). The xbox one system on a chip and kinect sensor. *IEEE Micro* 34, 44–53. doi: 10.1109/MM.2014.9

Torti, E., Fontanella, A., Musci, M., Blago, N., Pau, D., Leporati, F., et al. (2019). Embedding recurrent neural networks in wearable systems for real-time fall detection. *Microprocess. Microsyst.* 71:102895. doi: 10.1016/j.micpro.2019.102895

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based feature selection: introduction and review. *J. Biomed. Inform.* 85, 189–203. doi: 10.1016/j.jbi.2018.07.014

Verrini, F., Crispim, C. F., Chessa, M., Solari, F., and Bremond, F. (2018). "Recognition of daily activities by embedding hand-crafted features within a semantic analysis," in *Proceedings of the 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, Sophia Antipolis, 210–215. doi: 10.1109/IPAS.2018.8708854

Visutarrom, T., Mongkolnam, P., and Chan, J. H. (2014). "Postural classification using kinect," in *Proceedings of the 2014 International Computer Science and Engineering Conference (ICSEC)*, Khon Kaen, 403–408. doi: 10.1109/ICSEC.2014.6978231

Visutarrom, T., Mongkolnam, P., and Chan, J. H. (2015). "Multiple-stage classification of human poses while watching television," in *Proceedings of the 2014 2nd International Symposium on Computational and Business Intelligence*, New Delhi, 10–16. doi: 10.1109/ISCBI.2014.10

Wang, J., Liu, Z., and Wu, Y. (2014). *Human Action Recognition with Depth Cameras*. Cham: Springer International Publishing, doi: 10.1007/978-3-319-04561-4560

Weblet Importer (2019). Available online at: http://vlm1.uta.edu/~zhangzhong/fall_detection/(accessed October 18, 2019).

Yang, X., and Tian, Y. (2014). Effective 3D action recognition using EigenJoints. *J. Vis. Commun. Image Represent.* 25, 2–11. doi: 10.1016/j.jvcir.2013.03.001

Youness, C., and Abdelhak, M. (2016). Machine learning for real time poses classification using kinect skeleton data. *Proc. Comput. Graph. Imaging Vis. New Tech. Trends* 2016, 307–311. doi: 10.1109/CGiV.2016.66

Analysis of Postural Control Using Principal Component Analysis: The Relevance of Postural Accelerations and of Their Frequency Dependency for Selecting the Number of Movement Components

Arunee Promsri[1,2] and Peter Federolf[1]*

[1] Department of Sport Science, University of Innsbruck, Innsbruck, Austria, [2] Department of Physical Therapy, University of Phayao, Phayao, Thailand

One criterion when selecting the number of principal components (PCs) to be considered in a principal component analysis (PCA) is the fraction of overall variance that each PC represents. When applying a PCA to kinematic marker data in postural control research, this criterion relates to the amplitude of postural changes, recently often called "principal (postural) positions" (PPs). However, in the assessment of postural control, important aspects are also how fast posture changes and the acceleration of postural changes, i.e., "principal accelerations" (PAs). The current study compared how much of the total position variance each PP explained (PP_rVAR) and how much of the total acceleration variance each PA explained (PA_rVAR). Furthermore, the frequency content of PP and PA signals were evaluated. Postural movements of 26 participants standing on stable ground or balancing on a multiaxial balance board were analyzed by applying a PCA on 90 marker coordinates. For each PC, PP_rVAR, PA_rVAR, and the Fourier transformations of the PP and PA time series were calculated. The PP_rVAR and the PA_rVAR-distributions differed substantially. The PP-frequency domain was observed well below 5 Hz, the PA-frequency domain up to 5 Hz for stable standing and up to 10 Hz on the balance board. These results confirm that small-amplitude but fast movement components can have a higher impact on postural accelerations—and thus on the forces active in the system—than large-amplitude but slow lower-order movement components. Thus, PA variance and its dependence on filter frequencies should be considered in dimensionality reduction decisions.

Keywords: movement strategy, neuromuscular control, filtering, frequency analysis, principal component analysis PCA, principal acceleration

INTRODUCTION

Principal component analysis (PCA) is an unsupervised data analysis procedure often used as a preprocessing step, e.g., to improve performance or for dimensionality reduction, before more complex machine learning procedures are applied. If applied in the analysis of human motion, a PCA can by itself reveal interesting information about the coordinative structure of complex

whole-body movements. Accordingly, applying a PCA on kinematic data has received increasing attention in research on several kinds of human movements, such as reaching (Longo et al., 2019), karate kicking (Zago et al., 2017a), juggling (Zago et al., 2017b), skiing (Federolf et al., 2014; Gløersen et al., 2018; Pellegrini et al., 2018), or walking (Troje, 2002; Daffertshofer et al., 2004; Verrel et al., 2009; Zago et al., 2017c).

One of the main purposes for performing a PCA on kinematic data—or in fact on any dataset—is the idea that the entire variance in the data can often be approximated to high accuracy with only a limited number of principal components (PCs). One of the most common criteria for choosing the number of PCs is the eigenvalue spectrum, which represents the variance explained by each PC and which can be expressed in relative values, i.e., as a percentage of the entire variance in the data.

A research area where PCA has been particularly frequently applied on kinematic human movement data is research on postural control (Federolf et al., 2013; Federolf, 2016; Haid and Federolf, 2018; Haid et al., 2018, 2019; Promsri et al., 2018, 2019, 2020a,b; Wachholz et al., 2019a,b). In postural control studies, when PCA is applied to kinematic data, it decomposes the complex multi-segment whole-body movements into a set of one-dimensional movement components, called "principal movements" PM_k, where k is the order of the movement component (Federolf et al., 2013; Federolf, 2016). Previous research has shown that the lower-order PM_k represent in close approximation the classical motor strategies (Horak and Nashner, 1986; Winter, 1995), i.e., the ankle or hip strategies (Federolf, 2013). If PCA is calculated on normalized data from different volunteers, then a subject-specific *relative explained variance* can be calculated in analogy to the eigenvalues, which quantify the explained variance for the whole dataset and are thus not subject-specific. The *relative explained variance*-spectra provide one criterion for how many movement components PM_k one wants to consider in the analysis (Federolf, 2013; Haid et al., 2019).

However, analyzing the different postures observed during a measurement sequence may not be the only variable of interest. How fast the posture changes and how much a postural change is accelerated, also provide valuable information. We have shown in previous papers, that Newton's mechanics can be applied to the PCA-based posture space by defining a "principal (postural) position" (PP_k) for each PM and their time derivatives, principal velocity (PV_k) and principal acceleration (PA_k) (Federolf, 2016; Longo et al., 2019). The PA_k are of interest, since they relate to forces acting in the system and thus to the neuromuscular control of the postural movements (Federolf, 2016; Haid et al., 2018; Promsri et al., 2018, 2019, 2020a,b; Haid and Federolf, 2019; Wachholz et al., 2019a,b, 2020). We want to emphasize here that the PA_k obtained by double-differentiation of the PP_k time series (Federolf, 2016; Longo et al., 2019) are different variables than when a PCA is performed directly on acceleration data (Verheul et al., 2019): The former can be seen as an expansion of the movement strategy concept (Horak and Nashner, 1986; Winter, 1995), since the PA_k quantify the acceleration of the considered movement components/movement strategies; the latter PCA

identifies correlated patterns directly in acceleration data, which yields a different solution.

Differentiation is a non-linear operation and, consequently, the relative variance spectra of the PP_k differ from the PA_k relative variance spectra (Longo et al., 2019). Particularly in postural control it is likely that large-amplitude, yet slow movement components influence the PA-spectrum less than small-amplitude, but fast movement components. The PA-explained variance spectrum could be a second important criterion for the decision on how many PM_k should be considered in an analysis (Longo et al., 2019). Unfortunately, noise amplification in differentiation makes a filtering of the PP_k signals necessary before PV_k and PA_k are calculated (Winter et al., 1974), and since the PA_k variance spectra are speed-dependent, they will change with the filter cut-off frequency used before the differentiation.

In summary, when applying a PCA to investigate the coordinative structure of postural control movements, both the principal positions (PP_k) and the principal accelerations (PA_k) are of interest since they provide relevant information on the composition of the postural movements and on the control of the movement components, respectively (Promsri et al., 2020a). Both the PP_k- and PA_k-spectra should be considered when selecting the number of movement components to be analyzed, however, the PA_k-spectra are speed- and thus filter frequency-dependent. Thus, the purposes of the current Brief Research Report were (i) to compare the PP_k and PA_k relative variance spectra for postural control data; (ii) to evaluate the frequency content of the PP_k and PA_k time series; and (iii) to assess how the PA relative variance spectrum depends on the filtering cut-off frequency.

MATERIALS AND METHODS

Participants

Twenty-six physically active young adults (14/12 males/females, age 25.3 ± 4.2 years, weight 70.7 ± 11.4 kg, height 175.0 ± 8.1 cm, physical activity participation 8.4 ± 5.4 h/weeks [mean $\pm$ SD]) with no neuromuscular injuries/disorders and no specific balance training participated in the current study. All volunteers provided informed consent and the study protocol had been approved by the Board of Ethical Questions in Science of the University of Innsbruck, Austria (Certificate 16/2016).

Measurement Procedures

Participants were equipped with 39 reflective-markers according to the "Plug-In Gait" marker setup (Vicon Motion Systems Ltd., Oxford, UK). Two 80-s barefooted-bipedal balancing trials, one for each support surface, were completed in randomized order on a firm surface (FS) and on a wobble board (WB; Powrx Balance Board; POWRX GmbH., Germany). After completing the first trial, participants could rest for up to 3 min. For the WB condition, volunteers had a 15-s familiarization trial with no instruction or feedback. Postural movement trajectories were captured by a standard 8-camera motion tracking system (Vicon Bonita B10 cameras with Nexus 2.2.3 software; Vicon Motion Systems Ltd., Oxford, UK) using a sampling rate of 250 Hz.

To standardize the standing position (**Supplementary Figure 1**), participants were asked to place two marked points (base of each 2nd metatarsal bone) over a horizontal line taped on the floor for FS or over a horizontal diameter of the WB; to align the inside of the feet (the medial borders of each distal end of the first metatarsal bone) with tapes defining an individual inter-feet distance (15% of biacromial diameter); to rest their hands on the hips; and to look straight ahead at a 10-cm-diameter red-circle target on a wall at the individual eye level ~5 m away. To standardize the position of the wobble board, we placed the center of the wobble board over the center of a reticle cross-line marked on the floor. During testing, volunteers were asked to stand still for the FS or to keep the board horizontal for the WB; to avoid any voluntary movements; and to keep their eyes on the target.

Data Analysis
Kinematic Data Pre-processing
All data processing was conducted in Matlab (MathWorks Inc., Natrick, MA, USA). The pre-processing steps and the PCA analysis were conducted based on earlier studies (Federolf, 2016; Promsri et al., 2018, 2019, 2020a). Briefly, any gaps in marker trajectories were filled by a PCA-based reconstruction technique (Federolf, 2013; Gløersen and Federolf, 2016). Two PCAs were performed, one for each balancing condition (Promsri et al., 2020a). The middle 60 seconds of each balancing trial were extracted and nine asymmetrical markers placed on the upper arms, lower arms, right scapular, upper thighs, and the lower thighs were omitted. In analogy to previous studies (Troje, 2002; Daffertshofer et al., 2004; Verrel et al., 2009; Federolf, 2016), the 3D coordinates (x, y, z) of the remaining 30 markers of each dataset at a given time t were interpreted as 90-dimensional posture vectors:

$$\vec{p(t)} = \left[x_1(t), y_1(t), z_1(t), \ldots, x_{30}(t), y_{30}(t), z_{30}(t) \right] \quad (1)$$

Three pre-processing steps were then conducted. First, the posture vectors were centered by subtracting the subject's mean posture vector. For each subject, $subj$, a mean posture vector:

$$\overrightarrow{p^{subj}} = \left[\overline{x_1}, \overline{y_1}, \ldots, \overline{z_{30}} \right] \quad (2)$$

where the bar over the variable indicates the mean over time, $x = mean_t(x(t))$, was subtracted from each posture vector:

$$\vec{p'(t)} = \vec{p(t)} - \overrightarrow{p^{subj}} \quad (3)$$

This procedure is the first step toward removing anthropometric differences (Federolf, 2016). The PCA was, therefore, conducted on deviations from a subject's mean posture, i.e., on postural movements. Second, the centered posture vectors were normalized to the mean Euclidean distance d^{subj} (Federolf, 2013, 2016). Thus, for each posture vector $\vec{p'(t)}$ the Euclidean norm:

$$d^{subj}(t) = \left(\sqrt{x_1(t)^2 + y_1(t)^2 + z_1(t)^2 + \ldots + z_{30}(t)^2} \right) \quad (4)$$

was calculated and the $\vec{p'(t)}$ were then divided by the mean of these Euclidian distances:

$$\vec{p''(t)} = \frac{1}{d^{subj}} \vec{p'(t)}. \quad (5)$$

Third, the normalized posture vectors were weighted using sex-specific mass distributions (Gløersen et al., 2018). Specifically, for each marker i a weight factor w_i was calculated by dividing the relative weight of the segment to which the marker was attached, m_s, by the number n_s of markers on this segment. For markers placed on joints, the masses of both segments were added. For example, w_i for the knee markers was calculated as $w_i = \frac{m_{thigh}}{n_{thigh}} + \frac{m_{shin}}{n_{shin}}$ with $n_{thigh} = n_{shin} = 3$, $m_{thigh} = 14.16\%$, and $m_{shin} = 4.33\%$ for men (de Leva, 1996). Thus, the normalized postural movement vectors had the form:

$$\vec{p'''(t)} = \frac{1}{d^{subj}} \left[w_1 \left(x_1(t) - \overline{x_1^{subj}} \right), \quad w_1 \left(y_1(t) - \overline{y_1^{subj}} \right), \right. \\ \left. \ldots, w_{30} \left(z_{30}(t) - \overline{z_{30}^{subj}} \right) \right] \quad (6)$$

TABLE 1 | The relative explained variance of principal position PP_rVAR (%) and a qualitative description of the movement patterns represented by the first ten principal movements (PM_{1-10}).

PM	PP_rVAR (%)	Main movements
A: Firm surface		
1	64.4 ± 16.3	Anteroposterior ankle strategy
2	19.9 ± 13.5	Mediolateral ankle strategy (lateral weight shift)
3	4.8 ± 3.8	Anteroposterior hip strategy
4	2.3 ± 1.9	Transverse rotation of pelvis and upper body
5	1.8 ± 1.8	Vertical breathing movement patterns
6	1.3 ± 0.8	Vertical breathing movement patterns
7	1.0 ± 1.4	Anteroposterior trunk flexion coupled with knee flexion and extension
8	0.7 ± 1.0	Breathing; small chest movements
9	0.6 ± 0.7	Retraction and protraction of shoulders
10	0.4 ± 0.4	Upper body movement
B: Wobble board		
1	29.9 ± 8.5	Mediolateral ankle strategy (lateral weight shift)
2	24.2 ± 8.3	Anteroposterior ankle strategy
3	17.9 ± 10.5	Transverse; twisting of the board coupled with the whole-body rotation
4	11.7 ± 8.2	Anteroposterior hip strategy coupled with anteroposterior ankle strategy
5	5.0 ± 3.0	Vertical hip, knee, and ankle flexion/extension
6	3.7 ± 1.3	Mediolateral hip strategy coupled with mediolateral ankle strategy
7	1.6 ± 0.8	Diagonal lateral weight shift
8	1.1 ± 0.6	Anteroposterior ankle plantarflexion/dorsiflexion
9	1.1 ± 1.9	Twisting of the board coupled with the whole-body rotation
10	0.5 ± 0.3	Lateral weight shift and with small rotation

*Together, PM_{1-10} explained 97.2 and 96.7% of the overall postural variance when balancing on (**A**) firm surface, and (**B**) wobble board, respectively.*

Then, the normalized posture vectors $\overrightarrow{p'''}(t)$ from all volunteers were concatenated to form a 390,000 × 90-PCA input-matrix (250 [sampling rate] * 60 [trial duration] * 26 [number of subjects] × 90 [marker coordinates]).

Principal Component Analysis

The PCA was calculated by a singular-value decomposition of the input matrix's covariance matrix and produced a set of PC-eigenvectors, $\overrightarrow{PC_k}$, which form a new basis for the vector space of marker positions (Haid et al., 2019). All PC-eigenvectors are linear combinations of the original marker coordinates. Animated stick figures can be created from the mean postures and from each eigenvector to characterize the principal movements PM_k (Federolf et al., 2013; Federolf, 2016). The time evolution of each PM_k, i.e., the $PP_k(t)$, were obtained by a coordinate transformation of the normalized posture vectors onto

FIGURE 1 | Illustration of the first ten principal movements (PM_{1-10}) of bipedal standing on (**A**) the firm surface and bipedal balancing on (**B**) the wobble board. Gray and black lines/dots show the extreme posture in opposite directions. Movement amplitudes are amplified using the indicated factor for a better visualization (Firm surface: amplification 10× for PM_{1-5}, and 20× for PM_{6-10}; Wobble board: amplification 1× for PM_{1-5}, and 2× for PM_{6-10}). Movements are clearer and can be more easily characterized when viewed in animated stick figure videos: **Supplementary Videos S1, S2** for balancing on the firm and soft surfaces, respectively.

the PCA-eigenvectors.

$$PP_k(t) = \overrightarrow{p'''}(t) \cdot \overrightarrow{PC_k}$$

The $PP_k(t)$ represent positions in posture space, i.e., how much the posture at time t deviates in the direction of the PC_k-eigenvector from the mean posture (Federolf, 2016). In other words, the $PP_k(t)$ represent the amplitude of each movement component PM_k. The variance of each $PP_k(t)$, divided by the sum of the variances of all $PP_k(t)$, results in a variable *relative explained variance of principal position PP_rVAR_k* that quantifies for each volunteer and each order k, how much the specific PM_k contributed to the whole postural movements of the subject.

In analogy to Newton's mechanics and differentiation rules, the rate of postural change can be quantified by principal velocities $PV_k(t)$, i.e., by the first time derivative of the $PP_k(t)$, $\mathbf{PV_k} = \frac{\mathbf{d}}{\mathbf{dt}} \mathbf{PP_k}$; and the acceleration of postural movements can be quantified by principal accelerations $PA_k(t)$, i.e., by the second time derivative of the $PP_k(t)$, $\mathbf{PA_k} = \frac{\mathbf{d^2}}{\mathbf{dt^2}} \mathbf{PP_k}$ (Federolf, 2016). In case of unperturbed human postural control, $PA_k(t)$ are either a direct result of muscle activation, a result of the neuromuscular system utilizing gravity to produce desired accelerations, or an undesired result of gravity which the neuromuscular system was not able to prevent e.g., loss of stability (Promsri et al., 2020a). In this sense, the $PA_k(t)$ are the essential mechanical variables that the sensorimotor system must control in order to govern the body's motion and maintain its stability. Thus, each $PA_k(t)$ represents a variable that quantifies how the mechanical system is controlled (Federolf, 2016; Promsri et al., 2020a). In analogy to PP_rVAR_k, we calculated the variable *relative explained variance of principal acceleration PA_rVAR_k* to assess how much each movement component contributed to the overall postural accelerations in the individual subjects.

Due to noise amplification in the differentiation processes (Winter et al., 1974), filtering of the $PP_k(t)$ is needed before computing $PV_k(t)$ and $PA_k(t)$. The current study examined the effect of low-pass filtering using a 3rd-order, zero-phase, low-pass Butterworth filter. The Butterworth filter was selected, since it is free of ripples in the pass and stop band. The filter order (3rd) was selected arbitrarily, however, preliminary tests suggested that the filter order has a very small effect on the PA time series. Prior

FIGURE 2 | Box plots representing the data from all 26 participants of **(A)** the relative explained variance of principal postural positions (PP_rVAR_k) and **(B)** the relative explained variance of principal postural accelerations (PA_rVAR_k) of standing on the firm surface (FS) and balancing on the wobble board (WB). The PA_rVAR_k were determined after filtering the data with a 3rd-order 10 Hz low-pass Butterworth filter.

FIGURE 3 | Box plots of the relative explained variance of principal postural acceleration (PA_rVAR_k) of standing on the firm surface (FS) and balancing on the wobble board (WB) with different cut-off frequencies, including **(A)** 1 Hz, **(B)** 2 Hz, **(C)** 5 Hz, **(D)** 10 Hz, **(E)** 20 Hz, and **(F)** no filtering, which were observed from 26 participants (k displays order of principal components, PMs; k = 1 to 25). The letter, "A," and its arrows point to lower-order PAs, PA_rVAR_1, and PA_rVAR_2, whose contribution to the overall acceleration variance decrease with increasing cutoff frequencies. The letter "B" and its arrows highlight two medium-order PAs, PA_rVAR_8, and PA_rVAR_{10}, whose contribution to the overall acceleration variance increase as cutoff frequencies are increased.

to filtering, the frequency contents of the raw $PP_k(t)$ and $PA_k(t)$ were evaluated using a Fourier transformation. Then, the effect of cut-off frequency on PA_rVAR_k was evaluated for both balancing situations, FS and WB, with cut-off frequencies of 1, 2, 5, 10, and 20 Hz and with no filtering. Finally, explained variance spectra of PP_rVAR_k and PA_rVAR_k (10 Hz) were compared.

RESULTS

The first 10 principal movements (PM_{1-10}) of standing on a firm surface (FS) and balancing on a wobble board (WB) are described and shown in (**Table 1**, **Figure 1**), and in (**Supplementary Videos 1, 2**). Higher-order movement components were not included for the visualization and description, since their small movement amplitudes make them difficult to characterize, however, higher-order components were considered in the evaluation of the variance spectra. The spectra of explained variance, PP_rVAR_k and PA_rVAR_k (for a cut-off frequency of 10 Hz) are shown in (**Figure 2**). As expected, several movement components that contributed little to the postural variance did have an over-proportional contribution to the acceleration variance. Specifically, for standing on the FS, PM_3, PM_8, and PM_{10} which predominantly represented hip strategy and upper body movements, and for balancing on the WB PM_8 which predominantly quantified ankle plantar/dorsiflexion, were of particular interest.

Fourier transformations of the raw PP and PA time series of one arbitrarily selected, representative volunteer are shown in (**Supplementary Figure 2**) for FS and in (**Supplementary Figure 3**) for WB. The PP-frequency domain of both FS and WB conditions was observed well below 5 Hz. In contrast, in the PA-spectra, despite the strong and blue-shifted noise, signals are visible in the ranges 0–5 Hz for FS and in the range up to ~10 Hz for WB. In addition, (**Figure 3**) illustrates how the spectrum of explained variance PA_rVAR_k changes with increasing filter cut-off frequency.

DISCUSSION

Our analysis demonstrates that PA_k and PA_k-based variables, here PA_rVAR_k, depend on the filter cut-off applied in the PA calculation. Low cut-off frequencies (<5 Hz) lead to over-pronunciation of slow movement components. As filter frequency is increased (5–20 Hz) a new pattern emerges, in line with the expectation that some of the higher-order movement components contribute more than other movement components to the accelerations. The Fourier analysis of the underlying signals suggests that the pattern emerging with increasing filter cut-offs is not (not only) a consequence of noise increasingly affecting the signal: while the $PP_k(t)$ live in a very low frequency range (<3 Hz), several of the $PA_k(t)$ show a relevant frequency content up to ~5 Hz in FS and up to 10 Hz in the WB conditions. These observations suggest that filter cut-off frequencies of 5–7 Hz for FS and around 10 Hz for WB would be appropriate.

The current findings underpin that (i) when focusing only on the classical movement strategies (lower-order PMs), one might overlook movement components that are small in posture-amplitude, but that can be accelerated fast and thus provide an important contribution to postural control. Spectra of PA-explained variance should be considered when deciding on how many PC-components are included in an analysis. (ii) When interested in neuromuscular control and thus in the accelerations and forces controlling postural movements, then filter frequencies should not be selected below 5 Hz for stable situations and not below 10 Hz for more dynamic balancing trials. The current findings corroborate the findings of Longo et al. (2019), who assessed PA-relative variance in a cyclic upper-body motion. Moreover, Longo et al. (2019) also mathematically validated that all PA_k together (i.e., the sum of all PA_k) represent the entire marker accelerations present in the dataset. The current results also agree with previous studies in which the dependence of PA_k variables on filter cut-off frequencies was assessed, and which reported consistent results for cut-off frequencies in the range 5 to 12 Hz (Haid et al., 2018; Promsri et al., 2018, 2019, 2020a). Furthermore, recent studies on muscle synergies and on coherence between electromyographic signals from different muscles also reported spectra peaking around 9 Hz and posture-related coherence in frequency bands 5–20 Hz (Boonstra et al., 2008, 2015), which supports the assumption that the PA_k signals in this frequency range are of physiological origin and probably not an artifact or noise phenomenon.

The role of movement analysis in monitoring and diagnosing neurodegenerative conditions is increasingly recognized, particularly when combined with machine/deep learning approaches (Buckley et al., 2019). However, how successful such approaches can become depends largely on the information contained in the input data to these algorithms. Disregarding information at an early stage, e.g., due to dimensionality reduction or through filtering, is a form of investigator bias that likely affects even the performance of so-called unsupervised methods. Driven by biomechanical considerations, the current study evaluated what information might be contained in the often disregarded higher-order PC-components. The question, which specific PC_k components are relevant, depends on the specific movement, the specific boundary conditions that are present, and the research question that is studied. However, as general advise the current study suggest that PA_rVAR_k-spectra should be analyzed when deciding on how many PC components are to be considered; and the frequency content and suitable filters should be carefully assessed in the calculation of PAs.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study reported here was approved by the Board of Ethical Questions in Science of the University of Innsbruck, Austria

(Certificate 16/2016). All participants provided informed written consent prior to their participation.

AUTHOR CONTRIBUTIONS

AP and PF have contributed equally to the design and implementation of the research and to the writing of the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00480/full#supplementary-material

Supplementary Video 1 | Principal movements of bipedal standing.

Supplementary Video 2 | Principal movements of bipedal standing on a wobble board.

REFERENCES

Boonstra, T. W., Daffertshofer, A., van Ditshuizen, J. C., van den Heuvel, M. R. C., Hofman, C., Willigenburg, N. W., et al. (2008). Fatigue-related changes in motor-unit synchronization of quadriceps muscles within and across legs. *J. Electromyogr. Kinesiol.* 18, 717–731. doi: 10.1016/j.jelekin.2007.03.005

Boonstra, T. W., Danna-Dos-Santos, A., Xie, H. B., Roerdink, M., Stins, J. F., and Breakspear, M. (2015). Muscle networks: connectivity analysis of EMG activity during postural control. *Sci. Rep.* 5:17830. doi: 10.1038/srep17830

Buckley, C., Alcock, L., McArdle, R., Ur Rehman, R. Z., Del Din, S., Mazzà, C., et al. (2019). The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: insights from gait and postural control. *Brain Sci.* 9:34. doi: 10.3390/brainsci9020034

Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., and Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clin. Biomech.* 19, 415–428. doi: 10.1016/j.clinbiomech.2004.01.005

de Leva, P. (1996). Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. *J. Biomech.* 29, 1223–1230. doi: 10.1016/0021-9290(95)00178-6

Federolf, P., Reid, R., Gilgien, M., Haugen, P., and Smith, G. (2014). The application of principal component analysis to quantify technique in sports. *Scand. J. Med. Sci. Sports* 24, 491–499. doi: 10.1111/j.1600-0838.2012.01455.x

Federolf, P., Roos, L., and Nigg, B. M. (2013). Analysis of the multi-segmental postural movement strategies utilized in bipedal, tandem and one-leg stance as quantified by a principal component decomposition of marker coordinates. *J. Biomech.* 46, 2626–2633. doi: 10.1016/j.jbiomech.2013.08.008

Federolf, P. A. (2013). A novel approach to solve the "missing marker problem" in marker-based motion analysis that exploits the segment coordination patterns in multi-limb motion data. *PLoS ONE* 8:e78689. doi: 10.1371/journal.pone.0078689

Federolf, P. A. (2016). A novel approach to study human posture control: "principal movements" obtained from a principal component analysis of kinematic marker data. *J. Biomech.* 49, 364–370. doi: 10.1016/j.jbiomech.2015.12.030

Gløersen, Ø., and Federolf, P. (2016). Predicting missing marker trajectories in human motion data using marker intercorrelations. *PLoS ONE* 11:e0152616. doi: 10.1371/journal.pone.0152616

Gløersen, Ø., Myklebust, H., Hallén, J., and Federolf, P. (2018). Technique analysis in elite athletes using principal component analysis. *J. Sports Sci.* 36, 229–237. doi: 10.1080/02640414.2017.1298826

Haid, T., and Federolf, P. (2018). Human postural control: assessment of two alternative interpretations of center of pressure sample entropy through a principal component factorization of whole-body kinematics. *Entropy* 20:30. doi: 10.3390/e20010030

Haid, T., and Federolf, P. (2019). The effect of cognitive resource competition due to dual-tasking on the irregularity and control of postural movement components. *Entropy* 21:70. doi: 10.3390/e21010070

Haid, T. H., Doix, A.-C. M., Nigg, B. M., and Federolf, P. A. (2018). Age effects in postural control analyzed via a principal component analysis of kinematic data and interpreted in relation to predictions of the optimal feedback control theory. *Front. Aging Neurosci.* 10:22. doi: 10.3389/fnagi.2018.00022

Haid, T. H., Zago, M., Promsri, A., Doix, A.-C. M., and Federolf, P. A. (2019). PManalyzer: a software facilitating the study of sensorimotor control of whole-body movements. *Front. Neuroinform.* 13:24. doi: 10.3389/fninf.2019.00024

Horak, F. B., and Nashner, L. M. (1986). Central programming of postural movements: adaptation to altered support-surface configurations. *J. Neurophysiol.* 55, 1369–1381. doi: 10.1152/jn.1986.55.6.1369

Longo, A., Haid, T., Meulenbroek, R., and Federolf, P. (2019). Biomechanics in posture space: Properties and relevance of principal accelerations for characterizing movement control. *J. Biomech.* 82, 397–403. doi: 10.1016/j.jbiomech.2018.11.031

Pellegrini, B., Zoppirolli, C., Boccia, G., Bortolan, L., and Schena, F. (2018). Cross-country skiing movement factorization to explore relationships between skiing economy and athletes' skills. *Scand. J. Med. Sci. Sports* 28, 565–574. doi: 10.1111/sms.12938

Promsri, A., Haid, T., and Federolf, P. (2018). How does lower limb dominance influence postural control movements during single leg stance? *Hum. Mov. Sci.* 58, 165–174. doi: 10.1016/j.humov.2018.02.003

Promsri, A., Haid, T., and Federolf, P. (2020a). Complexity, composition, and control of bipedal postural control system adapts to unstable support surfaces or altered feet positions. *Neuroscience* 430, 113–124. doi: 10.1016/j.neuroscience.2020.01.031

Promsri, A., Haid, T., Werner, I., and Federolf, P. (2020b). Leg dominance effects on postural control when performing challenging balance exercises. *Brain Sci.* 10:128. doi: 10.3390/brainsci10030128

Promsri, A., Longo, A., Haid, T., Doix, A. C. M., and Federolf, P. (2019). Leg dominance as a risk factor for lower-limb injuries in downhill skiers—a pilot study into possible mechanisms. *Int. J. Environ. Res. Public Health* 16:3399. doi: 10.3390/ijerph16183399

Troje, N. F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *J. Vis.* 2, 371–387. doi: 10.1167/2.5.2

Verheul, J., Warmenhoven, J., Lisboa, P., Gregson, W., Vanrenterghem, J., and Robinson, M. A. (2019). Identifying generalised segmental acceleration patterns that contribute to ground reaction force features across different running tasks. *J. Sci. Med. Sport.* 22, 1355–1360. doi: 10.1016/j.jsams.2019.07.006

Verrel, J., Lövdén, M., Schellenbach, M., Schaefer, S., and Lindenberger, U. (2009). Interacting effects of cognitive load and adult age on the regularity of whole-body motion during treadmill walking. *Psychol. Aging* 24, 75–81. doi: 10.1037/a0014272

Wachholz, F., Kockum, T., Haid, T., and Federolf, P. (2019a). Changed temporal structure of neuromuscular control, rather than changed intersegment coordination, explains altered stabilographic regularity after

a moderate perturbation of the postural control system. *Entropy* 21:614. doi: 10.3390/e21060614

Wachholz, F., Tiribello, F., Mohr, M., van Andel, S., and Federolf, P. (2020). Adolescent awkwardness : alterations in temporal control characteristics of posture with maturation and the relation to movement exploration. *Brain Sci.* 10:E216. doi: 10.3390/brainsci10040216

Wachholz, F., Tiribello, F., Promsri, A., and Federolf, P. (2019b). Should the minimal intervention principle be considered when investigating dual-tasking effects on postural control? *Brain Sci.* 10:1 doi: 10.3390/brainsci10 010001

Winter, D. A. (1995). Human balance and posture control during standing and walking. *Gait Posture* 3, 193–214. doi: 10.1016/0966-6362(96)82849-9

Winter, D. A., Sidwall, H. G., and Hobson, D. A. (1974). Measurement and reduction of noise in kinematics of locomotion. *J. Biomech.* 7, 157–159. doi: 10.1016/0021-9290(74)90056-6

Zago, M., Codari, M., Iaia, F. M., and Sforza, C. (2017a). Multi-segmental movements as a function of experience in karate. *J. Sports Sci.* 35, 1515–1522. doi: 10.1080/02640414.2016.1223332

Zago, M., Pacifici, I., Lovecchio, N., Galli, M., Federolf, P. A., and Sforza, C. (2017b). Multi-segmental movement patterns reflect juggling complexity and skill level. *Hum. Mov. Sci.* 54, 144–153. doi: 10.1016/j.humov.2017.04.013

Zago, M., Sforza, C., Bona, A., Cimolin, V., Costici, P. F., Condoluci, C., et al. (2017c). How multi segmental patterns deviate in spastic diplegia from typical developed. *Clin. Biomech.* 48, 103–109. doi: 10.1016/j.clinbiomech.2017.07.016

Check for updates

Recognition of Foot-Ankle Movement Patterns in Long-Distance Runners With Different Experience Levels Using Support Vector Machines

Eneida Yuri Suda, Ricky Watari, Alessandra Bento Matias and Isabel C. N. Sacco*

Physical Therapy, Speech and Occupational Therapy Department, School of Medicine, University of São Paulo, São Paulo, Brazil

Running practice could generate musculoskeletal adaptations that modify the body mechanics and generate different biomechanical patterns for individuals with distinct levels of experience. Therefore, the aim of this study was to investigate whether foot-ankle kinetic and kinematic patterns can be used to discriminate different levels of experience in running practice of recreational runners using a machine learning approach. Seventy-eight long-distance runners (40.7 ± 7.0 years) were classified into less experienced ($n = 24$), moderately experienced ($n = 23$), or experienced ($n = 31$) runners using a fuzzy classification system, based on training frequency, volume, competitions and practice time. Three-dimensional kinematics of the foot-ankle and ground reaction forces (GRF) were acquired while the subjects ran on an instrumented treadmill at a self-selected speed (9.5–10.5 km/h). The foot-ankle kinematic and kinetic time series underwent a principal component analysis for data reduction, and combined with the discrete GRF variables to serve as inputs in a support vector machine (SVM), to determine if the groups could be distinguished between them in a one-vs.-all approach. The SVM models successfully classified all experience groups with significant crossvalidated accuracy rates and strong to very strong Matthew's correlation coefficients, based on features from the input data. Overall, foot mechanics was different according to running experience level. The main distinguishing kinematic factors for the less experienced group were a greater dorsiflexion of the first metatarsophalangeal joint and a larger plantarflexion angles between the calcaneus and metatarsals, whereas the experienced runners displayed the opposite pattern for the same joints. As for the moderately experienced runners, although they were successfully classified, they did not present a visually identifiable running pattern, and seem to be an intermediate group between the less and more experienced runners. The results of this study have the potential to assist the development of training programs targeting improvement in performance and rehabilitation protocols for preventing injuries.

Keywords: running, machine learning, running experience, biomechanics, fuzzy logic

INTRODUCTION

Running is a very popular activity and its practice has been increasing in the last decades because of its accessibility and several benefits (Lee et al., 2017). It is a repetitive activity which results in minor load variations at each step (Davis and Futrell, 2016). These repeated loads that occur during running have beneficial effects over the foot musculoskeletal structures, such as increases in foot muscle volume and cross-sectional area, and in bone density (Garofolini and Taylor, 2019). However, repetitive loading can also make musculoskeletal tissues from the lower limbs more susceptible to cumulative overload and, therefore, overuse injuries (Davis and Futrell, 2016).

Notably, the foot-ankle forms a dynamic link between the body and the ground (Rodgers, 1988), being the first segment to provide this interaction and being responsible for the accommodation for the irregularities of the supporting surfaces. Additionally, the foot-ankle complex contributes to the dissipation of energy returned from the ground and to the attenuation of high impact forces during foot strike in running by having many kinematic adjustments in its more than 33 joints, participating in propulsion generation during push-off (Holowka and Lieberman, 2018), and storing and releasing elastic energy during stance phase (Holowka and Lieberman, 2018; Kelly et al., 2018). Altogether, these features make the structure and function of the foot-ankle complex extremely important for running practice. There are evidences that barefoot running, which enhances strength of the plantarflexors and foot intrinsic muscles (Lieberman, 2012), might serve as protection against knee injuries (Altman and Davis, 2016), one of the most common running-related injury (van Gent et al., 2007). Therefore, the study of possible changes that might occur in foot-ankle biomechanics with running practice gains importance.

Running experience appears to be protective against injuries (Macera, 1992; Nielsen et al., 2012; van der Worp et al., 2015; Hulme et al., 2017; Linton and Valentin, 2018). Videbæk et al. (2015) showed that novice runners have a significantly greater risk of injury than recreational runners, who run regularly and participate in short races (less than 10 km), showing an incidence of 17.8 per 1000 h of running against 7.7, respectively. Besides that, novice runners have the majority of injuries in the lower leg (34.7%), 3.5% in the foot and 8.2% in the ankle, while recreational runners present the majority of injuries in the knee (26.3%), 10.1% in the foot and 7.8% in the ankle (Kluitenberg et al., 2015). Those differences in injury incidence and distribution point out that these might be distinct populations, who could present different running mechanics that can be related to experience level and could be a protective factor for running-related injuries.

However, most studies classify experience based solely on years of practice, or just classify the subjects between "novice" or "experienced," leaving the reasons behind this possible protective effect unknown. Although development of expertise can be achieved through deliberate practice, i.e., by performing a task in a way that provides effective skill acquisition (Ericsson et al., 1993), defining running experience just as years of practice does not take into account the structure of the deliberate practice (Iglesias et al., 2010). In this context, Roveri et al. (2017) developed a decision support system to classify experience in the deliberate practice of running that takes into account the structure of the practice, including training frequency, training volume, years of practice, and participation in races as inputs in the algorithm. Therefore, using this classification approach could provide a more complete and objective measure of experience levels, allowing the study of how the biomechanical adaptations in running might be transitioning across the experience acquisition process. Furthermore, future studies could take advantage of these findings to investigate what are the implications of the potential differences in biomechanics to running performance and injuries.

Some studies suggest that running experience does not influence running biomechanics. Agresta et al. (2018) did not find any influence of years of running experience in trunk and lower limb kinematics, spatiotemporal variables, nor GRF variables during running. Similarly, Schmitz et al. (2014) did not find differences between novice and experienced runners in impact peak, loading rate, nor peak hip adduction angle during running. However, Clermont et al. (2017) compared 3D running kinematics of the pelvis, hip, knee and ankle of recreational and competitive runners by means of a combination of principal component analysis and a SVM classifier and determined that it is possible to distinguish both groups based on the differences in pelvic tilt, knee flexion and ankle eversion. Hence, there is evidence that multivariate analysis combined with machine learning approaches can be an effective analysis method for identifying mechanical patterns of running. Nevertheless, there is still a lack of understanding of how the different segments of the foot-ankle complex participate in these biomechanical adaptations, given that there is a functional importance of this multi-joint complex during running regarding energy dissipation, attenuation of forces, energy storage and release during stance, and propulsion generation.

In this context, studying the effects of experience on the mechanical behavior of the foot-ankle can contribute to the better understanding of how it is influenced by the skill acquisition. This understanding could give insights on what type of training or rehabilitation protocol could enhance performance or prevent and treat injuries. Therefore, the main purpose of this study is to determine if it is possible to separate and classify groups with distinct levels of experience, determined by a system that takes into account different aspects of running practice, based on foot-ankle kinematics and kinetic patterns, and impact variables. We hypothesize that there will be different foot-ankle mechanical patterns that can distinguish between levels of running experience, identified by a machine learning approach.

MATERIALS AND METHODS

This study was a retrospective secondary analysis of a subset of data from a larger randomized clinical trial approved by the Research and Ethics Committee of the School of Medicine,

University of São Paulo (Protocol no. 031/15) and registered at ClinicalTrials.gov (Identifier NCT02306148).

Participants

Data from 78 recreational long-distance runners were selected based on the availability for this analysis, since this study is a secondary analysis from a larger randomized clinical trial. All runners consented to participate after receiving information on all details of the study. Participants were between 18 and 55 years of age, ran between 20 and 100 km per week, and had no lower limb injury or pain in the 3 months prior to assessment. Participants were excluded if they were under any physical therapy treatment at baseline, had a history of using minimalist shoes or barefoot running, presented any orthopedic or neurologic impairment or major vascular complication, had previous lower-limb surgery, or had diabetes mellitus.

Running experience was classified by a fuzzy decision-support system developed by Roveri et al. (2017), that is composed by two Mamdani subsystems based on expert running coaches' knowledge. The first subsystem uses the training frequency and volume as inputs, which are transformed into linguistic variables: (i) too low, (ii) low, (iii) medium, or (iv) high for each one of the inputs. These linguistic variables are combined to generate a score (0–10) of quality of practice. In the second subsystem, the quality of practice serves as input, along with the number of competitions and practice time, also transformed in linguistic variables: (i) very bad, (ii) bad, (iii) medium, (iv) good, or (v) very good for quality of practice; (i) few, (ii) medium or (iii) many for number of competitions; (i) very short, (ii) short, (iii) moderate, or (iv) long for practice time. The second subsystem generates, then, the final score of running experience. The running experience score (x) for was used to classify the subjects as less experienced ($x < 5.0$), moderately experienced ($5.0 \leq x \leq 7.0$), or experienced ($x > 7.0$). The characteristics of the subjects according to the experience level and anthropometry are shown in **Table 1**.

Data Collection

Foot biomechanics were assessed during barefoot running at a self-selected speed on an instrumented treadmill, which was leveled with the ground and embedded with two force plates in tandem position (AMTI Force-Sensing treadmill AMTI, Watertown, EUA; force plates at 1000 Hz). Foot kinematics were acquired using eight infrared cameras (Vicon® VERO, Vicon Motion System, Ltd., Oxford Metrics, United Kingdom; at 200 Hz) and 16 retroreflective markers (10 mm in diameter) were placed on the dominant foot according to the Rizzoli Orthopedics Institute Foot Model (Leardini et al., 2007; Portinaro et al., 2014). Subjects underwent a warmup period for habituation to the treadmill and laboratory environment, after which, kinematic data was recorded for 30 s in order to acquire at least 10 step cycles of the assessed limb. There was no statistically significant difference for running velocity across groups (**Table 1**).

Data Processing

The origin of the laboratory coordinate system was defined as one corner of the force plate and all segments were modeled as rigid bodies with the local coordinate system coinciding with the anatomical axes. All joints were considered to have a spherical shape (three rotational degrees of freedom), with rotations of each segment reported relative to the neutral positions defined during the initial static standing trial. All joint rotations were calculated based on the International Society of Biomechanics recommendations (Wu et al., 2002).

Kinematic and GRF data were analyzed and processed using a zero-lag, fourth-order Butterworth low-pass filter with cutoff frequencies of 15 and 100 Hz, respectively, based on residual analysis (Winter and Patla, 1997).

We extracted the eight kinematic time series from the following joints: ankle in all three movement planes; medial longitudinal arch in sagittal plane (Caravaggi et al., 2019); 1st metatarsophalangeal joint in sagittal plane (Met-Hal); the angle between the calcaneus and metatarsal bones (Cal-Met) in all planes.

Calculation of joint kinematics and kinetics were performed using Visual3D software (C-motion, Kingston, ON, Canada). A bottom–up inverse dynamics method was used to calculate the net ankle moment and power in the sagittal plane, with the human body modeled as 2 linked segments (foot and shank) and the inertial properties were based on Dempster's standard regression equations. Net ankle moment and power were calculated for the stance phase.

All nine analyzed GRF variables were normalized by each runner's body weight. From the vertical component, it was extracted the first and second peaks, calculated the loading rate (the force rate between 20 and 80% of the magnitude between the foot contact and the first peak), and the impulses from the beginning of the stance phase to the second peak, and from the second peak to the end of stance phase (**Figure 1**). From the anteroposterior component, it was extracted the negative and positive peaks and calculated the impulses from the decelerating and accelerating phases (**Figure 1**).

Machine Learning Analysis
Feature Extraction

The stance phase of all kinematic (eight variables) and kinetic (ankle sagittal moment and power) time series were determined by means of the vertical GRF using a 10 N threshold and normalized in time to 101 points. Then, the data waveforms (10 waveforms) were averaged across 10 consecutive stance phases of the dominant limb (101 data points per axis direction for each joint kinematics and kinetic waveform), combined into a 78 × 1010 matrix (78 runners X 1010 waveform data point), and standardized to a mean of 0 and a standard deviation of 1 (Kettaneh et al., 2005). Given the large number of dependent variables and potential for redundancy of data, this data matrix underwent a principal component analysis (PCA), resulting in 77 principal components (PCs). PCA is an orthogonal transformation technique used to convert a set of variables into a set of linearly uncorrelated variables by determining new bases (PCs) that maximize the variability in the original data set (Abdi and Williams, 2010). These PC scores were combined with nine standardized discrete GRF variables, resulting in a total set of 86 potential predictor variables.

TABLE 1 | Mean and standard deviation of participants' characteristics from the studied groups.

	Less Experienced (n = 24)	Moderately experienced (n = 23)	Experienced (n = 31)	P
Age (years)	40.1 ± 5.3	40.6 ± 7.1	41.8 ± 7.0	0.614[†]
Height (m)	1.66 ± 0.09	1.71 ± 0.09	1.69 ± 0.09	0.097[†]
Body mass (kg)	71.1 ± 15.4	74.7 ± 10.4	67.1 ± 11.6	0.136[†]
Body mass index (kg/cm^2)	25.6 ± 3.5[a]	25.2 ± 3.1	23.3 ± 2.5[a]	0.017[†*]
Sex (% women)	41.7	60.9	48.4	0.409[‡]
Training frequency (times/week)	3.1 ± 0.6[bc]	3.9 ± 1.0[b]	4.2 ± 1.6[c]	<0.001[†*¥]
Training volume (km/week)	20.0 ± 5.6[d]	29.1 ± 10.3[e]	54.1 ± 38.0[de]	<0.001[†*¥]
Quality of practice (fuzzy system score)	3.4 ± 0.9	4.6 ± 1.2	6.7 ± 1.6	<0.001[†*¥]
Years of practice (years)	2.8 ± 2.9[f]	7.7 ± 11.0	9.0 ± 5.8[f]	0.006[†*]
Participation in races	18.8 ± 26.0[g]	31.1 ± 29.4	53.1 ± 56.9[g]	0.012[†*]
Running experience level (fuzzy system score)	3.2 ± 0.8	5.9 ± 0.7	7.9 ± 0.6	<0.001[†*¥]
Running velocity at data collection (km/h)	9.4 ± 1.4	9.8 ± 1.5	9.7 ± 1.0	0.639[†]

[†]*ANOVA followed by Bonferroni post hoc tests.* [‡]*Chi-square test.* *Statistically significant differences.* [¥]*Statistically significant differences between all studied groups.* [a,b,c,d,e,f,g]*Statistically significant differences between the groups.*

Classification Procedures

The potential predictors were used as inputs for creating SVM models to classify the runners into each experience level group, using a one-vs.-all approach. The SVM approach was chosen because of its capability of overcoming the problem of high dimensionality with high discriminative power for group classification, even in cases with small sample sizes (Verplancke et al., 2008; Son et al., 2010). The SVM algorithm (Noble, 2006) defines an optimal separating hyperplane, creating a maximum-margin of separation between binary classes in a dataset. For that, the SVM projects the input feature's data into a higher dimensional space using kernel functions, and then, based on the data points located closest to the separating hyperplane (support vectors), constructs a linear hyperplane in this transformed space, which can be projected back to the original data space. In order to deal with possible misclassifications (datapoints in the wrong side of the separating hyperplane), SVM uses the soft margin concept, which allows these errors without affecting the final result. The trade-off between margin width and misclassification rate is defined by the C-parameter, wherein different values for C (0.1, 1, 10, 100, 1000) were used in the evaluation to test the dependence of the approach on the C-parameter. A linear kernel function was applied to the SVM algorithm, since it is less prone to overfitting to the dataset, and the current study has a limited sample size.

A sequential forward selection algorithm was applied for the identification of relevant features, in which a subset of potential predictor variables was defined by sequentially adding one new feature at a time to the SVM model, and the feature subset that rendered the best classification performance was selected. In order to assess the generalization performance of the classifier in identifying the label of unknown data and to avoid data overfitting, a 10-fold cross-validation was performed (Fukuchi et al., 2011), wherein the runners were randomly divided into 10 subsets, stratified proportionally by the experience level class. The SVM algorithm was trained by removing one subset at a time and the resulting model was applied to the holdout subset to determine the cross-validation performance. The evaluation function selected the feature that provided the highest Matthew's correlation coefficient (MCC) as the first criteria, because the group distribution becomes unequal in a one-vs.-all approach (Chicco and Jurman, 2020). When it was not possible to calculate the MCC, because of absence of predicted cases in one class, the feature with highest Cohen's d effect size and highest performance accuracy was selected. This process was repeated for all input variables in a greedy search approach, applying each C-parameter value, and the model with the highest MCC was chosen as the final classification model.

Classifier Performance and Interpretation

In order to assess the performance of the SVM models, the cross-validated classification results were used to calculate the MCC, accuracy, precision, recall, and F1-score. A critical binomial test indicated the minimum significant accuracy level for each model, considering a distribution probability equal to the ratio of cases in each experience level class and a confidence level of 0.95.

The squared coefficients of correlation between the PC scores and the joint kinematic and kinetic waveforms was calculated as a measure of proportion of variance of the data that was explained by the selected PCs (Abdi and Williams, 2010), and the relative proportion in each joint and axis was used to help with the interpretation of the selected PC features. The data waveforms were reconstructed based on the selected PCs for each model and the Cohen's d effect size between each group and the remaining runners was calculated. The regions of the waveforms with at least a medium effect size in the one-vs.-all comparison were also considered for interpretation of the distinguishing profile for each experience level class. All data analyses and variable calculations were performed using a custom-written MATLAB script (MathWorks, Natick, MA, United States).

For a better understanding of the discrete GRF variables that were included as possible inputs to the model they were compared between the experience groups using ANOVAs followed by Bonferroni *post hoc* tests ($P < 0.05$). The significance level was set at 5%. All the univariate analyses were performed with SPSS Statistics 23.0 (IBM, Armonk, NY, United States).
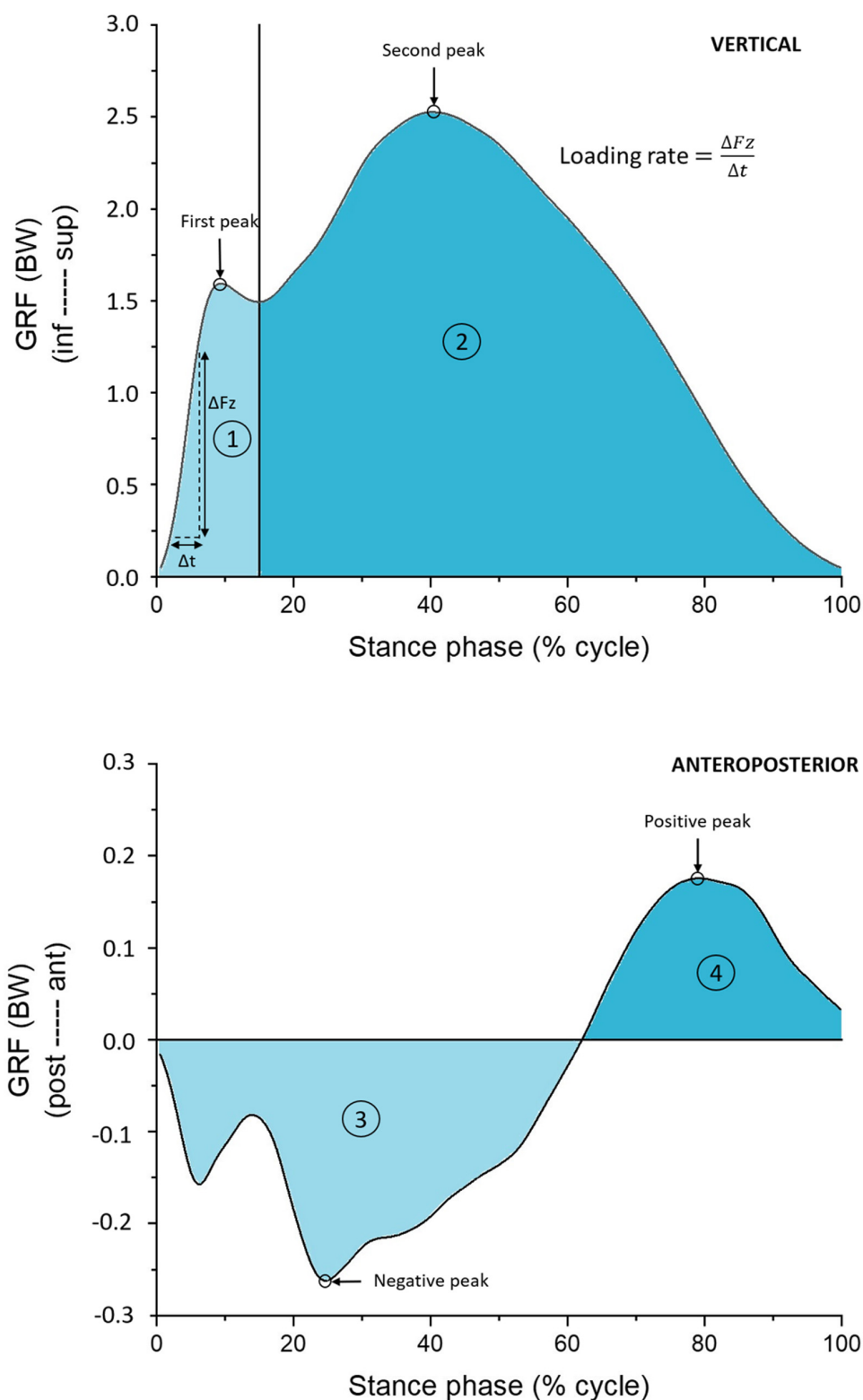
FIGURE 1 | Vertical and anteroposterior (AP) ground reaction forces (GRF) showing the extracted variables: force peaks for both vertical and AP forces, and loading rate for the vertical component. Loading rate was determined as the slope of the line between 20 and 80% of the first vertical peak. The colored areas correspond to the calculated impulses: 1 – from the beginning of the stance phase to the valley after the first vertical peak (light blue area, upper graph); 2 – from the valley after the first vertical peak to the end of the stance phase (blue area, upper graph); 3 – decelerating phase (light blue area, bottom graph); 4 – accelerating phase (blue area, bottom graph).

RESULTS

Performance of SVM Classification Models

All classification performance measures are presented in **Table 2**. The C-parameter values influenced the SVM performance, and the best model for classifying the less experienced group used C = 100, while the best classifiers for the other groups used C = 1. The SVM models for the classification of all experience groups presented a cross-validated accuracy that surpassed the minimum significant accuracy indicated by the critical binomial test. The best classification model for the less experienced group obtained the highest MCC value and the best balance between precision and recall rates, rendering an F1-score of 0.80. The best model for the moderately experienced presented a perfect precision rate, wherein all the runners identified as a member of this group were correctly classified, but almost half of the moderately experienced runners were misclassified, resulting in low recall rate (56.5%). The model for the experienced group presented a lower accuracy rate, but precision and recall rates reached 82.1 and 76.7%, respectively. Overall, the best classification models achieved an MCC score of strong to very strong relationships.

Discriminating Running Patterns

The SVM model for the less experienced group selected 81 variables as the input features, from which 72 of them were PCs from the foot-ankle kinematic and kinetic (ankle sagittal moment and power) time series, with a total variance explained of 99.9% The other nine features were GRF variables, which included the loading rate, 1st and 2nd vertical peaks, both vertical impulses, the anteroposterior negative and positive peaks, and impulses from decelerating and accelerating phases. However, between-group comparisons of the GRF variables did not find any significant differences across experience levels (**Table 3**). The reconstructed waveforms indicated that the kinematic features considered important for the identification of the less experienced runners were mainly related to greater Cal-Met plantar flexion, and greater Met-Hal dorsiflexion (**Figures 2**, **3**). Both of these joint movements are related to PC2, which was the first feature included in the forward feature selection process and represents 15.5% of variance explained of the total foot-ankle data, wherein the relative proportion of representation is highly loaded on Cal-Met (44.1% of PC2) and Met-Hal (28.1%) sagittal planes.

The classification model for the moderately experienced group selected eight PCs that were responsible for only 0.4% of variance explained, and two GRF variables, which were the loading rate and anteroposterior negative peak. The foot-ankle mechanics features distinguishing the moderately experienced runners were not as evident, since they were only PCs representing very low variance explained and did not present a visually identifiable pattern.

For the experienced group the best model selected six foot-ankle kinematic/kinetic waveform PCs representing 21.7% of variance explained, along with the 1st vertical GRF peak and the anteroposterior impulse from the acceleration phase. The experienced runners presented the opposite distinguishing features from the less experienced group, which were mainly related to smaller CalMet plantar flexion and smaller Met-Hal dorsiflexion (**Figures 2**, **3**). These differences were also related to PC2, which was the only selected with a relatively high variance explained.

DISCUSSION

The main purpose of the study was to determine if running experience level could be classified based on foot-ankle kinematic and kinetic patterns, and GRF variables. The results of this study showed that this classification is possible since the SVM models successfully separated all experience groups, with the less experienced and experienced runners presenting discriminating features with opposing motor patterns in the metatarsophalangeal and calcaneus-metatarsal joints, while the moderately experienced group did not present an explicitly visible pattern, although they were classified with significant accuracy.

One of the distinguishing kinematic features for the less experienced runners was a larger toe dorsiflexion angle throughout the whole stance phase. It is possible that the lack of experience is related to a greater use of hallux and toe extensor muscles in combination with tibialis anterior muscle as ankle dorsiflexors, in order to guarantee toe clearance and dorsiflexion throughout the running stance phase, especially during the weight acceptance phase, which could result in the observed greater toe dorsiflexion. Interestingly, the experienced group presented smaller toe dorsiflexion angles than the other runners, suggesting that there is a change in motor strategy with skill acquisition, possibly reducing muscle activation that are unnecessary and causing less energy consumption. There is evidence that runners with more years of running experience show different lower-limb coordination patterns, more specifically, in the variability of coordination, as measured by non-linear analysis (Agresta et al., 2019; Hafer et al., 2019; Mo and Chow, 2019), showing that deliberate practice seems to cause motor strategy adaptations that modifies biomechanical patterns. Unfortunately, to our knowledge there are no studies comparing muscular activation patterns in runners with different experience levels showing evidences of that. A study in cyclists showed that novice cyclists present longer periods of primary activity in leg muscles and more extensive coactivation between muscles, as opposed to trained cyclists, who display shorter bursts with consistent timing of muscle activity (Chapman et al., 2008). Although cycling is a different motor task, this finding suggests that deliberate practice of a motor task seems to change muscle coordination.

The metatarsophalangeal joints function as a dissipater of large amounts of energy during running and sprinting, particularly when a passive dorsiflexion occurs at the foot contact transition from the metatarsal heads onto the toes, but fail to generate any energy at push off by remaining in that position (Stefanyshyn and Nigg, 1997). Since the less experienced runners displayed larger angles of hallux dorsiflexion

	Less experienced vs. all	Moderately experienced vs. all	Experienced vs. all
Minimum significant accuracy	76.9%	78.2%	70.5%
Cross-validation performance			
Accuracy	88.5%	87.2%	84.6%
Recall	72%	56.5%	76.7%
Precision	90%	100%	82.1%
F1-score	0.80	0.72	0.79
MCC[a]	0.73	0.69	0.67

[a]*Matthew's correlation coefficient.*

TABLE 3 | Mean and standard deviation of ground reaction force variables extracted from the stance phase of running and results from the between-group comparisons and correlation analysis.

Discrete variable	Less experienced (*n* = 24)	Moderately experienced (*n* = 23)	Experienced (*n* = 31)	ANOVA *F*	*P*
Vertical component					
Loading rate (N/s)	78.86 ± 43.45	61.94 ± 25.95	69.00 ± 36.79	1.295	0.28
First peak (BW)	1.19 ± 0.40	1.09 ± 0.37	1.08 ± 0.39	0.598	0.55
Second peak (BW)	2.15 ± 0.29	2.26 ± 0.30	2.21 ± 0.28	0.990	0.38
Impulse 1 (N.s)	0.033 ± 0.020	0.027 ± 0.016	0.031 ± 0.019	0.590	0.56
Impulse 2 (N.s)	0.318 ± 0.028	0.327 ± 0.024	0.311 ± 0.030	1.898	0.16
Anteroposterior component					
Positive peak (BW)	0.226 ± 0.050	0.238 ± 0.045	0.239 ± 0.053	0.502	0.61
Negative peak (BW)	−0.230 ± 0.080	−0.226 ± 0.614	−0.236 ± 0.053	0.160	0.85
Impulse 3 (N.s)	−0.017 ± 0.006	−0.018 ± 0.005	−0.016 ± 0.005	1.027	0.36
Impulse 4 (N.s)	0.016 ± 0.004	0.016 ± 0.004	0.017 ± 0.005	0.254	0.78

BW, body weight.

during this transition between midstance to push off, they would be expected to have greater soft tissue tension, such as in the toe flexor muscles and tendons, and plantar fascia (Bruening et al., 2018). However, since there was little or no metatarsophalangeal plantar flexion, there was more energy dissipated by the passive structures, leading to less efficient propulsion in the following phase.

There was also an opposing behavior between the less experienced and experienced groups regarding the calcaneus-metatarsal, with the former displaying greater plantar flexion. This higher plantar flexion in less experienced runners could be associated to the greater toe dorsiflexion angles, because of the windlass mechanism, in which the dorsiflexion moments at the metatarsophalangeal joints cause tension to the plantar aponeurosis, pulling the calcaneus toward the metatarsal heads (Holowka and Lieberman, 2018).

Evidence shows that the passive structures of the foot arch have an important role in the metabolic energy-saving of the foot by the reduction of the mechanical work that would be needed from muscle activation (Stearne et al., 2016). One of the exclusion criteria for the current study was the use of minimalist shoes, which are known to be associated with higher intrinsic foot muscle volume after a running training regime (Miller et al., 2014; Chen et al., 2016). Since the subjects of this study all ran with traditional shoes, the function of the foot intrinsic muscles is probably reduced due to the support given to the medial longitudinal arch and the stiffness

of the midsole in the traditional running footwear, possibly causing a higher reliance on the passive properties of the plantar tissues. Therefore, it is expected that there would be no disparity in the function of these muscles between groups. The fact that the experienced runners presented a less plantar flexed calcaneus-metatarsal joint potentially indicate the presence of a higher tension in the plantar structures, such as the plantar fascia, and intrinsic and extrinsic foot muscles, since this higher angle would correspond to a greater distance between the rearfoot and the forefoot. Although the arch compression/recoil property is not changed due to experience level, since the calcaneus-metatarsal range of motion is similar between all experience groups (**Figure 2**), the foot plantar structures would be working in a position with higher strain in the experienced runners. If experienced runners rely more on the passive structures, less muscle activation is needed, and less energy is spent (Stearne et al., 2016), showing that experienced runners might have a more efficient foot-ankle biomechanical pattern. The question remains whether this pattern could be related to protective factors against injuries in runners with more experience.

Moderately experienced runners did not show a clearly distinct biomechanical pattern since the SVM model selected only higher order PCs with low explained variance (0.4%) as discriminating factors, indicating that differences were very subtle and complicated to be visualized. Still, these complex patterns were able to successfully distinguish this particular

FIGURE 2 | Angle between calcaneus and metatarsal bones in the sagittal plane during stance phase of running for all three running experience level groups. The foot schematics in the left represent the calculated angles.
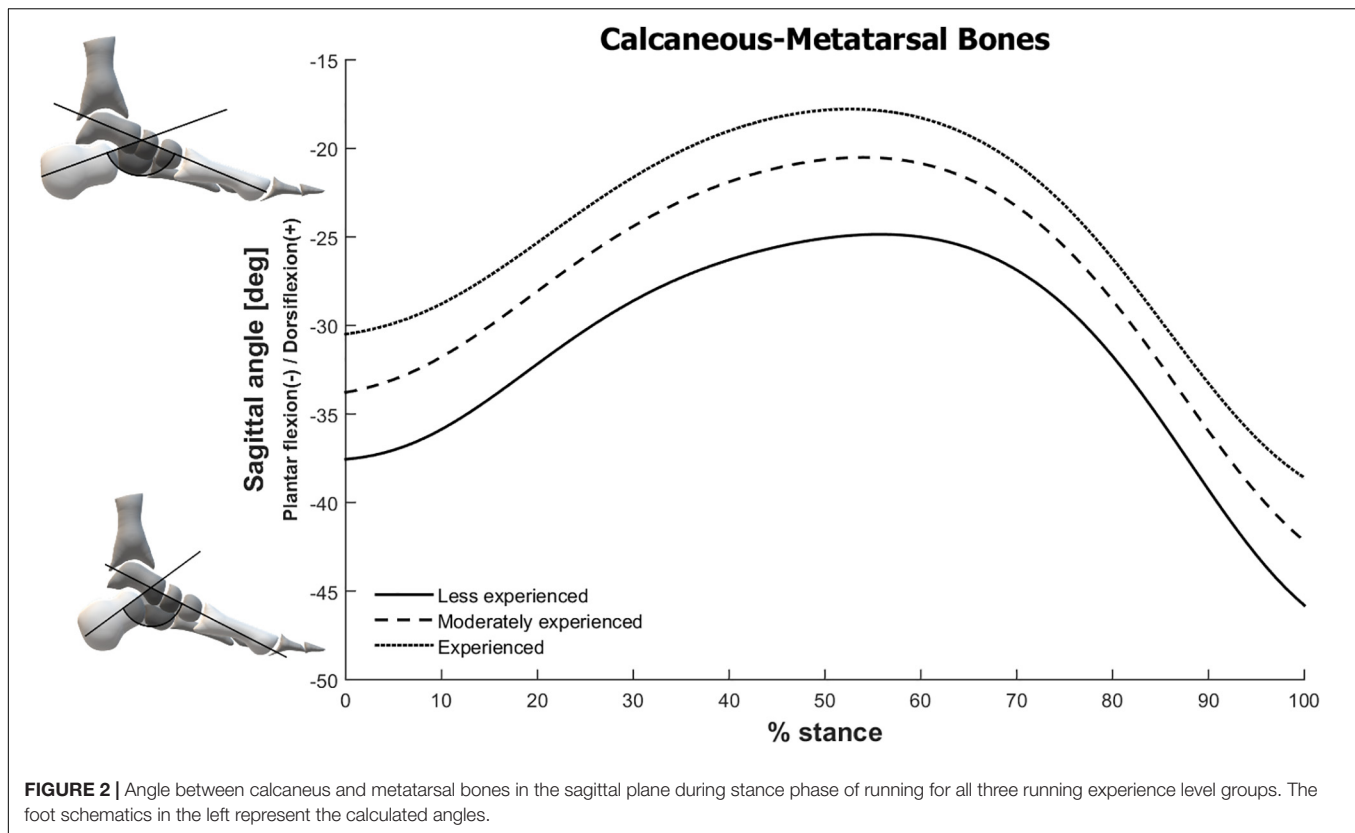
group from the other runners. It is possible that this is a transitioning group that is composed of a more heterogeneous sample, which would hinder the identification of a specific movement pattern. This model presented a perfect precision rate, i.e., all the subjects identified as moderately experienced runners were correctly classified. However, it had a low recall rate (56.5%), meaning that it fails to detect almost half of the runners from the moderately experienced group. Therefore, there is a specific mechanical pattern for this group, which is probably responsible for the high precision rate of the model, but the high rate of false negatives supports the assumption of a heterogeneous and intermediate group, with a great portion of runners possibly behaving similarly to either experienced or less experienced runners. It is possible that a further stratification of the experience levels could improve the discrimination among these subjects, but that would not be possible with the current sample size, which is a limitation of this analysis.

Although there were no significant differences between experience levels for the GRF variables, they were still crucial for the performance of the classification models, since they were selected as discriminating features by the SVM models. This shows that the combination of different biomechanical features is a better representation of the motor behavior and necessary for an improved identification of the mechanical patterns. Unfortunately, it was not possible to identify how these GRF factors are specifically related to the distinguishing mechanical pattern within each group.

The identified mechanical patterns should be considered with caution, since the participants were asked to run barefoot because it was needed for the multi-segment foot model implementation (Leardini et al., 2007; Portinaro et al., 2014), which could have altered the runners' habitual movement. Nevertheless, they still reflect distinct motor strategies across experience levels, highlighting the importance of foot-ankle mechanics to discriminate the experimental groups. On a similar matter, we did not define a fixed running velocity for the subjects, which could lead to mechanical differences that are due to speed effects. However, there were no significant differences in running velocity across experience levels, thus we do not expect that such effects influenced our results.

A PCA was applied to the foot-ankle kinematic and kinetic data for dimensionality reduction and feature extraction purposes, and this strategy allows the identification of patterns across the foot joints. However, all the PCs were included as possible features in the SVM, even though the movement patterns are only visible in the low-order PCs and the high-order PCs could be including noise and signal artifacts. This approach was chosen because there are studies that were not able to identify differences between experience levels, indicating that the changes due to running experience could be more complex and subtle, and such differences have been shown to be identified only by the high-order PCs (Phinyomark et al., 2015). Although it is possible that a portion of the PCs are representing noise in the data, the SVM models still successfully classified the different experience levels by including these high-order PCs, and presented an

FIGURE 3 | First metatarsophalangeal joint sagittal plane angle during stance phase of running for all three running experience level groups. The foot schematics in the left represent the calculated angles.
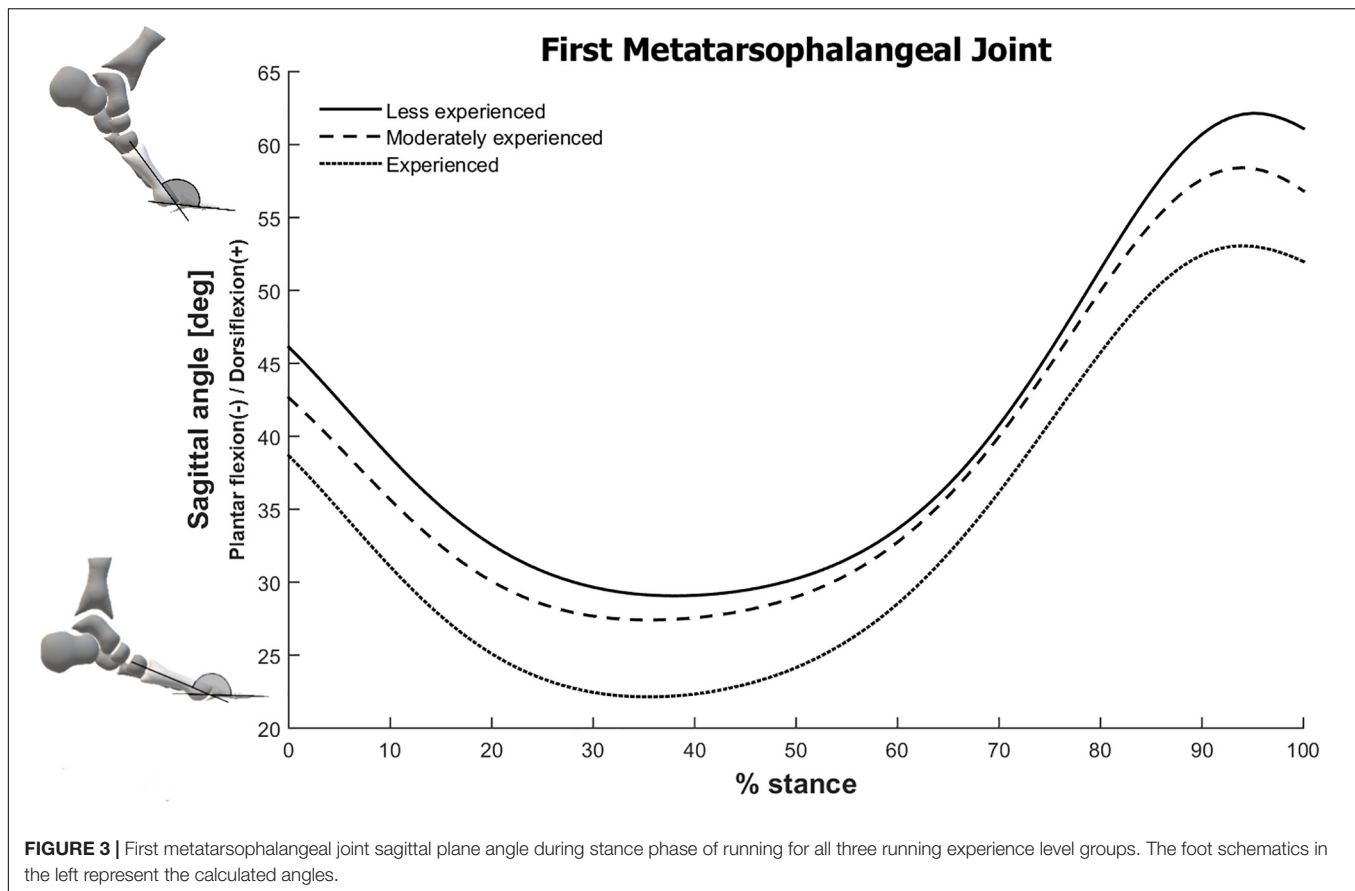
accuracy rate that surpassed the minimal threshold determined by a critical binomial test.

Another limitation of this study is its relatively small sample size, since the training data set for a machine learning classifier approach should increase exponentially in size for each added input dimension (Altman and Krzywinski, 2018). For that reason, the SVM was chosen for the analysis, because it considers the data points in the margins of each class to determine the separating hyperplane, thus it is not influenced by the distribution of the data points and can effectively separate binary classes even with limited sample size (Noble, 2006). In addition, the use of SVM has shown very high classification performance in clinical settings (Golub et al., 1999; Son et al., 2010) and with biomechanical data (Lai et al., 2009; Fukuchi et al., 2011).

Another problem that a reduced sample size can encounter is data overfitting, which causes the classifying function to be too specific to the training dataset and not generalizable. Although it would be ideal to have a test dataset to ensure the generalizability of the analysis, it was not possible to use this strategy because of the limited sample size coming from the major randomized clinical trial. Thus, in order to avoid overfitting, a 10-fold cross-validation was applied in the feature selection process and the performance measurement, increasing the robustness of the results, and a linear kernel was applied to the algorithm, again to prevent overfitting by not adapting the hyperplanes to irregular margins. Still, the current results

should be considered with caution since extrapolation for general population may not be suitable. Furthermore, since SVM is intended for binary classifications, it was necessary to use a one-vs.-all approach, which causes the group sizes to be unequal and interferes in the classifier performance measures. Nevertheless, this issue was dealt by applying the MCC as the main performance measure because it is more reliable and informative when evaluating binary classifications, especially on imbalanced datasets (Chicco and Jurman, 2020).

This is the first study to apply a machine learning approach to investigate how running experience affects foot-ankle biomechanics. The results indicate that using foot-ankle kinematic and kinetic waveforms associated with GRF variables as inputs in an SVM classifier can successfully separate and classify runners with different levels of experience. The main identifiable features that are important for the discrimination were the toe dorsiflexion and calcaneus-metatarsal plantar flexion angles. The less experienced group presented greater metatarsophalangeal dorsiflexion throughout the whole stance phase, which could cause higher energy dissipation and a less efficient propulsion. The more experienced group displayed smaller calcaneus-metatarsal plantar flexion that might be related to a more efficient biomechanical pattern regarding energy expenditure. As for the moderately experienced runners, although they were successfully classified, they did not present a visually identifiable running pattern, and seem

to be an intermediate group between the less and more experienced runners. The current findings can potentially guide the development of training programs and rehabilitation protocols aimed at foot-ankle mechanics for different levels of experience in running. Furthermore, future studies could investigate what are the implications of those different biomechanical patterns according to the experience level in running performance and injuries.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research and Ethics Committee of the School of Medicine of the University of São Paulo (Protocol no. 031/15). The participants provided their written informed consent to

participate in this larger randomized clinical trial registered at ClinicalTrials.gov (Identifier NCT02306148).

AUTHOR CONTRIBUTIONS

All authors have made substantial contributions to the manuscript. ES, RW, and IS were responsible for the conception and design of the study. AM was responsible for data acquisition and data processing. ES and RW were responsible for data analysis and interpretation, and drafting the article. IS and AM revised the manuscript critically. All authors read, provided feedback, and approved the submitted version.

FUNDING

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdiscip. Rev. Comp. Stat.* 2, 433–459. doi: 10.1002/wics.101

Agresta, C. E., Goulet, G. C., Peacock, J., Housner, J., Zernicke, R. F., and Deneweth, J. (2019). Years of running experience influences stride-to-stride fluctuations and adaptive response during step frequency perturbations in healthy distance runners. *Gait Posture* 70, 376–382. doi: 10.1016/j.gaitpost.2019.02.034

Agresta, C. E., Peacock, J., Housner, J., Zernicke, R. F., and Zendler, J. D. (2018). Experience does not influence injury-related joint kinematics and kinetics in distance runners. *Gait Posture* 61 13–18. doi: 10.1016/j.gaitpost.2017.12.020

Altman, A. R., and Davis, I. S. (2016). Prospective comparison of running injuries between shod and barefoot runners. *Br. J. Sports Med.* 50, 476–480. doi: 10.1136/bjsports-2014-094482

Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. *Nat. Methods* 15, 399–400.

Bruening, D. A., Pohl, M. B., Takahashi, K. Z., and Barrios, J. A. (2018). Midtarsal locking, the windlass mechanism, and running strike pattern: a kinematic and kinetic assessment. *J. Biomech.* 73, 185–191. doi: 10.1016/j.jbiomech.2018.04.010

Caravaggi, P., Matias, A. B., Taddei, U. T., Ortolani, M., Leardini, A., and Sacco, I. C. N. (2019). Reliability of medial-longitudinal-arch measures for skin-markers based kinematic analysis. *J. Biomech.* 88, 180–185. doi: 10.1016/j.jbiomech.2019.03.017

Chapman, A. R., Vicenzino, B., Blanch, P., and Hodges, P. W. (2008). Patterns of leg muscle recruitment vary between novice and highly trained cyclists. *J. Electromyogr. Kinesiol.* 18, 359–371. doi: 10.1016/j.jelekin.2005.12.007

Chen, T. L. W., Sze, L. K. Y., Davis, I. S., and Cheung, R. T. H. (2016). Effects of training in minimalist shoes on the intrinsic and extrinsic foot muscle volume. *Clin. Biomech.* 36, 8–13. doi: 10.1016/j.clinbiomech.2016.05.010

Chicco, D., and Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. doi: 10.1186/s12864-019-6413-7

Clermont, C. A., Osis, S. T., Phinyomark, A., and Ferber, R. (2017). Kinematic gait patterns in competitive and recreational runners. *J. Appl. Biomech.* 33, 268–276. doi: 10.1123/jab.2016-0218

Davis, I. S., and Futrell, E. (2016). Gait retraining: altering the fingerprint of gait. *Phys. Med. Rehabil. Clin. N. Am.* 27, 339–355. doi: 10.1016/j.pmr.2015.09.002

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295x.100.3.363

Fukuchi, R. K., Eskofier, B. M., Duarte, M., and Ferber, R. (2011). Support vector machines for detecting age-related changes in running kinematics. *J. Biomech.* 44, 540–542. doi: 10.1016/J.JBIOMECH.2010.09.031

Garofolini, A., and Taylor, S. (2019). The effect of running on foot muscles and bones: a systematic review. *Hum. Mov. Sci.* 64, 75–88. doi: 10.1016/j.humov.2019.01.006

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537. doi: 10.1126/science.286.5439.531

Hafer, J. F., Peacock, J., Zernicke, R. F., and Agresta, C. E. (2019). Segment coordination variability differs by years of running experience. *Med. Sci. Sports Exerc.* 51, 1438–1443. doi: 10.1249/MSS.0000000000001913

Holowka, N. B., and Lieberman, D. E. (2018). Rethinking the evolution of the human foot: insights from experimental research. *J. Exp. Biol.* 221(Pt 17):jeb174425. doi: 10.1242/jeb.174425

Hulme, A., Nielsen, R. O., Timpka, T., Verhagen, E., and Finch, C. F. (2017). Risk and protective factors for middle- and long-distance running-related injury. *Sports Med.* 47, 869–886. doi: 10.1007/s40279-016-0636-4

Iglesias, G. D., García-González, L., Calvo, T. G., del Barco, B. L., and del Villar, F. (2010). Expertise development in sport: contributions under cognitive psychology perspective. *J. Hum. Sport Exerc.* 5, 462–475. doi: 10.4100/jhse.2010.53.16

Kelly, L. A., Cresswell, A. G., and Farris, D. J. (2018). The energetic behaviour of the human foot across a range of running speeds. *Sci. Rep.* 8:10576. doi: 10.1038/s41598-018-28946-1

Kettaneh, N., Berglund, A., and Wold, S. (2005). PCA and PLS with very large data sets. *Comp. Stat. Data Anal.* 48, 69–85. doi: 10.1016/j.csda.2003.11.027

Kluitenberg, B., van Middelkoop, M., Diercks, R., and van der Worp, H. (2015). What are the differences in injury proportions between different populations of runners? A systematic review and meta-analysis. *Sports Med.* 45, 1143–1161. doi: 10.1007/s40279-015-0331-x

Lai, D. T. H., Levinger, P., Begg, R., Gilleard, W. L., and Palaniswami, M. (2009). Automatic recognition of gait patterns exhibiting patellofemoral pain syndrome

using a support vector machine approach. *IEEE Trans. Inform. Technol. Biomed.* 13, 810–817. doi: 10.1109/TITB.2009.2022927

Leardini, A., Benedetti, M. G., Berti, L., Bettinelli, D., Nativo, R., and Giannini, S. (2007). Rear-foot, mid-foot and fore-foot motion during the stance phase of gait. *Gait Posture* 25, 453–462. doi: 10.1016/J.GAITPOST.2006.05.017

Lee, D.-C., Brellenthin, A. G., Thompson, P. D., Sui, X., Lee, I.-M., and Lavie, C. J. (2017). Running as a key lifestyle medicine for longevity. *Prog. Cardiovas. Dis.* 60, 45–55. doi: 10.1016/j.pcad.2017.03.005

Lieberman, D. E. (2012). What we can learn about running from barefoot running: an evolutionary medical perspective. *Exerc. Sport Sci. Rev.* 40, 63–72. doi: 10.1097/JES.0b013e31824ab210

Linton, L., and Valentin, S. (2018). Running with injury: a study of UK novice and recreational runners and factors associated with running related injury. *J. Sci. Med. Sport* 21, 1221–1225. doi: 10.1016/j.jsams.2018.05.021

Macera, C. A. (1992). Lower extremity injuries in runners. advances in prediction. *Sports Med.* 13, 50–57.

Miller, E. E., Whitcome, K. K., Lieberman, D. E., Norton, H. L., and Dyer, R. E. (2014). The effect of minimal shoes on arch structure and intrinsic foot muscle strength. *J. Sport Health Sci.* 3, 74–85. doi: 10.1016/j.jshs.2014.03.011

Mo, S., and Chow, D. H. K. (2019). Differences in lower-limb coordination and coordination variability between novice and experienced runners during a prolonged treadmill run at anaerobic Threshold speed. *J. Sports Sci.* 37, 1021–1028. doi: 10.1080/02640414.2018.1539294

Nielsen, R. O., Buist, I., Sørensen, H., and Lind, M. (2012). Training errors and running related injuries: a systematic review. *Int. J. Sports Phys. Ther.* 7, 58–75.

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Phinyomark, A., Hettinga, B. A., Osis, S., and Ferber, R. (2015). Do intermediate- and higher-order principal components contain useful information to detect subtle changes in lower extremity biomechanics during running? *Hum. Mov. Sci.* 44, 91–101. doi: 10.1016/j.humov.2015.08.018

Portinaro, N., Leardini, A., Panou, A., Monzani, V., and Caravaggi, P. (2014). Modifying the rizzoli foot model to improve the diagnosis of pes-planus: application to kinematics of feet in teenagers. *J. Foot Ankle Res.* 7:754. doi: 10.1186/s13047-014-0057-2

Rodgers, M. M. (1988). Dynamic biomechanics of the normal foot and ankle during walking and running. *Phys. Ther.* 68, 1822–1830. doi: 10.1093/ptj/68.12.1822

Roveri, I., de Jesus Manoel, E., Onodera, A. N., Ortega, N. R. S., Tessutti, V. D., Vilela, E., et al. (2017). Assessing experience in the deliberate practice of running using a fuzzy decision-support system. *PLoS One* 12:e0183389. doi: 10.1371/journal.pone.0183389

Schmitz, A., Russo, K., Edwards, L., and Noehren, B. (2014). Do novice runners have weak hips and bad running form? *Gait Posture* 40, 82–86. doi: 10.1016/j.gaitpost.2014.02.014

Son, Y. J., Kim, H. G., Kim, E. H., Choi, S., and Lee, S. K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Inform. Res.* 16, 253–259. doi: 10.4258/hir.2010.16.4.253

Stearne, S. M., McDonald, K. A., Alderson, J. A., North, I., Oxnard, C. E., and Rubenson, J. (2016). The foot's arch and the energetics of human locomotion. *Sci. Rep.* 6:19403. doi: 10.1038/srep19403

Stefanyshyn, D. J., and Nigg, B. M. (1997). Mechanical energy contribution of the metatarsophalangeal joint to running and sprinting. *J. Biomech.* 30, 1081–1085. doi: 10.1016/S0021-9290(97)00081-X

van der Worp, M. P., ten Haaf, D. S. M., van Cingel, R., de Wijer, A., Nijhuis-van der Sanden, M. W. G., and Staal, J. B. (2015). Injuries in runners; a systematic review on risk factors and sex differences. edited by Amir A. Zadpoor. *PLoS One* 10:e0114937. doi: 10.1371/journal.pone.0114937

van Gent, R. N., Siem, D., van Middelkoop, M., Van Os, A. G., Bierma-Zeinstra, S. M. A., Koes, B. W., et al. (2007). Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *Br. J. Sports Med.* 41, 469–480. doi: 10.1136/bjsm.2006.033548

Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F., et al. (2008). Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically Ill patients with haematological malignancies. *BMC Med. Inform. Decision Mak.* 8:56. doi: 10.1186/1472-6947-856

Videbæk, S., Bueno, A. M., Nielsen, R. O., and Rasmussen, S. (2015). Incidence of running-related injuries per 1000 H of running in different types of runners: a systematic review and meta-analysis. *Sports Med.* 45, 1017–1026. doi: 10.1007/s40279-015-0333-8

Winter, D. A., and Patla, A. E. (1997). *Signal Processing and Linear Systems for the Movement Sciences.* Waterloo: Waterloo Biomechanics.

Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., et al. (2002). ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion–part I: ankle, hip, and spine. International society of biomechanics. *J. Biomech.* 35, 543–548. doi: 10.1016/S0021-9290(01)00222-6

Check for
updates

CNN-Based Estimation of Sagittal Plane Walking and Running Biomechanics From Measured and Simulated Inertial Sensor Data

Eva Dorschky[1]*, Marlies Nitschke[1], Christine F. Martindale[1], Antonie J. van den Bogert[2], Anne D. Koelewijn[1] and Bjoern M. Eskofier[1]

[1] *Machine Learning and Data Analytics Lab, Department of Computer Science, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany,* [2] *Mechanical Engineering Department, Cleveland State University, Cleveland, OH, United States*

Machine learning is a promising approach to evaluate human movement based on wearable sensor data. A representative dataset for training data-driven models is crucial to ensure that the model generalizes well to unseen data. However, the acquisition of sufficient data is time-consuming and often infeasible. We present a method to create realistic inertial sensor data with corresponding biomechanical variables by 2D walking and running simulations. We augmented a measured inertial sensor dataset with simulated data for the training of convolutional neural networks to estimate sagittal plane joint angles, joint moments, and ground reaction forces (GRFs) of walking and running. When adding simulated data, the root mean square error (RMSE) of the test set of hip, knee, and ankle joint angles decreased up to 17 %, 27 % and 23 %, the RMSE of knee and ankle joint moments up to 6 % and the RMSE of anterior-posterior and vertical GRF up to 2 and 6 %. Simulation-aided estimation of joint moments and GRFs was limited by inaccuracies of the biomechanical model. Improving the physics-based model and domain adaptation learning may further increase the benefit of simulated data. Future work can exploit biomechanical simulations to connect different data sources in order to create representative datasets of human movement. In conclusion, machine learning can benefit from available domain knowledge on biomechanical simulations to supplement cumbersome data collections.

Keywords: biomechanics, biomechanical simulation and analysis, gait analysis, musculoskeletal simulation, inertial sensors, optimal control, machine learning, convolutional neural networks - CNN

1. INTRODUCTION

Due to technological advances in wearable computing, it is now possible to measure human movement outside the lab, in the natural environment (Seshadri et al., 2019). This facilitates a continuous monitoring of patients and athletes supporting medical diagnosis, performance assessment in sports, prevention of falling or sport-related injuries, tracking of disease progression and evaluating the efficiency of treatment. Extracting useful information from sensor data remains challenging as uncontrolled natural conditions imply variations in sensor placement, in data quality, and a wide range of movement patterns. Typically, only discrete variables are

computed from sensor data, such as speed, stride length, and step frequency (Hannink et al., 2017; Falbriard et al., 2018; Zrenner et al., 2018). However, a comprehensive biomechanical analysis, which involves the evaluation of joint angles, joint moments, muscle forces, and ground reaction forces (GRFs), would be beneficial to gain a deeper understanding of the movement mechanics and underlying causes.

However, low-quality sensor data and sparse measurements make it difficult to achieve a comprehensive analysis that is comparable to laboratory results, where optical motion capture (OMC) systems and force plates are available. Different methods were developed to address the challenge of extracting the kinematic and kinetic parameters of movements from sensor data, commonly inertial sensor data. These methods can be divided into physics-based or data-driven approaches.

Physics-based approaches use kinematic chain models or musculoskeletal models in combination with Kalman filters or global optimization to constrain the solution space (Roetenberg et al., 2009; Koning et al., 2013; Kok et al., 2014; Miezal et al., 2017; Karatsidis et al., 2018; Dorschky et al., 2019). Physical models can act as a filter to the noisy sensor data. Moreover, reconstructing the movement with a musculoskeletal model yields a comprehensive analysis including muscle forces, kinematics, and kinetics. In contrast to data-driven approaches, no lab measurements are necessary to train the model. However, global optimization methods require a relatively high computation time (Kok et al., 2014; Dorschky et al., 2019) and are thus less suitable for real-time applications. In addition, model inaccuracies such as simplified ground contact lead to errors in GRF and joint moment estimations.

Data-driven approaches can directly learn a mapping between sensor data and target biomechanical variables based on lab measurements (Wouda et al., 2018; Komaris et al., 2019; Stetter et al., 2019; Zell and Rosenhahn, 2019). Machine learning algorithms can reveal hidden relationships between sensor data and biomechanical variables, in particular, deep learning is a promising approach to model time series data of human movement (Halilaj et al., 2018). Trained models can be exploited in real-time to provide instantaneous feedback to the patient, athlete, or coach. For example, an early warning system monitoring the internal joint loads during sports could potentially prevent catastrophic non-contact knee injuries (Johnson et al., 2019). Furthermore, low-latency feedback on joint moments could help gait retraining in osteoarthritis patients to reduce the knee adduction moment (Preece et al., 2009). However, training data-based models requires a representative dataset, which is cumbersome to acquire as it typically involves synchronized recordings of inertial sensors and OMC systems. It is often impractical to collect a dataset large enough to train deep neural networks. Variations in movement patterns, different sensor positions, and movement or sensor artifacts can lead to high generalization errors within data-based models (Wouda et al., 2018).

Strategies like data augmentation and transfer learning have been applied to improve robustness and generalization of data-based models. Um et al. (2017) used label-preserving transformations of the sensor data (e.g., rotations, permutations, and time-warping) to augment the training dataset. This improved the robustness of the model with respect to sensor position and noise, but did not account for variations in movement patterns as the target variables remained unchanged. Veiga et al. (2017) and Johnson et al. (2019) utilized pre-trained deep neural networks from the image domain as a feature extractor. The former authors used images showing line curves of sensor signals. However, characteristic features of one dimensional inertial sensor signals likely differ from photographic images extracted from the ImageNet database. Johnson et al. (2019) transformed the data of five accelerometers into two-dimensional images: one dimension representing the sensor locations and the other dimension the normalized time. The acceleration magnitude was quantized to greyscale or RGB colorspace, what probably caused information loss.

To learn from sufficient data and incorporate variations of movement, Johnson et al. (2019) synthesized accelerometer data via double-differentiation of marker trajectories from their OMC archive. Huang et al. (2018) also synthesized inertial sensor data from motion capture datasets using a 3D model of the human body shape and pose (SMPL) together with a virtual sensor model. Mundt et al. (2020a,b) used OMC data from several studies of their lab together with a biomechanical model to create a large simulated dataset, which was used for training feedforward neural networks to estimate joint kinematics and kinetics. One drawback of these approaches is that additional datasets containing OMC data or SMPL poses of the movement of interest were required. Notably, Huang et al. (2018) reported that combining these datasets was non-trivial. Moreover, each recorded motion trajectory led to only one synthetic sensor trajectory. An infinite number of random samples can be generated using statistical modeling. Norgaard et al. (2018) synthesized inertial sensor data from random vectors using a generative adversarial network. Their approach did not include biomechanical constraints to extract physically plausible samples.

Our goal is to use physical knowledge of biomechanics to alleviate the issue of data limitation. We contribute a new method to expand a training dataset via biomechanical simulations created by solving optimal control problems. We simulated musculoskeletal models to follow walking and running trajectories which were randomly sampled from a "small" measured training dataset. In principle, an infinite number of simulations could be obtained with matching inertial sensor data and biomechanical variables. The constraints in the optimal control problem ensured that simulated motions were physically possible and dynamically consistent.

We evaluated if learning on simulated data can decrease generalization errors, how much simulated data is necessary, and what happens in the case of even smaller training datasets. Therefore, we trained convolutional neural networks (CNNs) to map inertial sensor data of walking and running cycles to joint angles, joint moments and GRFs. We compared the performance of the CNNs for training on only measured data with training on measured and simulated data.

FIGURE 1 | We trained CNNs to estimate sagittal lower body kinematics and kinetics from accelerometer and gyroscope data from four inertial sensors which were placed on the lower body. Therefore, we created simulated data based on the measured training dataset (described in section 2.1): we drew random samples from measured joint angles, GRFs, and walking/running speeds (see section 2.2), which were then tracked by musculoskeletal models solving optimal control problems (see section 2.3). Simulated movements yielded biomechanics with matching inertial sensor data using a virtual inertial sensor model.

2. MATERIALS AND METHODS

Figure 1 shows the overview of the proposed methods. We trained CNNs (LeCun et al., 1989) to estimate sagittal lower body kinematics and kinetics from accelerometer and gyroscope data from four inertial sensors which were placed on the lower body. Therefore, we created simulated data based on the measured training dataset (described in section 2.1): we drew random samples from measured joint angles, GRFs, and walking/running speeds (see section 2.2), which were then tracked by musculoskeletal models solving optimal control problems (see section 2.3). Simulated movements yielded biomechanics with matching inertial sensor data using a virtual inertial sensor model. We explain the network architecture of the CNNs in section 2.4 and the evaluation process in section 2.5.

2.1. Measured Data

We used the data recorded by Dorschky et al. (2019), which consisted of data from 10 subjects (denoted by S01-S10) walking and running at six different speeds with 10 trials each. The walking speeds were: 0.9 to 1.0 m s^{-1}, 1.2 to 1.4 m s^{-1}, and 1.8 to 2.0 m s^{-1}. The running speeds were: 3.1 to 3.3 m s^{-1}, 3.9 to 4.1 m s^{-1}, and 4.7 to 4.9 m s^{-1}. The dataset comprises 595 (valid) walking and running cycles in total. It includes data from seven custom-built inertial sensors (Portabiles GmbH, Erlangen,

DE) (Blank et al., 2015) including tri-axial accelerometers (± 16 g) and gyroscopes (± 2.000 deg/s) sampled at 1.000 Hz. Corresponding lower body joint angles, moments, and GRFs in the sagittal plane were computed from data measured with an OMC system with 16 infrared cameras (Vicon MX, Oxford, UK) and one force plate (Kistler Instruments Corp, Winterhur, CH), which were sampled at 200 and 1,000, respectively. The speed was measured by two light barriers at a distance of 2 m. In order to analyze right-sided biomechanics, data from four inertial sensors were used; located at the lower back, the lateral right thigh, the lateral right shank, and over the 2nd to 4th metatarsal of the right foot. Sensor positions are shown in **Figure 2**. Sensor data was aligned with segmental axes based on calibrating movements. Eight sagittal plane biomechanical variables were used as a reference: the right-side hip, knee, and ankle flexion angles and moments, and the anterior-posterior (A-P) and vertical GRFs. Biomechanical variables and sensor data were segmented into isolated segments of data from initial contact to initial contact and resampled to 100 time points using linear interpolation. For evaluation in section 2.5, the data from three subjects (S01, S02, and S03) were left out for testing and the data of the remaining subjects (S04-S10) were used for training the CNNs. Simulated data was created from the measured biomechanics of the training subjects.
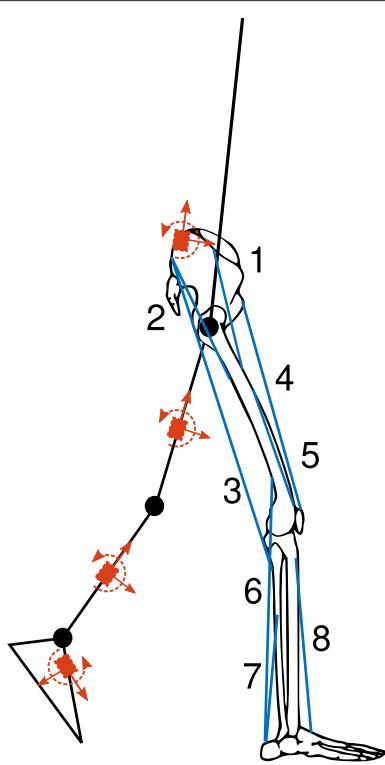
FIGURE 2 | Conceptual drawing of musculoskeletal model consisting of seven rigid segments and 16 Hill-type muscles (blue) with seven virtual inertial sensors (red). The muscles are drawn for the right leg only: 1—iliopsoas, 2—glutei, 3—hamstrings, 4—rectus femoris, 5—vasti, 6—gastrocnemius, 7—soleus, and 8—tibialis anterior. The virtual sensors are drawn for the left leg only simulating sagittal inertial sensor signals: anterior-posterior accelerations, longitudinal accelerations, and medial-lateral angular velocities indicated with red arrows. The figure is taken and modified from Dorschky et al. (2019).

2.2. Random Sampling

We estimated the joint distribution of measured joint angles, GRFs, and walking and running speeds of individual training subjects and drew random samples from these distributions. To achieve this, we concatenated for each walking and running cycle the 100 time points of right-sided hip, knee, and ankle joint angle and the A-P and vertical GRF and the corresponding speed. Thus, every walking and running cycle was described by a vector of $\mathbb{R}^{501}$. For each subject S_i, the vectors of the (approximately) 30 walking and 30 running cycles were stacked to matrices of $\mathbb{R}^{30 \times 501}$, $\mathbf{Z}_{S_i,\text{walking}}$ and $\mathbf{Z}_{S_i,\text{running}}$, whose rows represented observations of the random variable vectors $\mathbf{z}_{S_i,\text{walking}}$ and $\mathbf{z}_{S_i,\text{running}}$, respectively. We assumed multivariate normal distributions: $\mathbf{z}_{S_i,\text{walking}} \sim \mathcal{N}(\boldsymbol{\mu}_{S_i,\text{walking}}, \boldsymbol{\Sigma}_{S_i,\text{walking}})$ and $\mathbf{z}_{S_i,\text{running}} \sim \mathcal{N}(\boldsymbol{\mu}_{S_i,\text{running}}, \boldsymbol{\Sigma}_{S_i,\text{walking}})$. Therefore, we computed the sample means $\boldsymbol{\mu}_{S_i,\text{walking}}$ and $\boldsymbol{\mu}_{S_i,\text{running}} \in \mathbb{R}^{501}$ over the rows of $\mathbf{Z}_{S_i,\text{walking}}$ and $\mathbf{Z}_{S_i,\text{running}}$ and the sample covariance matrices $\boldsymbol{\Sigma}_{S_i,\text{walking}}$ and $\boldsymbol{\Sigma}_{S_i,\text{running}} \in \mathbb{R}^{501 \times 501}$ estimating the covariance between the random variables (the columns of $\mathbf{Z}_{S_i,\text{walking}}/\mathbf{Z}_{S_i,\text{running}}$). We drew 1,000 random samples from each distribution to serve as tracking data for the optimal

control simulation in section 2.3 using Matlab R2018a `mvnrnd` function (Kotz et al., 2004). Random samples of $\mathbf{z}$ were partitioned into joint angles, GRFs, and speed. Joint angles and GRFs were parted in the middle such that they could be used as tracking data for the right and left leg, as only a half symmetric cycle was simulated.

2.3. Simulated Data

We created seven planar musculoskeletal models (Van den Bogert et al., 2012), one for each of the training subjects. Each musculoskeletal model consisted of seven rigid segments (trunk, thighs, shanks, and feet) connected by six hinge joints (hip, knee, ankle in each limb) resulting in nine kinematic degrees of freedom. In addition, each model had 16 Hill-type muscles which are shown in **Figure 2**. The segments of the model were scaled using the bodyweight (BW) and bodyheight (BH) of each subject according to Winter (2009). The multi-body dynamics and muscle dynamics are described in previous publications (Van den Bogert et al., 2011; Dorschky et al., 2019). The unknowns of the model, which were the generalized coordinates and velocities, the muscle activations, muscle lengths, and the contact state, were summarized in state vector $\mathbf{x}(t)$. The control vector $\mathbf{u}(t)$ described the neural excitations of the muscles at time t. The model was simulated to follow random trajectories $m(t)$ of the right and left hip, knee, and ankle angles and anterior-posterior and vertical GRFs while minimizing average muscular effort. We simulated a half walking/running cycle of duration T assuming left-right symmetry, to speed up simulation. The simulation was formulated as the following optimal control problem:

$$
\underset{\mathbf{x}(t),\mathbf{u}(t)}{\text{minimize}} \quad \mathcal{J}(\mathbf{x}(t), \mathbf{u}(t))
$$

$$
= \frac{1}{T} \int_0^T \left(\underbrace{\frac{1}{10} \sum_{j=1}^{10} \frac{(s_j(t) - m_j(t))^2}{\sigma_j(t)^2}}_{\text{track random trajectories}} + \underbrace{\frac{W_{\text{effort}}}{16} \sum_{i=1}^{16} u_i(t)^2}_{\text{muscular effort}} \right) \mathrm{d}t
$$

$$
+ W_{\text{reg}} \mathcal{J}_{\text{reg}} \tag{1a}
$$

subject to

$$
\mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U \tag{1b}
$$

$$
\mathbf{u}_L \leq \mathbf{u} \leq \mathbf{u}_U \tag{1c}
$$

$$
\mathbf{f}(\mathbf{x}(t), \dot{\mathbf{x}}(t), \mathbf{u}(t)) = 0 \tag{1d}
$$

$$
\mathbf{x}(0) + vT\mathbf{e}_x - \mathbf{x}^*(T) = 0. \tag{1e}
$$

The objective function $\mathcal{J}(\mathbf{x}(t), \mathbf{u}(t))$ consisted of a tracking, an effort, and a regularization term with the weights $W_{\text{effort}} = 0.1$ and $W_{\text{reg}} = 0.00001$. The weighting was chosen empirically so that tracking and effort term had about the same magnitude and the regularization term was of lower magnitude. In the tracking term, the quadratic deviation of simulated trajectory $s(t)$ to the prescribed trajectory $m(t)$, normalized to the measured variance $\sigma(t)$, was minimized. Average muscular effort, the mean squared value of muscle excitations, was minimized to resolve muscle ambiguity and to allow the model to deviate from the random trajectories finding a more efficient and potentially more natural movement path. In the regularization term, $\mathcal{J}_{\text{reg}}$, the integral of

TABLE 1 | Architecture of convolutional neural networks with tuned hyperparameters.

Layer	Name	Hyperparameter	Search space	Selected value	Size of output
1	Convolution-ReLU	Kernel_size1, filters1	{3×1, 5×1, 7×1, 3×3, 5×3, 7×3}×{8, 16, 32, 64, 128}	5×3, 64	100×12×64
2	Max-Pooling	Pool_size1	{2×1, 2×2}	2×2	50×6×64
3	Convolution-ReLU	Kernel_size2, filters2	{3×1, 5×1, 7×1, 3×3, 5×3, 7×3}×{16, 32, 64, 128, 256}	5×3, 128	50×6×128
4	Max-Pooling	Pool_size2	{2×1, 2×2}	2×2	25×3×128
5	Flattening	-	-	-	9600
6	Dense-ReLU	-	-	-	100
7	Dense	l2_reg	{0.01,0.001,0.0001}	0.001	100

TABLE 2 | Hyperparameters related to training the convolutional neural networks.

Parameter	Considered values	Selected value
Batch size	{32,64,128,256,516}	64
Learning rate	{0.01,0.001,0.0001}	0.001
Number of epochs	{500,1000,2000,3000}	1000

the sum of squares of the time derivatives of all state and control variables was minimized helping the optimization to converge more quickly.

Equations (1b) and (1c) were the lower (L) and upper (U) bounds of the state vector $\mathbf{x}$ and the control vector $\mathbf{u} \in [0, 5]$ [the same bounds as in Dorschky et al. (2019)]. Dynamic equilibrium was constrained in Equation 1d. To do so, the dynamic equations, which were the multi-body dynamics, muscle dynamics, and contact dynamics (Van den Bogert et al., 2011; Dorschky et al., 2019), were formulated implicitly. In constraint Equation 1e, we enforced symmetry of the right and left leg with a forward translation in direction $\mathbf{e_x}$, where v is the randomly sampled speed (see section 2.2) and $\mathbf{x}^*$ is the mirrored state vector of the right and left leg. The optimal control problem, Equation (1), was solved using direct collocation. The state and control vector were sampled to 50 time points using the Backward Euler method. We used the open source optimizer IPOPT (Wächter and Biegler, 2006) and ran the simulations on a high performance cluster.

The simulation results were expanded to a whole symmetric walking/running cycle with 100 time points. We used the simulated biomechanics of the right leg for training the CNNs in section 2.5. Given the simulated movements, we could extract accelerometer and gyroscope signals at any position of the models. In this work, we used the measured sensor position for each subject from section 2.1 and calculated virtual inertial sensor data as introduced in Dorschky et al. (2019). Gyroscope signals were computed from global trunk orientation and relative joint angular rates. Accelerometer signals were computed from the segment accelerations adding gravity and centrifugal acceleration dependent on sensor position.

2.4. Convolutional Neural Network

We trained CNNs to learn a mapping between inertial sensor data and sagittal plane biomechanical variables for walking/running cycle defined from initial contact to initial contact sampled

at 100 time points. The sampling was chosen to match the simulated data. We trained eight separate CNNs, one for each output variable, namely the right hip, knee, and ankle angles and moments and A-P and vertical GRFs. As input, we used the sagittal plane sensor data of the hip sensor, right thigh sensor, right shank sensor and right foot sensor. We used two accelerometer axes (A-P and longitudinal) and one gyroscope axis (medial-lateral) of each sensor, resulting in an input dimension of 100 × 12. We scaled the data using min-max normalization.

The CNN architecture is based on previous work performing gait analysis from inertial sensor data of segmented strides (Hannink et al., 2017; Zrenner et al., 2018). They used two or three 1D convolutional layers to extract temporal features from accelerometer and gyroscope data. We found that 2D convolutional layers filtering over time and sensor channels were superior to 1D convolutional layers performing just temporal convolutions. They estimated single spatio-temporal gait parameters instead of biomechanical variables over gait cycles. Thus, the number of output nodes was adapted to 100 time points in our work.

Table 1 provides an overview of the network, which consisted of two convolutional layers for feature extraction with zero padding, a stride length of one, and a rectified linear activation function. After each convolutional layer, max-pooling was applied. Two convolutional layers seemed to yield superior performance in comparison to one or three convolutional layers because underfitting occurred in the first case and overfitting in the other case. The data was flattened before passing it to two dense layers for non-linear multivariate regression. The first dense layer had a non-linear rectified linear activation function and 100 nodes. The output layer was a dense layer with linear activation function and 100 nodes. To prevent the model from overfitting, we used L2 kernel regularization. During cross-validation (CV), we inspected the learning curves for overfitting verifying that the validation error did not increase with the number of iterations. We used the ADAM optimizer (Kingma and Ba, 2015) and the mean squared error loss function to train the CNNs. The batch size, learning rate, number of epochs, and L2 regularization factor were empirically set based on the measured training dataset considering specifically the values in **Table 2**. The number of filters, kernel size, and max-pooling were tuned using leave-one-subject-out CV within the seven training subjects (S4-S10) testing the hyperparameters
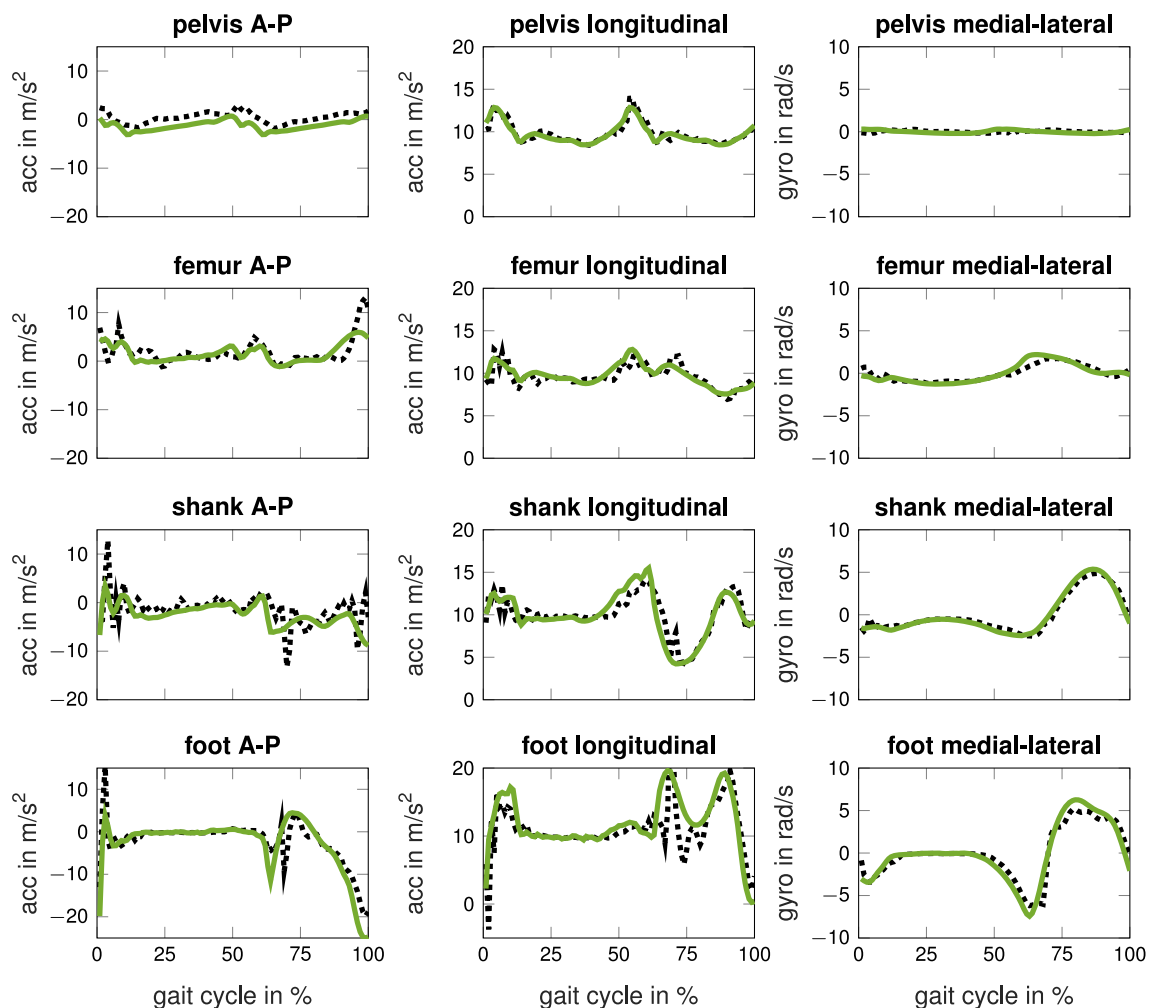
FIGURE 3 | Measured (black dotted) and simulated (green solid) accelerometer (acc) and gyroscope (gyro) data in the sagittal-plane of one subject running at fast speed. The inertial sensors were located at the lower back, the lateral right thigh, the lateral right shank, and at the span of the right foot.

in **Table 1**. The network was implemented in Python using Keras with Tensorflow backend (Chollet, 2015; Abadi et al., 2016). Our implementation of the CNN can be found in the **Supplementary Material**.

2.5. Evaluation

The chosen hyperparameters were fixed for all further evaluations. We trained every CNN with 10 random seeds to test the robustness of results with respect to different random samples of simulated data and random initializations of CNN layers. For comparison purposes, we used the same random seeds for all different training sets. First, we trained the CNNs using only measured data of subjects S04-S10 (training dataset) and tested them with the data of subjects S01-S03 (test dataset). Then, we evaluated how simulated data influences the resulting evaluation metrics. Therefore, we subsequently added simulated data to the training dataset (418 samples) to obtain twice (836 samples), four times (1,672 samples), eight times (3,344 samples), and 16 times (6,688 samples) the amount of training

samples. Simulated data was picked randomly and equally from the 1,000 simulations of each training subject of the walking and running simulations. Thus, the same amount of simulated data was taken from each normal distribution in section 2.2. We used the Python's random module to randomly pick simulated data (Matsumoto and Nishimura, 1998). As we trained every CNN 10 times with different random samples, we made sure that results were robust to random sampling. We trained the networks jointly on simulated and measured training data, which was randomly shuffled at each epoch.

Secondly, we evaluated the model when using less training subjects. We used only four subjects (S07-S10) and two subjects (S09 and S10) for training and tested it with the same three test subjects (S01-S03). For each amount of training subjects, we expanded the respective measured dataset to obtain twice, four times, eight times, and 16 times the amount of training samples. The simulated data was used from the training subjects only: from four subjects (S07-S10) and two subjects (S09 and S10), respectively.

FIGURE 4 | Simulated biomechanics data (green) created from a measured training dataset of seven subjects (blue). Simulated and measured data were used to train data-based models which were tested using the measured data of three independent subjects (red). The anterior-posterior (A-P) and vertical ground reaction force (GRF) are normalized to the bodyweight (BW) of each subject.

For evaluation, we computed the root mean square error (RMSE) and the Pearson correlation coefficient between estimated biomechanics and reference biomechanics. The RMSE was expressed in degrees for joint angles, in BW times BH in percent for joint moments, and in BW percent for GRFs. GRFs were only evaluated over the stance phase using the time points from force plate measurements. For evaluating individual results, performance metrics were computed using all 100 samples of all walking and running cycles of each individual subject and the results were averaged over the 10 random seeds. We used the Fisher-transform to estimate the mean of the Pearson correlation coefficient. For evaluating overall results, performance metrics were computed using all test samples

without separating the results of individual subjects and averaged over the 10 random seeds.

3. RESULTS

Each simulation had a mean CPU time of (3.6 ± 2.0) min on Intel Xeon processors E3-1240, whereas multiple simulations ran in parallel on a cluster. **Figure 3** shows the simulated inertial sensor data and the corresponding measured data. The pattern is similar, while the simulated data is smoother than the measured data. **Figure 4** shows the joint angles, moments, and GRFs of measurements and simulations used for training and the measured data used for testing. The simulated data covers a wider

range than measured data and is more dense. The simulated joint moments show more oscillations, especially in the hip flexion moment. Testing data lies outside of the training data distribution for hip joint angle for S01, knee angle during stance for S02 and peak knee moment for S02.

Training all CNNs including the hyperparameter search took about two weeks on a Nvidia GeForce GTX 1080 Ti. However, inference time of each CNN was less than 1ms per gait cycle.

Tables 3, 4 summarize the individual results of the test subjects for training with the data of all seven training subjects and a different amount of simulated data. In addition, the results of the leave-one-subject-out CV of the seven training subjects are presented using the selected hyperparameters from Tables 1, 2. For all three test subjects, the performance of the CNNs for joint angles increased adding simulated data to the training dataset. The estimation of the hip joint moment was best without using simulated data. Simulated data improved the RMSE of the knee joint moment for all test subjects, whereas the Pearson correlation coefficient only slightly improved for test subjects S01 and S03. The A-P and vertical GRF improved for test subject S01 and S02 adding simulated data, while the performance decreased for test subject S03. Adding more simulated data led to a decrease in performance. Looking at results of the CV, the RMSE of joint angles is lower and Pearson correlation coefficients are higher when simulated data is added. Simulated data did not increase performance for joint moments and vertical GRFs in the CV.

Figure 5 shows the estimated biomechanics for S03 running at fast speed using no simulated data and using seven times more simulated than measured data. The estimated hip angle, ankle angle, and knee moment are closer to the reference when simulated data was added to the training dataset. For example, the peak knee extension moment is higher and the estimated ankle angle is closer to the reference during swing phase.

Figure 6 summarizes the overall results for the cases where the number of training subjects was decreased from seven to four and to two subjects. Reducing the amount of training samples led to higher RMSE values except for the hip angle when training with four instead a seven subjects. Simulated data improved the results for joint angles independent of the amount of training subjects. When increasing the dataset by 16 times, the RMSE of hip, knee, and ankle angle decreased by 17, 27, and 23 % for training with all seven subjects. In the case of training with four subjects, the RMSE of the knee joint angle could even be reduced by 31 %. Moreover, the RMSE of the hip and ankle joint angle was lower when training with simulated and measured data of four subjects compared to training with only measured data of seven subjects. The RMSE of the knee joint angle was lower when training with simulated and measured data of two subjects compared to training with only measured data of seven subjects. However, hip flexion moment was worse for all training data configurations using simulated data. The knee extension moment and vertical GRF improved using simulated data for testing all training data configurations. The RMSE of ankle moment and A-P GRF improved using simulated data, unless data of only two subjects was used for training. When doubling the dataset, the RMSE of knee and ankle moment and vertical GRF decreased by about 6 % for training with all seven subjects. When increasing

the dataset by four times, the RMSE of A-P GRF decreased by about 2 % for training with all seven subjects. Adding more simulated data worsen the estimation of joint kinetics and GRFs.

We added heat-maps, like Figure 6, for the Pearson correlation coefficient to the Supplementary Material. When increasing the dataset by 16 times, Pearson correlation coefficients increased from 0.967 to 0.975 for the hip angle, from 0.988 to 0.992 for the knee angle, and from 0.956 to 0.976 for the ankle angle when training with all seven subjects. The correlations of kinetics were above 0.97 without using simulated data when training with all seven subjects, except for the hip moment with 0.94. Correlations above 0.90 can already be classified as excellent (Taylor, 1990) and are higher than previous work (Dorschky et al., 2019). Correlation coefficients only increased for knee joint moment from 0.970 to 0.971 and for vertical GRF from 0.983 to 0.985 when adding simulated data.

We added individual results of all subjects to the Supplementary Material comparing the RMSE, relative RMSE (Ren et al., 2008), and the Pearson correlation coefficient for a different amount of simulated data. We differentiated between walking and running to allow a better comparison to other work which only focuses on walking or running.

4. DISCUSSION

In this work, we presented a machine learning approach to extract joint angles, joint moments, and GRFs from a combination of simulated and experimental inertial sensor data. The goal was to combine the benefits of physics-based and data-driven approaches: We used simulated data from a physics-based model to reduce exhaustive collection of training data and used this to train data-driven models which can provide low-latency feedback on biomechanics.

The simulated data decreased the generalization error (here RMSE) of the joint angles by up to 31%. Pearson correlation coefficients of joint angles were already between 0.96-0.99 without using simulated data and were ≥ 0.98 with simulated data. Simulated data had a greater effect on RMSE than on correlation coefficients as the RMSE is more sensitive to outliers, and simulated data improved especially the results of outlying subjects. For example, the RMSE of the knee angle improved by 38% for S03 whose ankle dorsiflexion angle was smaller at toe-off compared to the other subjects (compare Figure 4 and Figure 5). For joint moments, the simulated data decreased the generalization error only partly when estimations based on measured data were above average (i.e., above the mean CV error). Simulated data worsened the performance for hip joint moment estimates. This could be explained by the discrepancy between simulated hip joint moments and its reference. This difference is visible in Figure 4, which shows noisy oscillating joint moments for the simulations. One reason may be that only joint angles and GRFs, and no joint moments, were tracked by the musculoskeletal model in Equation 1. Thus, the model tried to follow the predefined joint angles and GRFs using unrealistic (min-max switching) muscle activation patterns. This likely led to the noisy joint moment estimations. A higher

TABLE 3 | The root mean square error (RMSE) of sagittal plane joint angles, joint moments, and anterior-posterior (A-P) and vertical ground reaction force (GRF) is presented for varying ratios between simulated (sim) and measured (meas) data.

	Sim/meas data	Hip angle degree	Knee angle degree	Ankle angle degree	Hip moment BWBH%	Knee moment BWBH%	Ankle moment BWBH%	A-P GRF BW%	Vertical GRF BW%
CV	0	5.38 (1.57)	5.22 (1.22)	5.50 (1.64)	**1.62 (0.23)**	**1.14 (0.13)**	**1.32 (0.42)**	4.33 (0.48)	**14.44 (2.08)**
	1	5.19 (1.36)	4.95 (1.38)	5.00 (1.52)	1.66 (0.20)	1.21 (0.08)	1.38 (0.42)	4.36 (0.49)	14.75 (4.58)
	3	**5.08 (1.77)**	**4.81 (1.19)**	4.86 (1.53)	1.75 (0.23)	1.27 (0.08)	1.35 (0.39)	4.51 (0.31)	15.28 (3.55)
	7	5.17 (1.44)	5.09 (1.65)	4.72 (1.35)	1.76 (0.29)	1.35 (0.16)	1.36 (0.37)	4.35 (0.41)	15.10 (3.06)
	15	5.37 (1.57)	4.93 (1.20)	**4.60 (1.32)**	1.78 (0.30)	1.28 (0.15)	1.39 (0.34)	4.63 (0.52)	16.07 (3.75)
S01	0	9.42 (0.48)	4.45 (0.41)	3.29 (0.25)	**1.71 (0.11)**	1.21 (0.12)	**0.88 (0.12)**	4.52 (0.26)	11.74 (0.88)
	1	8.98 (0.59)	4.28 (0.55)	3.54 (0.47)	1.88 (0.12)	**1.07 (0.07)**	0.98 (0.07)	4.70 (0.28)	10.46 (0.86)
	3	9.11 (0.26)	3.87 (0.31)	3.23 (0.38)	1.97 (0.13)	1.31 (0.12)	1.00 (0.12)	4.23 (0.18)	**9.99 (0.76)**
	7	8.94 (0.55)	3.57 (0.27)	3.49 (0.23)	2.01 (0.10)	1.30 (0.11)	1.03 (0.10)	4.22 (0.13)	12.33 (0.80)
	15	**8.77 (0.49)**	**3.31 (0.34)**	**2.87 (0.30)**	2.07 (0.11)	1.36 (0.15)	1.05 (0.10)	**3.76 (0.31)**	13.53 (1.04)
S02	0	6.49 (0.59)	10.44 (1.31)	4.40 (0.57)	**1.44 (0.10)**	2.06 (0.25)	1.86 (0.21)	4.41 (0.38)	13.24 (1.20)
	1	6.32 (0.89)	8.69 (0.49)	4.24 (0.28)	1.71 (0.20)	**2.04 (0.21)**	**1.59 (0.14)**	4.03 (0.46)	**12.16 (0.77)**
	3	5.39 (0.57)	7.70 (0.44)	4.24 (0.46)	1.81 (0.06)	2.08 (0.20)	1.67 (0.15)	4.21 (0.18)	13.67 (1.01)
	7	4.47 (0.46)	**7.26 (0.45)**	4.47 (0.32)	1.89 (0.10)	2.27 (0.13)	1.78 (0.15)	4.40 (0.30)	12.42 (0.83)
	15	**3.69 (0.19)**	7.29 (0.40)	**4.18 (0.50)**	1.95 (0.12)	2.39 (0.25)	1.73 (0.17)	4.21 (0.15)	15.34 (1.21)
S03	0	3.71 (0.24)	5.52 (0.56)	6.31 (0.49)	**1.32 (0.04)**	1.96 (0.08)	**1.05 (0.07)**	4.29 (0.23)	**12.91 (0.62)**
	1	3.43 (0.31)	4.82 (0.39)	4.43 (0.24)	1.61 (0.13)	1.76 (0.15)	1.11 (0.03)	5.10 (0.29)	13.75 (0.93)
	3	3.10 (0.18)	4.47 (0.23)	4.30 (0.31)	1.62 (0.07)	1.79 (0.09)	1.24 (0.11)	4.36 (0.25)	14.25 (1.17)
	7	**3.00 (0.14)**	**4.36 (0.34)**	4.01 (0.35)	1.72 (0.09)	**1.75 (0.13)**	1.20 (0.09)	4.83 (0.27)	14.75 (0.72)
	15	3.06 (0.19)	4.62 (0.12)	**3.94 (0.38)**	1.78 (0.11)	1.77 (0.08)	1.21 (0.04)	4.95 (0.21)	16.24 (1.08)

Joint moments and GRFs are normalized to bodyweight (BW) and bodyheight (BH). The first rows show the mean RMSE and its standard deviation of the leave-one-subject-out cross-validation (CV) on the training dataset for the chosen hyperparameter. The subsequent rows show the mean RMSE and standard deviation over 10 random seeds for the three test subjects S01-S03 using the data of seven subjects for training. Bold highlighting indicates the lowest mean value in the respective column.

weighting of the effort term in the optimal control simulation might lead to smoother muscle activations and thus muscle forces and joint moments. Joint moments could also be tracked in the optimal control simulations. However, the results for joint angles and GRFs might get worse. Another reason may be that the reference joint moments are too smooth, as filtering of marker data and force plate data was applied before computing joint moments (Dorschky et al., 2019). Overall, the reference joint moments were not directly measured but estimated using inverse dynamics. Thus, error accumulation lead to inaccuracies especially for the hip joint moment.

The estimation result of GRFs was already better without simulated data compared to previous work with Pearson correlations >0.97. In Dorschky et al. (2019), the RMSE of A-P and vertical GRF was 5% BW and 15% BW. In this work, the mean RMSE of A-P and vertical GRF was about 3% BW and 10% BW using only measured data (analyzing the GRFs over the complete cycle). The simulations were created using the same musculoskeletal model as in Dorschky et al. (2019), who reported errors in the estimation of GRFs and ankle joint moments due to model inaccuracies, as the foot was modeled with a single rigid segment. Consequently, simulated data only partly enhanced the estimation of GRFs and ankle joint moments in **Tables 3**, **4**.

A direct comparison to previous work is difficult as different datasets of varying number of sensors, sensor positions, subjects,

and movements were used for evaluation. Machine learning models dedicated to one single task, for example, for estimating single joint angles or specialized for walking only, will probably outperform our machine learning models which were jointly tuned for different output variables. In order to fairly compare different approaches, they would all need to be tested using the same datasets. The presented machine learning approach outperforms our previous physics-based approach (Dorschky et al., 2019) evaluated on the same data-set. In contrast to physics-based approaches, machine learning models require representative training data. Combining simulated and measured data seems a promising approach (Mundt et al., 2020a). In this work, we focused on the comparison between learning on measured and learning on simulated data to evaluate whether simulations can decrease the generalization error by incorporating variations of movement. Future work should expand this method to 3D analysis and evaluate against state of the art methods (Stetter et al., 2019; Mundt et al., 2020a). 3D biomechanical optimal control simulations are more expensive to compute, but are advancing recently (Falisse et al., 2019).

The network architecture was specialized for pre-segmented walking and running cycles and a fixed input and output dimension. The segmentation and sampling was chosen to match with the simulation with a fixed number of collocation nodes. We trained the CNNs separately in order reduce

TABLE 4 | The Pearson correlation coefficient of sagittal plane joint angles, joint moments, and anterior-posterior (A-P) and vertical ground reaction force (GRF) is presented for varying ratios between simulated (sim) and measured (meas) data.

	Sim/meas data	Hip angle	Knee angle	Ankle angle	Hip moment	Knee moment	Ankle moment	A-P GRF	Vertical GRF
CV	0	0.969	0.989	0.962	**0.940**	**0.975**	**0.981**	0.970	**0.980**
	1	**0.974**	**0.990**	0.967	0.937	0.969	0.974	**0.971**	0.979
	3	0.973	**0.990**	0.970	0.931	0.964	0.974	0.968	0.977
	7	0.973	0.990	0.972	0.931	0.958	0.972	0.970	0.977
	15	0.973	**0.990**	**0.975**	0.927	0.958	0.971	0.967	0.975
S01	0	0.953	0.991	0.975	**0.920**	0.976	**0.985**	0.979	0.988
	1	0.958	0.990	0.981	0.899	**0.977**	0.983	0.980	**0.991**
	3	0.960	0.992	0.985	0.900	0.970	0.983	**0.982**	**0.991**
	7	**0.962**	0.993	0.985	0.880	0.968	0.982	0.980	0.989
	15	0.959	**0.994**	**0.987**	0.865	0.965	0.982	0.980	0.985
S02	0	0.970	0.989	0.962	**0.948**	**0.947**	**0.982**	0.972	0.979
	1	0.972	0.990	0.969	0.932	0.946	0.979	0.966	**0.983**
	3	**0.975**	0.990	0.975	0.938	0.941	0.973	0.971	0.980
	7	**0.975**	0.992	0.978	0.935	0.946	0.972	0.973	0.981
	15	**0.975**	**0.993**	**0.980**	0.936	0.941	0.972	**0.974**	0.980
S03	0	0.975	0.982	0.910	**0.948**	0.979	**0.978**	**0.976**	**0.981**
	1	0.982	0.988	0.941	0.924	**0.981**	0.970	0.970	**0.981**
	3	**0.984**	0.990	0.940	0.918	0.977	0.962	0.971	0.976
	7	0.983	**0.991**	0.948	0.906	0.974	0.961	0.969	0.974
	15	0.982	0.990	**0.949**	0.899	0.973	0.959	0.966	0.974

The first rows show the mean Pearson correlation coefficient of the leave-one-subject-out cross-validation (CV) on the training dataset for the chosen hyperparameter. The subsequent rows show the mean Pearson correlation coefficient over ten random seeds for the three test subjects S01-S03 using the data of seven subjects for training. Bold highlighting indicates the lowest mean value in the respective column.

the output dimension and consequently the amount of trainable parameters in the network to avoid overfitting. It has been shown that individual CNNs can outperform bigger networks with multiple output variables (Hannink et al., 2017). However, the first layers of the different networks, which act as feature extractors, probably share some common features such that multi-task learning or transfer learning might improve results (Caruana, 1997). Future work should consider different network architectures which avoid pre-processing (segmentation into walking and running cycles and resampling) of sensor data like fully (circular) convolutional networks and allow a continuous estimation of movement biomechanics using recurrent architectures like long short-term memory networks (Mundt et al., 2020b). In addition, the feature extraction using convolutional layers should be explored. In the CV, two dimensional convolution yielded superior results compared to one dimensional convolutions over time which are typically used for inertial sensor data (Hannink et al., 2017). The 2D convolution was applied over time and over adjacent sensor axes, where data is likely to be correlated. The order of sensor axes was not optimized and data of accelerometers and gyroscopes were not split, although different feature extractors for different sensor types may yield better results. As CNNs were tuned on measured data, we assume that different architectures would not influence the comparison between learning on measured and simulated data.

A reality gap was apparent between simulated and measured inertial sensor data. Simulated inertial sensor data were less noisy than measured data (e.g., **Figure 3** longitudinal acceleration of foot sensor). We modeled a rigid attachment of virtual sensors on the musculoskeletal model. In reality, the connection is loose due to soft tissue, which could be considered by a wobbling mass model. Another option is to use domain adaptation learning. For example, generative adversarial networks could be trained to learn a mapping between simulated and measured data (Shrivastava et al., 2017). In preliminary work, we learned a direct mapping between simulated and measured data using supervised learning. This yielded worse results which might be explained because end-to-end learning is typically superior. Further investigations and evaluations are necessary here.

In this work, we jointly learned from simulated and measured data. In our case, this approach worked better than training on simulated data and fine-tuning on real data. We assume that robust features were learned which were invariant to noise and movement artifacts. However, overfitting to simulated data was observed, for example for the vertical GRF where the performance decreased when adding three to fifteen times as much simulated as measured data. Instead of random sampling (see section 2.2), simulated data could only be created for those points where the current model is least certain. Thus, outliers could be covered with simulated data, whereas the performance of data that lies within the measurements would
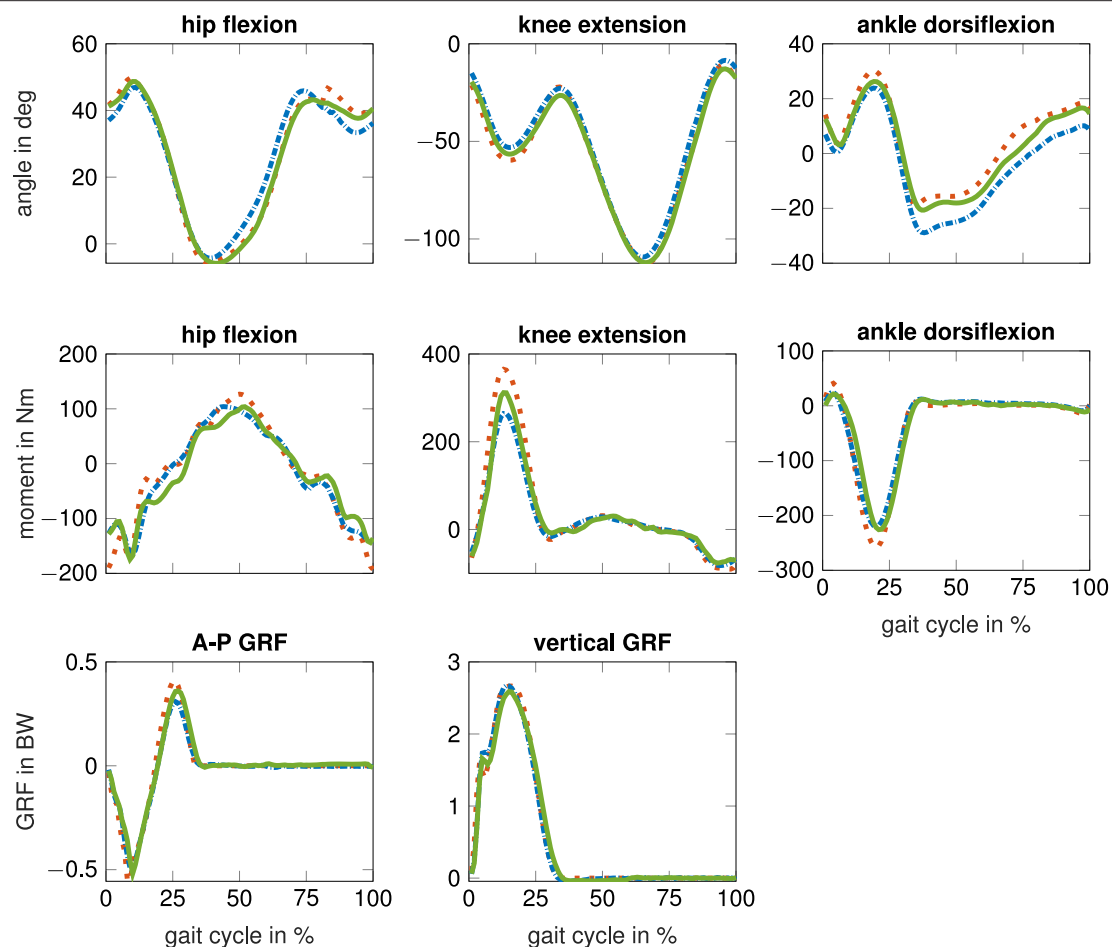
FIGURE 5 | Results for test subject S03 running at fast speed: reference biomechanics from optical motion capturing (dotted red) compared to estimated biomechanics from inertial sensor data using no simulated data (blue dashed dotted) and seven times as much simulated as measured data (green solid). The anterior-posterior (A-P) and vertical ground reaction force (GRF) are normalized to the bodyweight (BW).

ideally not be affected. Future work should consider methods, where simulated data is generated iteratively during training within a closed loop. For example, Ruiz et al. (2018) proposed a meta-learning algorithm to learn how to simulate. The algorithm should adjust parameters of a simulator to generate synthetic data such that a machine learning model achieves a higher accuracy.

Data augmentation is commonly used to artificially expand a data set for training deep neural networks, but most approaches use only label-preserving transformations of input data (e.g., adding noise or rotating sensor axes, Um et al., 2017). In contrast, the presented method creates new pairs of input and output data such that a wider range of movement mechanics is covered. In this work, we generated the simulated data based on the training data distribution of the individual training subjects to take into account intra-subject variability. The simulated data filled the sparsely populated space of measured training data, as more variations of movements and speeds were included in the training set. This can be seen in **Figure 4** where the simulated data covers a wider range of biomechanics and less space between curves is

apparent. However, on the one hand not all test data is covered within the simulated and measured data (see e.g., maximum knee extension moment) because of inter-subject variability. On the other hand, we surmise that the simulated data was spread too widely for GRFs as the estimated variance was high especially for initial contact. When we used simulated data closer to the mean of measured data for training the CNNs, the estimation of joint moments and GRFs was slightly better, but the estimation of joint angles was slightly worse. Future work may consider to use more light-tailed data distributions than multivariate normal distributions.

Results depended on the training data distribution. For example, the hip angle improved when training with four instead of seven subjects, likely because the testing data distribution better matched that of the training data distribution of the four subjects. To cover a wider range of movement variations and to achieve a representative dataset, different data sources could be combined using the biomechanical simulation. Public datasets of movement biomechanics could be tracked with the musculoskeletal model to obtain corresponding inertial sensor

FIGURE 6 | Overall results of the root mean square error (RMSE) for the estimated sagittal plane biomechanical variables. The vertical axis indicates the ratio between simulated (sim) and measured (meas) data used for training. The horizontal axis indicates the number of training subjects whose data were used for training. In addition, the mean RMSE of the leave-one-subject-out cross-validation (CV) is shown. Joint moments and the anterior-posterior (A-P) and vertical ground reaction force (GRF) are normalized to bodyweight (BW) and bodyheight (BH).

data. Instead of tracking joint angles and GRFs, video data or inertial sensor data could be tracked with the model (Heinrich et al., 2014; Dorschky et al., 2019). This shows the potential of using optimal control simulations to create labeled training data (corresponding inertial sensor data and biomechanics). Simulated inertial sensor data at different sensor positions could easily be obtained.

While the recording of measured data (without post-processing) took about two weeks, it only took a few hours to create the same amount of ready-to-use simulated data with the implemented simulation framework. As shown in **Figure 6**, the estimation of joint angles was even better using

a reduced dataset with simulated data compared to using all measured data without simulated data. On the one hand, using simulated data increases the number of samples and thus minimizes the risk of overfitting. On the other hand, simulated data includes additional variations of movement such that unseen data is covered with a higher probability. Simulated data would be of great advantage for rare events and abnormal movements where training data is hard to acquire, for example, for detecting an impending fall. Overall, biomechanical simulations can supplement time-consuming and expensive data collections to achieve a better generalization of machine learning models.

In summary, we presented a novel approach to generate an (in principle) infinite set of inertial sensor data with corresponding biomechanical variables using optimal control simulations of walking and running. We evaluated training on simulated data compared to solely learning on measured data. The simulated data improved the estimation of joint angles. The simulation-aided estimation of joint moments and GRFs was limited by inaccuracies of the musculoskeletal model. Improving the physics-based model or domain adaptation learning may help to reduce the gap between real and simulated data. The current method is a first step of using optimal control simulation for training deep neural networks and was evaluated for sagittal plane biomechanics only. In future work, this method should be evaluated for 3D biomechanical analysis. In addition, different datasets could be combined using the optimal control simulation in order to create representative datasets of human movement.

In conclusion, machine learning can benefit from available domain knowledge on biomechanical simulations to supplement cumbersome data collections. This enables the training of robust data-driven models that can provide real-time feedback on biomechanics "in the wild," for example, to reduce injury risk, for rehabilitation movement training, or for controlling active assisting devices such as exoskeletons.

DATA AVAILABILITY STATEMENT

Please contact the corresponding author to request the datasets.

AUTHOR CONTRIBUTIONS

ED performed the biomechanical simulations and trainings of the neural networks and wrote the paper. MN and AB supported the implementation of the biomechanical simulations. CM supported the conception end evaluation of the machine learning approach. AK and BE supervised the overall conception and design of the work. All authors reviewed the paper and approved the final manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00604/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]. arXiv:1603.04467.*

Blank, P., Kugler, P., and Eskofier, B. M. (2015). "miPod-A wearable sports and fitness sensor," in *10th Symposium der DVS Sektion Sportinformatik der Deutschen Vereinigung für Sportwissenschaft* (Wien), 78–79.

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi: 10.1023/A:1007379606734

Chollet, F. (2015). *Keras.* Available online at: https://github.com/fchollet/keras (accessed April 22, 2020).

Dorschky, E., Nitschke, M., Seifer, A.-K., van den Bogert, A. J., and Eskofier, B. M. (2019). Estimation of gait kinematics and kinetics from inertial sensor data using optimal control of musculoskeletal models. *J. Biomech.* 95:109278. doi: 10.1016/j.jbiomech.2019.07.022

Falbriard, M., Meyer, F., Mariani, B., Millet, G. P., and Aminian, K. (2018). Accurate estimation of running temporal parameters using foot-worn inertial sensors. *Front. Physiol.* 9:610. doi: 10.3389/fphys.2018.00610

Falisse, A., Serrancolí, G., Dembia, C. L., Gillis, J., Jonkers, I., and De Groote, F. (2019). Rapid predictive simulations with complex musculoskeletal models suggest that diverse healthy and pathological human gaits can emerge from similar control strategies. *J. R. Soc. Interface* 16:20190402. doi: 10.1098/rsif.2019.0402

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., and Delp, S. L. (2018). Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J. Biomech.* 81, 1–11. doi: 10.1016/j.jbiomech.2018.09.009

Hannink, J., Kautz, T., Pasluosta, C. F., Gaßmann, K.-G., Klucken, J., and Eskofier, B. M. (2017). Sensor-based gait parameter extraction with deep convolutional neural networks. *IEEE J. Biomed. Health Inform.* 21, 85–93. doi: 10.1109/JBHI.2016.2636456

Heinrich, D., van den Bogert, A. J., and Nachbauer, W. (2014). Relationship between jump landing kinematics and peak ACL force during a jump in downhill skiing: a simulation study. *Scand. J. Med. Sci. Sports* 24, 180–187. doi: 10.1111/sms.12120

Huang, Y., Kaufmann, M., Aksan, E., Black, M. J., Hilliges, O., and Pons-Moll, G. (2018). "Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time," in *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018* (New York, NY). doi: 10.1145/3272127.3275108

Johnson, W. R., Mian, A., Robinson, M. A., Verheul, J., Lloyd, D. G., and Alderson, J. A. (2019). Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning. *arXiv [Preprint]. arXiv:1903.07221.*

Karatsidis, A., Jung, M., Schepers, H. M., Bellusci, G., de Zee, M., Veltink, P. H., et al. (2018). Predicting kinetics using musculoskeletal modeling and inertial motion capture. *arXiv [Preprint]. arXiv:1801.01668.* doi: 10.1016/j.medengphy.2018.12.021

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (San Diego, CA), 1–15.

Kok, M., Hol, J. D., and Schön, T. B. (2014). An optimization-based approach to human body motion capture using inertial sensors. *IFAC Proc. Vol.* 47, 79–85. doi: 10.3182/20140824-6-ZA-1003.02252

Komaris, D., Pérez-Valero, E., Jordan, L., Barton, J., Hennessy, L., O'Flynn, B., et al. (2019). Predicting three-dimensional ground reaction forces in running by using artificial neural networks and lower body kinematics. *IEEE Access* 7, 156779–156786. doi: 10.1109/ACCESS.2019.2949699

Koning, B. H. W., van der Krogt, M. M., Baten, C. T. M., and Koopman, B. F. J. M. (2013). Driving a musculoskeletal model with inertial and magnetic measurement units. *Comput. Methods Biomech. Biomed. Eng.* 18, 1003–1013. doi: 10.1080/10255842.2013.867481

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2004). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. Hoboken, NJ: John Wiley & Sons.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Matsumoto, M., and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* 8, 3–30. doi: 10.1145/272991.272995

Miezal, M., Taetz, B., and Bleser, G. (2017). "Real-time inertial lower body kinematics and ground contact estimation at anatomical foot points for agile human locomotion," in *Proceedings - IEEE International Conference on Robotics and Automation* (Singapore: IEEE), 3256–3263. doi: 10.1109/ICRA.2017.7989371

Mundt, M., Koeppe, A., David, S., Witter, T., Bamer, F., Potthast, W., et al. (2020a). Estimation of gait mechanics based on simulated and measured IMU data using an artificial neural network. *Front. Bioeng. Biotechnol.* 8:41. doi: 10.3389/fbioe.2020.00041

Mundt, M., Thomsen, W., Witter, T., Koeppe, A., David, S., Bamer, F., et al. (2020b). Prediction of lower limb joint angles and moments during gait using artificial neural networks. *Med. Biol. Eng. Comput.* 58, 211–225. doi: 10.1007/s11517-019-02061-3

Norgaard, S., Saeedi, R., Sasani, K., and Gebremedhin, A. H. (2018). "Synthetic sensor data generation for health applications: a supervised deep learning approach," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 1164–1167. doi: 10.1109/EMBC.2018.8512470

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., Howard, D., Meijer, K., and Crompton, R. (2009). Activity identification using body-mounted sensors-a review of classification techniques. *Physiol. Meas.* 30, R1–R33. doi: 10.1088/0967-3334/30/4/R01

Ren, L., Jones, R. K., and Howard, D. (2008). Whole body inverse dynamics over a complete gait cycle based only on measured kinematics. *J. Biomech.* 41, 2750–2759. doi: 10.1016/j.jbiomech.2008.06.001

Roetenberg, D., Luinge, H., and Slycke, P. (2009). XSENS MVN: full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV*. Technical Report.

Ruiz, N., Schulter, S., and Chandraker, M. (2018). Learning to simulate. *arXiv preprint arXiv:1810.02513*.

Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., et al. (2019). Wearable sensors for monitoring the internal and external workload of the athlete. *NPJ Digital Med.* 2:71. doi: 10.1038/s41746-019-0149-2

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). "Learning from simulated and unsupervised images through adversarial

training," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (Honolulu, HI), 2242–2251. doi: 10.1109/CVPR.2017.241

Stetter, B. J., Ringhof, S., Krafft, F. C., Sell, S., and Stein, T. (2019). Estimation of knee joint forces in sport movements using wearable sensors and machine learning. *Sensors* 19:3690. doi: 10.3390/s19173690

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *J. Diagnost. Med. Sonogr.* 6, 35–39. doi: 10.1177/875647939000600106

Um, T. T., Pfister, F. M. J., Pichler, D., Endo, S., Lang, M., Hirche, S., et al. (2017). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *arXiv [Preprint]. arXiv:1706.00527*. doi: 10.1145/3136755.3136817

Van den Bogert, A. J., Blana, D., and Heinrich, D. (2011). Implicit methods for efficient musculoskeletal simulation and optimal control. *Proc, IUTAM* 2, 297–316. doi: 10.1016/j.piutam.2011.04.027

Van den Bogert, A. J., Hupperets, M., Schlarb, H., and Krabbe, B. (2012). Predictive musculoskeletal simulation using optimal control: effects of added limb mass on energy cost and kinematics of walking and running. *Proc. Instit. Mech. Eng. P J. Sports Eng. Technol.* 226, 123–133. doi: 10.1177/1754337112440644

Veiga, J. J. D., O'Reilly, M., Whelan, D., Caulfield, B., and Ward, T. E. (2017). Feature-free activity classification of inertial sensor data with machine vision techniques: method, development, and evaluation. *JMIR mHealth and uHealth* 5:e115. doi: 10.2196/mhealth.7521

Wächter, A., and Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.* 106, 25–57. doi: 10.1007/s10107-004-0559-y

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470549148

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., van Beijnum, B. J. F., et al. (2018). Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors. *Front. Physiol.* 9:218. doi: 10.3389/fphys.2018.00218

Zell, P., and Rosenhahn, B. (2019). Learning inverse dynamics for human locomotion analysis. *Neural Comput. Appl.* 1–15. doi: 10.1007/s00521-019-04658-z

Zrenner, M., Gradl, S., Jensen, U., Ullrich, M., and Eskofier, B. M. (2018). Comparison of different algorithms for calculating velocity and stride length in running using inertial measurement units. *Sensors* 18:4194. doi: 10.3390/s18124194

Check for updates

Supervised Machine Learning Applied to Wearable Sensor Data Can Accurately Classify Functional Fitness Exercises Within a Continuous Workout

Ezio Preatoni[1], Stefano Nodari[2] and Nicola Francesco Lopomo[2]*

[1] *Department for Health, University of Bath, Bath, United Kingdom,* [2] *Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Brescia, Italy*

Observing, classifying and assessing human movements is important in many applied fields, including human-computer interface, clinical assessment, activity monitoring and sports performance. The redundancy of options in planning and implementing motor programmes, the inter- and intra-individual variability in movement execution, and the time-continuous, high-dimensional nature of motion data make segmenting sequential movements into a smaller set of discrete classes of actions non-trivial. We aimed to develop and validate a method for the automatic classification of four popular functional fitness drills, which are commonly performed in current circuit training routines. Five inertial measurement units were located on the upper and lower limb, and on the trunk of fourteen participants. Positions were chosen by keeping into account the dynamics of the movement and the positions where commercially-available smart technologies are typically secured. Accelerations and angular velocities were acquired continuously from the units and used to train and test different supervised learning models, including k-Nearest Neighbors (kNN) and support-vector machine (SVM) algorithms. The use of different kernel functions, as well as different strategies to segment continuous inertial data were explored. Classification performance was assessed from both the training dataset (k-fold cross-validation), and a test dataset (leave-one-subject-out validation). Classification from different subsets of the measurement units was also evaluated (1-sensor and 2-sensor data). SVM with a cubic kernel and fed with data from 600 ms windows with a 10% overlap gave the best classification performances, yielding to an overall accuracy of 97.8%. This approach did not misclassify any functional fitness movement for another, but confused relatively frequently (2.8–18.9%) a fitness movement phase with the transition between subsequent repetitions of the same task or different drills. Among 1-sensor configurations, the upper arm achieved the best classification performance (96.4% accuracy), whereas combining the upper arm and the thigh sensors obtained the highest level of accuracy (97.6%) from 2-sensors movement tracking.

We found that supervised learning can successfully classify complex sequential movements such as those of functional fitness workouts. Our approach, which could exploit technologies currently available in the consumer market, demonstrated exciting potential for future on-field applications including unstructured training.

Keywords: automatic classification, inertial measurement unit, sport, on-field testing, activity monitoring, machine learning, wearable sensors

INTRODUCTION

The problem of tracking, identifying and classifying human actions has received increasing interest over the years, as it plays a key role in many applied contexts, such as human-computer interface (Popoola and Wang, 2012; Sarig Bahat et al., 2015; Quitadamo et al., 2017; Bachmann et al., 2018), daily-life activity monitoring (Mannini and Sabatini, 2010; Cheng et al., 2015; Chetty and White, 2016), clinical assessment (Rawashdeh et al., 2016; Arifoglu and Bouchachia, 2017; Howell et al., 2017) and sports performance (Attal et al., 2015; Ghazali et al., 2018; Hsu et al., 2018). The development of unobtrusive technologies for motion capture (e.g., wearable inertial measurement units—IMUs), their widespread integration in relatively cheap, commercially available devices (e.g., smartphones, watches, activity trackers, heart rate monitors, sensorized insoles), and the push toward healthier, more active life styles, have generated a multitude of existing and potential applications where automatic movement classification and assessment is fundamental (Attal et al., 2015; Cheng et al., 2015; Cust et al., 2019).

Sport coaching and training still largely rely on visual observation and subjective feedback, and they could benefit from quantitative input supporting decision making. Having quantitative real-time information about the amount, quality and intensity of the work carried out may play an important role at multiple levels. It could inform coaching and strength & conditioning planning, help monitoring training load, and evaluating the quality of movement performance (i.e., the outcome achieved) and movement execution (i.e., technique). It could also help improving injury prevention, as continuous monitoring could enable systematic screening of movement behavior, help identifying risk factors and mechanisms of injury, and support decision making in terms of pre- and rehabilitation programmes (Jones and Wallace, 2005).

Motion capture has traditionally relied on optical-based solutions, but recent development in microelectronics has generated increased interest and research efforts into wearable technologies (Adesida et al., 2019). Wearable systems are particularly suitable to sport-specific needs (van der Kruk and Reijne, 2018), since: (1) sport usually takes place in uncontrolled and unstructured settings, with environmental conditions difficult to be predicted *a priori* (e.g., weather, interaction with equipment and other people) and many possible measurement interferences (e.g., electromagnetic noise); (2) the size of the acquisition volume inherently depends on the type of practiced sport (e.g., team vs. individual, indoor vs. outdoor); (3) sensors used to capture sports movements should be both robust and non-obtrusive for the athlete (i.e., ecologically transparent). Systems based on wearable devices, including low-cost activity trackers, smartwatches and smartphones (Ahmad et al., 2017), have kept evolving and are widely available for the consumer market, including clinical uses and sports applications (Ghazali et al., 2018; Hsu et al., 2018). Wearable technologies for motion analysis are predominantly inertial measurement units (IMUs) (Davila et al., 2017), which, thanks to their low cost and minimal obtrusiveness, represent an optimal solution for tracking and assessing sports movement on-field (Hsu et al., 2018; van der Kruk and Reijne, 2018; Adesida et al., 2019).

Despite the widespread of wearable technology in both applied and research environments, the use of wearable data as input of algorithms for the detection and classification of human actions remains non-trivial, especially in sport. Indeed, sport activities typically involve a large variety of movements, execution technique demonstrates inherent inter- and intra-individual variability, and data is of high-dimensionality (Endres et al., 2012; Hsu et al., 2018). For this reason, no "one-size-fits-all" approach exists (Crema et al., 2019), and bespoke solutions have been reported to address only specific needs, including: recognition/classification (i.e., "what type" of task a subject performs) or identification of the achieved performance (i.e., "how good" the subject performs the task, with respect to a specific reference). In this perspective, the literature has focused the analysis on very specific sport activities and tasks (Cust et al., 2019).

Among fitness activities, functional training combines aerobic conditioning, weightlifting, interval training, balancing, gymnastics, and functional fitness movements (i.e., exercises that mimic daily life requirements, such as lifting weights) performed at high level of intensity (Liebenson, 2006). Functional fitness has been shown to improve cardiovascular capacity, muscle tone and central nervous system efficiency (Barbieri et al., 2019; Singh and Saini, 2019), but may also increase risk of musculoskeletal injuries affecting shoulder, lower back and knee joints (Gianzina and Kassotaki, 2019). It is therefore important to provide athletes with reliable feedback about their efforts, and guide them toward safe movement technique. The availability of a quantitative system for the monitoring of movement completion and overall performance would aid coaching and judging. Functional fitness workouts often consist of continuous sequences of movements, and the identification and assessment of individual elements within the sequence currently relies on visual observation and the expertise of the coach. The wide spectrum of situations in terms of dynamics and body part involved represents a difficult challenge for the automatic classifications of activities and makes

it a good proof of concept for the scope of our study. Being able to identify specific movements within a complex movement sequence could be the starting point of a number of useful applications such as counting the number of movement tasks completed, and hence assessing technique and training load for both performance and injury prevention purposes.

We aimed to develop and validate a bespoke algorithm for the automatic recognition and classification of four popular functional fitness drills, when performed in a continuous workout. In particular, we wanted to test the capability of supervised machine learning approaches when fed with data from a network of five wearable inertial sensors on the body. Also, we carried out a sensitivity analysis, which could indicate whether subsets of the available measuring units could still provide acceptable classification performance.

MATERIALS AND METHODS

Population
Fourteen healthy participants (11 males and three females, age 18–50) with at least 6-months experience in functional training activities volunteered to take part in this study. All participants were physically active, free from any neurological disease and musculoskeletal condition at the time of testing, and familiar with the movement tasks to be performed. The study protocol received ethical approval by the local research ethics committee (reference number EP 17/18 247). Volunteers were informed about experimental procedures and signed informed consent before participating. Based on the existing literature (Cust et al., 2019) and the exploratory nature of the study, a sample size >12 was deemed adequate to address the research objectives.

Experimental Setup
Five wearable units (Trigno Avanti Wireless EMG System, Delsys Inc., USA) were secured to the participants via double-sided hypoallergenic tape and elastic straps. IMUs were located onto specific anatomical landmarks (**Figure 1**), which included the left ankle, thigh, upper arm and wrist, and trunk (L5-S1 level). These positions were chosen to: (a) reproduce the locations where commercially available devices with embedded motion monitors (e.g., smart watches, smart phones, shoe-sensors) could be positioned; and, (b) to capture whole body information and drill dynamics whilst allowing the natural execution of movements, avoiding obstruction or discomfort for the subject. The wearable units embed tri-axial accelerometers, gyroscopes and magnetometers and were able to synchronously communicate with the system base station via Bluetooth Low Energy (BLE) wireless protocol ensuring an acquisition rate of 148.15 Hz.

Accelerations and angular velocities (±16 g, $\pm2,000^\circ$/s) were acquired continuously throughout the workout by means of the wearable units; magnetometer measurements were excluded due to the presence of ferromagnetic materials, very close to the acquisition volume. Data coming from the sensors were
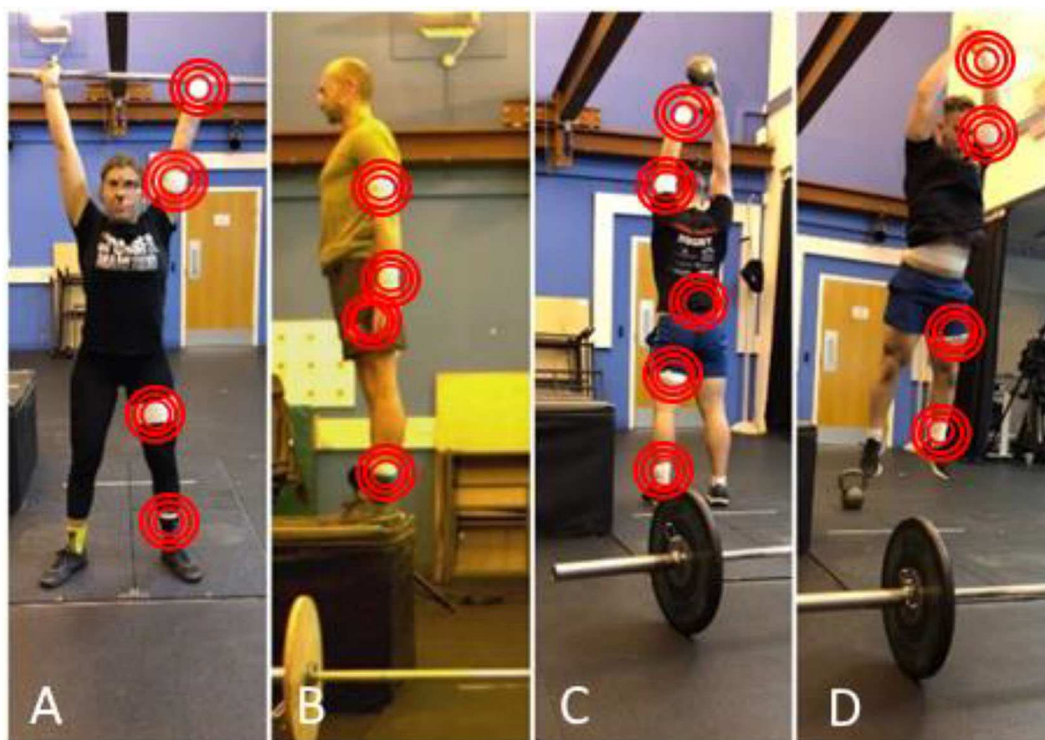
FIGURE 1 | Experimental setup and movement tasks, where the position of IMU sensors has been highlighted. **(A)** "Clean and Jerk," **(B)** "Box Jump," **(C)** "American Swing," and **(D)** "Burpee". All the five sensors were worn by the participants throughout the execution of the protocol.

FIGURE 2 | Schematic representation of the execution stages of the fitness training drills used in this study: **(A)** "Clean and Jerk," **(B)** "American Swing," **(C)** "Box Jump," and **(D)** "Burpee".
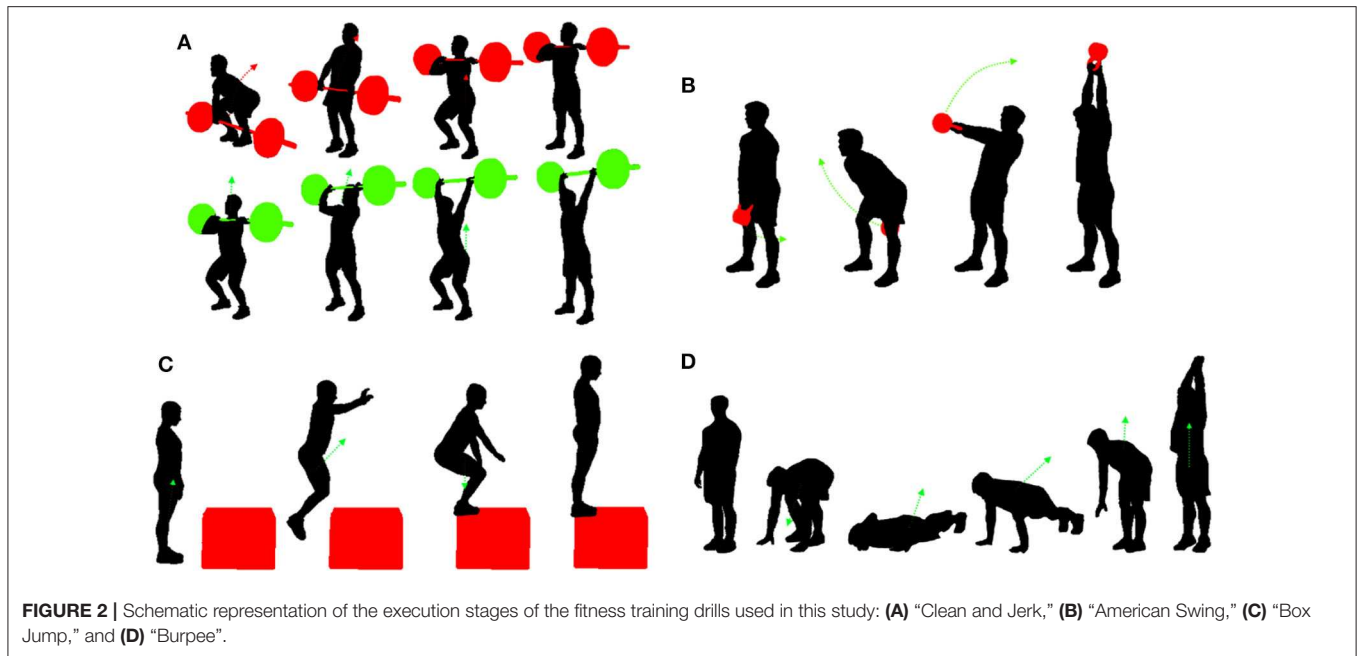
synchronized with a commercial video-camera (Oqus Video 210c, Qualisys AB, Sweden; 50 Hz) via a dedicated trigger module (Trigger Module, Delsys Inc. USA) connected to both systems. The video camera was positioned in front of workout station, thus allowing the correct acquisition of all the performed movements.

Experimental Protocol

Participants were asked to execute a workout session including four popular functional training drills (**Figure 2**). These consisted of:

- *"Clean and Jerk"* (*C&J*). A weighted barbell is lifted from the ground to over the head in two subsequent movements: the "clean," where the barbell is pulled from the floor to a racked position across the shoulders, and "jerk," where the barbell is raised above the head, and a stable position is achieved by keeping straight legs and arms, and feet, torso and barbell lie in the same plane.
- *"American Swing"* (*AS*). A kettlebell is grasped with both hands and swung from below the groin to above the head, keeping the arms straight. The upward momentum of the kettlebell is predominantly generated by the explosive extension of the hip.
- *"Box Jump"* (*BJ*). The participant start from a standing position in front of a box, performs a countermovement jump to land on top of it, achieves a stable upright position, and completes the task by returning to the start position.
- *"Burpee"* (*BP*). A four-stage exercise, where the participant starts from a standing position, squats placing the hands on the floor, kicks back into a plank position while keeping the arms extended, returns in the squat position and, jumps up extending the upper limbs overhead.

All the movement tasks were illustrated to the participants at the start of the session, following the standards approved for competition (CrossFit, 2019; WODstar, 2019). A 50 cm box was used in the Box Jump exercise for all participants, whereas drills with an added resistance were differentiated between female and male participants, and set to, respectively: 20 and 40 kg in the Clean & Jerk; 12 and 16 kg in the American Swing.

After a self-directed warm up, and some repetitions to familiarize with the experimental setup, each participant performed 3 sets of functional fitness activities structured as follows:

Set 1 (classifier training dataset):

- 3× C&J + 3× BJ + 3× AS + 3× BP

Set 2 (workout simulation session, classifier test dataset):

- *1st Round:* 1× C&J + 1× BJ + 1× AS + 1× BP
- *2nd Round:* 2× C&J + 2× BJ + 2× AS + 2× BP
- *3rd Round:* 3× C&J + 3× BJ + 3× AS + 3× BP
- *4th Round:* 4× C&J + 4× BJ + 4× AS + 4× BP

Set 3 (classifier training dataset):

- 3× C&J + 3× BJ + 3× AS + 3× BP

Five-minute recovery was allowed between sets, whereas movements were executed sequentially with no rest allowed between repetitions of the same exercise, different exercises or rounds. This was done to ensure ecological validity with respect to a real functional fitness training session, and to challenge the capability of the classifying algorithm to recognize movements when they are performed without clear breaks in-between them. The order of movement execution was randomized between participants, to avoid possible bias due to repetitive patterns.

TABLE 1 | Features extracted from each time window of each signal collected (from accelerometers and gyroscopes), and then used as input of the classification algorithm.

Time domain	Frequency domain
Mean value (Magnitude)	Mean value
Standard deviation	Power
Root mean square	Higher frequency
Mean absolute deviation	Lower frequency
Max value	Median frequency
Min value	Mean frequency
Kurtosis	Spectral entropy
Skewness	
Quartile (25th, 50th, 75th)	

The workout simulation (Set 2) was preceded (Set 1) and followed (Set 3) by a sequence of three repetitions of each task. The pre- and post-workout session were used as training sets for the machine learning algorithms, and were both included so that the classification method could be robust to the possible changes in movement execution caused by fatigue or learning effects in the participants.

Data Analysis

The three components of acceleration and angular velocity from the five IMUs ($6 \times 5 = 30$ continuous timeseries) were used as input data for the classifying algorithm. Kinematic quantities were not filtered, and frequency-domain signals were attained through transforming the time-domain signals via Fast Fourier Transform. Data features in the time and frequency domain (**Table 1**) were extracted from data windows moving across the original kinematics timeseries. This process aimed to reduce the signals into distinctive characteristics of specific movement tasks or part of them. The more each movement can be separated in feature space, the higher the achieved recognition and classification performance is Zhang and Sawchuk (2013) and Hoettinger et al. (2016).

To set suitable ranges for window duration and decide the amount of window overlap, we analyzed the distribution of movement durations across the population (**Figure 3**). Values between 300 and 600 ms (in increments of 100 ms) for window length, and of 0, 10, and 20% for the amount of overlap were chosen to study the sensitivity of the classification to the choice of windowing parameters. This allowed to have at least three time windows covering the execution of each movement or the transitions between subsequent movements. A [N × 540] feature matrix was generated for each participant, where N indicates the number of time windows in each session, and 540 is the overall number of features included in the analysis (5 sensors × 2 kinematic quantities per sensor × 3 directions per quantity × 18 features per quantity).

Data Labeling

A supervised approach to automatic classification was adopted, with video-based classification used as the gold standard for labeling each data window as a transition phase or as a part of one of the four possible functional fitness movements (**Figure 2**). Camera footage was used to identify the start and end of each movement and for their classification (i.e., labeling), as required by the supervised learning model. Movement recognition, timing and labeling were visually carried out by a single expert operator using freeware video editing software (VirtualDub, virtualdub.org). When a window spanned between a transition phase and one of the four movement tasks, a "majory" criterion was used. This implied assigning a movement label (i.e., "C&J," "AS," "BJ," or "BP") to a window where the movement covered more than 66% of its length. Otherwise, the transition label (i.e., "TRANS") was allocated.

Classifier Training

After the extraction of the features and the labeling of associated windows, we trained different type of automatic classifiers using data Set 1 and 3. k-Nearest Neighbors (kNN), with different types of metrics (Euclidean, cosine, cubic or weighted distance) and number of neighbors (fine, k = 1; medium, k = 10; and coarse, k = 100), and Support Vector Machine (SVM) with several types of kernel functions (i.e., linear, quadratic, cubic and fine-medium-coarse Gaussian), were selected as the classifying algorithms to be tested. This choice was driven by the existing literature in the area of machine learning approaches addressing human motion (Camomilla et al., 2018; Cust et al., 2019) and sport (Cust et al., 2019) classification. At this stage, all the reported features (**Table 1**) were used to train the models.

Classifier Assessment

Two levels of classifier evaluation were carried out. Firstly (Stage 1), we performed a 5-fold cross-validation on the classifier training dataset (N = 14 participants, Set 1 and 3); this approach was used to mitigate the risk of overfitting by partitioning the dataset into k-folds and estimating the accuracy of each fold (Taha et al., 2018). We used this stage to select the most promising algorithm amongst the many tested. Finally (Stage 2), assessed the classifier performance on new data (i.e., the workout simulation dataset, Set 2) in a Leave-On-Subject-Out (LOSO) fashion (Hagenbuchner et al., 2015; Willetts et al., 2018). In this stage, the classifier was trained with data from Set 1 and 3 (including N-1 participants), and validated against data of the N-th participant, from Set 2; the N-th participant was iteratively changed, and results were reported averaging the multiple iterations. This approach guaranteed having independent data, in terms of both trials and individuals, between training and testing sets.

Classification accuracy (Equation 1) was evaluated as follows (Hoettinger et al., 2016; Davila et al., 2017):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP, TN, FP and FN represent True Positive, True Negative and False Positive, respectively.

Once the optimal classifier was identified, the corresponding confusion matrix and Receiver Operating Characteristic (ROC) curves (in a multi-label "one-vs.-rest" assessment) were analyzed
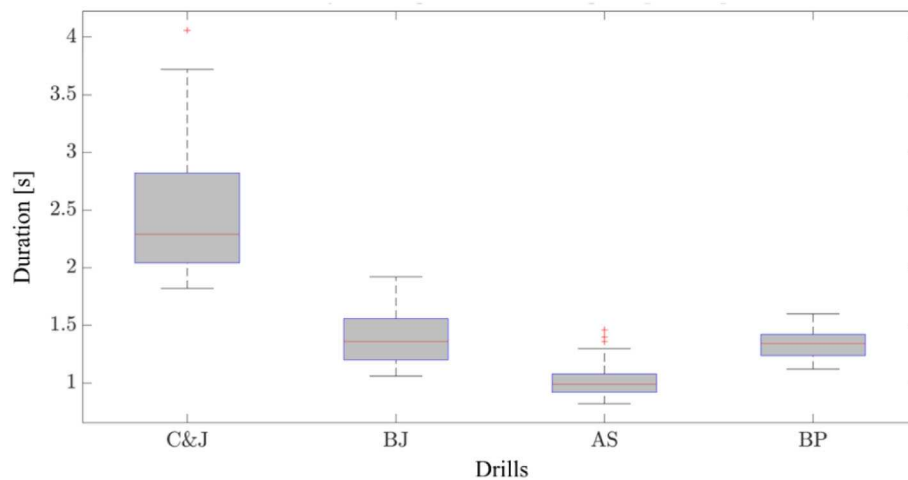
FIGURE 3 | Distribution of individual movement durations in the analyzed population. C&J, Clean and Jerk; AS, American Swing; BJ, Box Jump; BP, Burpee.

to assess the ability of the algorithm to recognize and correctly classify each functional fitness exercise. From the confusion matrix, for each exercise, we evaluated:

- the Positive Predictive Value (PPV), representing the precision of the classifier (Equation 2):

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

- the True Positive Rate (TPR), representing the sensitivity (also called recall) of the classifier (Equation 3):

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

Sensitivity Analysis
Two type of sensitivity analysis were carried out: (a) the effect of window length and overlapping, where all classifier types and all the features were included; and, (b) the effect of selecting a subset of the five available IMUs, which was analyzed starting from the classifier previously identified as giving the best outcome performance (as highlighted in validation Stage 1). For (b), the analysis was carried out starting from the data provided by each sensor in isolation and by considering data from pairs of sensors, as follows:

- wrist and ankle;
- wrist and lumbar area;
- wrist and thigh;
- wrist and upper arm;
- upper arm and ankle;
- upper arm and lumbar area;
- upper arm and thigh.

Feature Selection Analysis
Once the best performing subset of the five measurement units was identified, an exploratory analysis of the most significant features extracted was carried out. We used the minimum Redundancy Maximum Relevance (mRMR) filter-based algorithm applied to the standardized feature matrix, due to its trade-off between performance and efficiency (Peng et al., 2005; Wang et al., 2016). To compare the overall accuracy, a fixed number of features was identified starting from the analysis of the predictor importance scores performed on the training dataset; these features were then used to train the models and to test them following Stage 2 validation.

Training of the supervised learning models and analysis of classification performance were carried out through the Statistics and Machine Learning Toolbox and bespoke functions developed in Matlab (v R2019b, The Mathworks Inc.).

RESULTS

k-Fold Cross-Validation of Classifier Performance and Sensitivity Analysis: Time Window and Overlap Parameters
When data input included all the five available sensors, both SVM- and kNN-type classifiers achieved good level of overall accuracy (**Tables 2**, **3**, respectively). Accuracy ranged from 82.5% (SVM classifier with fine gaussian kernel, and 300 ms−10% overlap windows) to 97.8% (cubic kernel SVM classifier, with 600 ms−10% overlap windows).

Testing SVM Performance With Training and Test Datasets
Considering the overall accuracy, the training time (a ratio of more than 20 between the slowest and the fastest classifier) and the computational costs (a ratio of more than 80 between the fastest and slowest classifiers, in terms of prediction speed), the SVM with cubic kernel applied to 600 ms−10% overlap

TABLE 2 | Overall classification performance (accuracy, in %) for Support Vector Machine (SVM) algorithms, as a factor of different kernel functions, window lengths, and percentage of window overlap.

		Overlap [%]			Overlap [%]			Overlap [%]		
		0	10	20	0	10	20	0	10	20
		Linear			**Quadratic**			**Cubic**		
Window [ms]	300	95.6	95.5	95.7	96.7	96.7	96.8	97.1	97.0	97.3
	400	95.7	96.0	96.2	97.2	97.1	97.1	97.3	97.3	97.3
	500	96.3	96.6	96.4	97.1	97.1	97.1	97.4	97.4	97.2
	600	96.2	97.0	96.4	97.0	97.7	97.7	97.0	**97.8**	97.7
		Fine gaussian			**Medium gaussian**			**Coarse gaussian**		
Window [ms]	300	82.6	**82.5**	82.7	95.5	95.5	95.7	93.5	93.7	94.0
	400	82.8	82.9	82.8	95.4	95.6	96.0	93.8	94.0	94.0
	500	83.6	83.2	83.3	95.2	95.4	95.5	93.6	94.6	94.3
	600	83.6	83.7	83.4	95.2	95.8	95.5	93.6	94.6	94.2

Data from all the five IMUs available were used as input. Green bold numbers = best performance; red bold numbers = worst performance.

TABLE 3 | Overall classification performance (accuracy, in %) for k-Nearest Neighbors (kNN) algorithms, as a factor of different kernel functions, window lengths and percentage of window overlap.

		Overlap [%]			Overlap [%]			Overlap [%]		
		0	10	20	0	10	20	0	10	20
		Fine Class			**Medium Class**			**Coarse Class**		
Window [ms]	300	96.4	96.4	96.7	96.3	96.0	96.5	89.2	89.8	90.6
	400	96.3	96.3	97.0	96.1	96.5	96.8	**89.0**	90.1	90.5
	500	96.1	97.0	97.0	96.0	96.5	96.1	89.7	89.8	90.4
	600	96.2	**97.2**	96.7	95.7	96.4	96.2	**89.0**	89.9	90.5
		Cosine			**Cubic**			**Weighted**		
Window [ms]	300	96.4	96.3	96.6	94.0	93.7	94.1	96.4	96.3	96.8
	400	96.3	96.5	96.5	93.9	94.2	94.9	96.4	96.7	96.9
	500	96.1	96.7	96.3	94.6	94.7	94.8	96.3	96.6	96.4
	600	96.1	**97.2**	96.5	94.7	94.9	94.8	96.3	96.9	96.5

Data from all the five IMUs available were used as input. Green bold numbers = best performance; red bold numbers = worst performance.

windows appeared as the optimal learning model. The confusion matrix for this classifier (**Table 4**) showed that the trained model yielded to almost no (validation Stage 1) or few (validation Stage 2) misclassifications between different functional fitness movements. Specific accuracy ranged from 99.7% for burpees in the 5-fold cross-validation to 94.3% for the transition phase when tested on new data. All but one erroneous classification in the 5-fold cross-validation were from movement tasks identified as transition phases (64, 1.6% of the total) and, less frequently, from transitions confused for functional fitness drills (19, 0.5%). We had up to 18.9% of false negative rates in the AS drill, which reported the lowest level of precision (93.0%) and sensitivity (81.1%) (**Table 5**). Similar outcomes, but with lower percentage values, were reported by the LOSO validation on the test dataset. Precision and sensitivity values were always highest in the transition movements (94.9 and 97.8%, respectively), whereas the

Clean & Jerk (89.3 and 82.2%) and American Swing (93.0 and 79.3%) showed the lowest performance results (**Table 5**).

The analysis of ROC curves gives us the power of our classifier in a multi-label classification problem, as a function of the Type I error (i.e., 1—specificity), as it was a binary predictor. Considering the validation stages, the selected SVM classifier showed an almost null value for FPR in each functional fitness movement (<1%), with the TPR ranging from 84% (BJ classification) to 96% (BP classification). The highest value of TPR was reached in the classification of transition phases (99%), although, in TRANS the classifier also reported the highest level of FPR (7%). Finally, the Area Under the Curve (AUC), which describes the capability of the supervised learning model to distinguish between one class and the others, ranged between 0.98 and 1, therefore showing good overall classification performances.

TABLE 4 | Confusion matrixes for the cubic kernel SVM algorithm with a 600 ms window length and 10% overlap.

Stage 1: 5-fold Cross-Validation – Training Dataset

		Predicted class				
		C&J	AS	BJ	BP	TRANS
True class	C&J	228 (5.9%)				22 (0.6%)
	AS		107 (2.8%)			25 (0.6%)
	BJ			92 (2.4%)		13 (0.3%)
	BP		1 (0.0%)		138 (3.6%)	4 (0.1%)
	TRANS	2 (0.1%)	7 (0.2%)	5 (0.1%)	5 (0.1%)	3215 (83.2%)

Stage 2: LOSO – Test Dataset

		Predicted class				
		C&J	AS	BJ	BP	TRANS
True class	C&J	434 (8.1%)	4 (0.1%)			90 (1.7%)
	AS	3 (0.1%)	214 (4.0%)		2 (0.0%)	51 (0.9%)
	BJ			174 (3.2%)		29 (0.5%)
	BP	2 (0.0%)			280 (5.2%)	43 (0.8%)
	TRANS	47 (0.9%)	12 (0.2%)	12 (0.2%)	20 (0.4%)	3961 (73.3%)

Classification performance is reported from the two stages of validation as total counts and % of total. Blank cells correspond to a count of zero. LOSO = "leave one subject out." Training = sensor data from movements of Set 1 and Set 3 of the experimental protocol, used to train the classifier. Test = sensor data from movements of Set 2 of the experimental protocol, not used to train the algorithm. Results are the average across multiple iterations. Green = correct prediction, red = misclassification.

TABLE 5 | Accuracy (ACC), precision (PPV, Positive Predictive Value) and sensitivity (TPR, True Positive Rate) for each functional fitness movement, related to the cubic kernel SVM algorithm with a 600 ms length windows and 10% overlap.

	ACC		PPV		TPR	
	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2
C&J	99.4	97.3	99.1	89.3	91.2	82.2
AS	99.1	98.7	93.0	93.0	81.1	79.3
BJ	99.5	99.2	94.8	93.5	87.6	85.7
BP	99.7	98.8	96.5	92.7	96.5	86.2
TRANS	97.9	94.3	98.0	94.9	99.4	97.8

Data from all the five IMUs available were used as input. Classification performance is reported from the two stages of validation as %, and results are the average across multiple iterations. Stage 1= 5-fold Cross-Validation—Training Dataset; Stage 2 = "leave one subject out" (LOSO) on test dataset. Green bold numbers = best performance; red bold numbers = worst performance.

Sensitivity Analysis: Number of Sensors

When considering the data coming from a single sensor, the selected SVM classifier achieved good values of recognition rates in most cases (**Table 6**), with an overall accuracy between 83.2% (data from the ankle sensor, validation Stage 2) and 96.4% (data from the upper arm sensor, cross-fold validation). Using input data from pairs of IMUs generally improved the overall classification accuracy, pushing it up of several percentage point when testing on new data (Stage 2: from 83.2−91.0% to 92.0−93.0%). However, using two sensors did not match the performances obtained when data from all the sensors were utilized (93.0 vs. 97.8%).

In relation to the contribution of each sensor to the correct classification of individual functional drills, including data from the sensor placed on the upper limb (upper arm or wrist), or from a combination of a sensor on the upper limb and a sensor on the lumbar area or thigh, seemed to improve classifier performance, in at least 3 out of 4 movements and in the transition phases (**Tables 7**, **8**). Only in the AS, the classifier seemed to perform relatively better when using data from the sensor placed on the lumbar spine (single sensor configuration). The worst overall performance was obtained when considering the data acquired by the only sensor placed on the ankle. Only for the AS, the algorithm did worse considering the data registered by the sensor placed on the wrist.

Feature Selection Analysis

From the sensitivity analysis we identified two configurations to be further tested by using the feature selection. We considered the data collected by the sensor on the upper arm (UA configuration, for a total of 108 features) and by the combination of sensors on the upper arm and thigh (UA+T configuration, for a total of 216 features). After a qualitative analysis of the trend in prediction scores, from the most important predictor to the less significant, we set the number of the features to keep to 20.

The reduction of the number of the features did not compromise the overall accuracy of the classifier, thus underlining the reliability of the approach. In particular, the

TABLE 6 | Overall classification performance (accuracy, in %) for Support Vector Machine (SVM) algorithms, with 600 ms−10% overlap windows.

Validation	W	UA	T	A	L	W+A	W+L	W+T	W+UA	UA+A	UA+L	UA+T
Stage 1	94.5	**96.4**	93.5	**92.4**	93.5	**96.4**	96.5	97.0	96.8	96.8	97.4	**97.6**
Stage 2	89.3	91.0	86.8	**83.2**	87.5	92.0	92.0	92.6	92.2	92.2	**93.0**	**93.0**

Data from individual or pairs of IMUs were used as input. Stage 1 = 5-fold Cross-Validation−Training dataset; Stage 2 = "leave one subject out" (LOSO) on workout dataset. Green bold numbers = best performance; red bold numbers = worst performance. W, Wrist; UA, Upper Arm; T, Thigh; A, Ankle; L, Lumbar Segment.

TABLE 7 | Accuracy (ACC), precision (PPV, Positive Predictive Value) and sensitivity (TPR, True Positive Rate) for each functional fitness movement, related to the cubic kernel SVM algorithm with a 600 ms length windows and 10% overlap.

	W			UA			T			A			L		
	ACC	PPV	TPR	ACC	PPV	TPR	ACC	PPV	TPR	ACC	PPV	TPR	ACC	PPV	TPR
C&J	**96.4**	85.7	**75.9**	96.2	**86.7**	72.9	93.0	65.1	62.9	**90.9**	**53.6**	**55.7**	93.2	67.6	58.5
AS	**96.8**	81.3	**48.1**	97.3	**78.5**	63.7	97.5	81.7	65.9	97.6	**82.6**	65.2	**97.8**	82.1	**71.1**
BJ	98.6	83.9	76.8	**99.1**	**90.9**	**83.7**	96.9	61.3	45.3	**95.4**	**36.5**	**28.1**	98.1	87.9	57.1
BP	97.0	**88.3**	**58.2**	97.9	83.6	**81.5**	97.5	85.3	69.8	97.1	79.5	69.2	**97.0**	**78.9**	68.0
TRANS	89.7	90.2	**96.8**	91.5	**92.7**	96.3	88.6	90.6	94.7	**85.3**	**89.0**	**91.8**	89.0	90.5	95.5

Data from individual IMUs were used as input. Classification performance is reported as %. For sake of clarity, only results from the most stringent validation (Stage 2 = "leave one subject out" on test dataset) are reported. Green bold numbers = best performance; red bold numbers = worst performance. W, Wrist; UA, Upper Arm; T, Thigh; A, Ankle; L, Lumbar Segment.

highest value of accuracy was maintained when considering the UA configuration (99.1% for BJ), whereas UA+T configuration reported a reduction of only 0.3% (99.1 vs. 98.8%) (**Table 9**). Furthermore, in both configurations, all the values of accuracy were >89.5% (TRANS in UA). Larger differences concerned precision and recall in classifying the AS task, which decreased to 59.3% and 49.6% (UA), and 71.9% and 62.6% (UA+T), respectively. For AS alone, both PPV and TPR decreased by 15–20%, showing risk of misclassification.

Most of the identified features were time-domain features (15 out of 20 for the UA configuration and 16 out of 20 for the UA+T configuration), and was information coming from gyroscope data (13 out of 20 for the UA configuration and 12 out of 20 for the UA+T configuration). In UA+T, the identified features were equally spread between the sensor placed on the upper arm and on the thigh (10 out of 20, each).

DISCUSSION

We developed and tested a supervised learning approach to recognizing and classifying functional fitness movements within a continuous workout, combining four different drills. Accelerations and angular velocities from a set of wearable inertia sensors were used as input of the classifier. Different machine learning algorithms, time segmentation strategies and combination of sensors were assessed. Classification accuracy was generally high in both Support Vector Machine (SVM) and k-Nearest Neighbors approaches (>82.5% in the worst case); the SVM model with cubic kernel and applied to 600 ms−10% overlap data windows gave the best performance overall (94.4– 97.8% accuracy, depending on the type of validation carried

out). Information coming from sensors from the upper limb, alone or in combination with a wearable unit in the lumbar area or on the thigh, appeared to be key to achieve optimal classification performance.

By using SVM on the whole dataset, misclassifications (as False Negative Rate—NFR) were lower in the "Transition" phase (0.6–2.2%) and higher in the other four drills, particularly in the "American Swing" (18.9–20.7%). *Ex post* analysis highlighted that the higher percentages of errors could be typically related to three main factors. (1) The overall smaller number of windows associated with functional movements as opposed to transitions. (2) The choice made for the "majority" criterion, whereby up to 34% of a functional movement could still belong to a window labeled as TRANS. This may have an influence on the capability of the classifier to assign a window to one of the four drills instead of TRANS. (3) The difficulty in labeling windows as belonging to a movement or TRANS between repetitions of the same exercise, when the dynamics of the task makes it difficult to establish with certainty the start and end of the movement. Combining these three items, the problem appeared more evident for the "American Swing," possibly for the inherent dynamics of the task.

When analyzing the contribution of each sensor independently (1-sensor input) or in combination with another IMU (2-sensor input), the overall classification performance decreased of few percentage points, but still achieved an accuracy >83.2% in the worst case (i.e., IMU on the ankle, with the most stringent validation approach). Ankle kinematics may contain less information when feet are not moving; this situation may happen in a number of movement- and transition-related situations, such as during the "Clean and Jerk," thus explaining the decreased performance of the classifier.

TABLE 8 | Accuracy (ACC), precision (PPV, Positive Predictive Value) and sensitivity (TPR, True Positive Rate) for each functional fitness movement, related to the cubic kernel SVM algorithm with a 600 ms length windows and 10% overlap.

		C&J	AS	BJ	BP	TRANS
W+A	ACC	96.4	98.4	98.8	98.0	92.5
	PPV	83.5	91.8	89.1	84.8	93.7
	TPR	79.5	74.4	**76.8**	80.6	**96.5**
W+L	ACC	96.6	**98.5**	**98.7**	**97.8**	92.5
	PPV	86.1	91.3	**84.5**	**84.6**	93.6
	TPR	77.7	**77.4**	80.8	77.5	96.6
W+T	ACC	**97.3**	98.1	98.7	97.9	**93.0**
	PPV	87.8	90.4	87.7	88.4	93.7
	TPR	**84.7**	69.6	77.3	**75.1**	97.3
W+UA	ACC	97.0	**97.8**	99.1	98.0	**92.4**
	PPV	**89.5**	**88.1**	90.9	87.6	93.0
	TPR	79.0	**65.6**	83.3	78.2	97.2
UA+A	ACC	**95.9**	98.2	99.1	**98.6**	92.6
	PPV	**81.2**	90.6	**92.8**	89.1	93.8
	TPR	**75.4**	71.1	82.3	**87.7**	96.6
UA+L	ACC	96.7	98.4	99.0	98.5	**93.3**
	PPV	88.6	**93.3**	89.0	89.9	93.8
	TPR	76.3	72.6	83.7	84.9	**97.6**
UA+T	ACC	96.7	98.2	**99.1**	98.5	**93.3**
	PPV	84.6	89.2	90.6	**91.6**	**94.4**
	TPR	80.9	73.7	**85.7**	83.7	96.9

Data from pairs of IMUs were used as input. Classification performance is reported as %. For sake of clarity, only results from the most stringent validation (Stage 2 = "leave one subject out" on workout dataset) are reported. Green bold numbers = best performance; red bold numbers = worst performance. W= Wrist, UP= Upper Arm, T= Thigh, A= Ankle, L= Lumbar Segment.

In fact, collecting upper arm kinematics alone yielded 91.0–96.4% accuracy (depending on the validation approach). Also, adding information from a second sensor generally improved the capability of the algorithm to identify classes correctly, narrowing the performance gap between using two IMUs or the whole sensor network. The best combinations resulted from adding one further IMU to one sensor on the upper arm, i.e., upper arm and lumbar area (93.0–97.4%) or thigh (93.0–97.6%), which further confirms the need for the system to cover the widest range of movement dynamics. Similarly to what observed for the whole sensor network, misclassifications were more common in the "American Swing" (31.8–36.3% and 21.2–26.3% FNR for the UA and UA+T configurations, respectively).

To explore the translation of the selected algorithm into more easily applicable framework, a subset of features, consisting of the best 20 identified through a filter-based algorithm (mRMR), was used in a 1-sensor or 2-sensor configuration, and its classification ability tested (LOSO validation on the test dataset). The overall accuracy resulted better than 90% for all the performed task, although the confusion matrixes highlighted difficulties in distinguishing "similar" gestures (AS misclassified with TRANS). Further analysis of feature selection suggested that the most informative characteristics of the dataset were mainly related to time domain (i.e., kurtosis and skewness). Although

TABLE 9 | Accuracy (ACC), precision (PPV, Positive Predictive Value) and sensitivity (TPR, True Positive Rate) for each functional fitness movement, related to the cubic kernel SVM algorithm with a 600 ms length windows and 10% overlap.

		ACC (%)	PPV (%)	TPR (%)
UA	C&J	95.4	79.6	70.8
	AS	95.8	**59.3**	**49.6**
	BJ	**99.1**	90.0	84.2
	BP	97.3	78.8	75.4
	TRANS	**89.5**	**91.7**	**94.6**
UA+T	C&J	96.0	79.3	80.5
	AS	96.9	**71.9**	**62.6**
	BJ	**98.8**	85.5	81.3
	BP	97.5	82.7	75.1
	TRANS	**91.9**	**93.9**	**95.4**

Data from the IMU placed on the upper arm (UA) and the combination of upper arm and thigh (UA+T) were used as input. Classification performance is reported from the only Stage 2 = "leave one subject out" (LOSO) validation on test dataset. Green bold numbers = best performance; red bold numbers = worst performance in each stage.

these preliminary findings support the use of feature reduction in the pipeline of data processing, a more in-depth analysis of feature selection and outcomes derived thereof is advisable, especially for 1-sensor solutions with lower-end technology (Fan et al., 2019).

Supervised machine learning appeared a suitable tool for the automatic classification of different functional fitness exercises. Our study addressed a scenario that for number and type of movements involved appears more challenging than what has been assessed by other works in the field. Also, we located our sensors according to where existing consumer technologies would be placed, and not thinking of what the best configuration for motion capture would be. Despite these added complexities, our approach obtained similar performance to what reported by the literature as the current state of the art. Ghazali et al. (2018) achieved 91.2% accuracy in tracking several common sporting activities such as walking, sporting, jogging sprinting and jumping. Using wearable sensors and SVM/kNN methods, Mannini and Sabatini (2010) were able to distinguish between elementary physical activities such as standing, sitting, lying, walking, climbing and identify activities within sequences of sitting-standing-walking-standing-sitting with an accuracy between 97.8 and 98.3%.

Within fitness activities, Adelsberger and Troster (2013), studied 16 participants performing a squat press, and via SVM managed to detect movements with 100% accuracy and differentiate between expert and beginner performance (94% accuracy). Research on weightlifting has used different approaches, mainly aiming at recognizing the type of exercise performed (Pernek et al., 2015; Hausberger et al., 2016; O'Reilly et al., 2017), or identifying performance metrics (e.g., quality of execution, intensity, deviation from a standard pattern) for each exercise (Pernek et al., 2015; O'Reilly et al., 2017a,b,c). Approaches looking at performance metrics focus on the possibility of using personalized classifiers to monitor the quality

of movement execution; they are more complex and demanding in terms of computational resources and sample sizes than what we presented in our study. On the other hand, the solutions presented in literature to address the movement recognition problem are very similar to what we have proposed. Different algorithms (such as the Random Forest—RF—and the Linear Discriminant Analysis—LDA) have been explored in the existing literature, but the overall accuracies appear comparable to the values we obtained. For the size of our dataset, SVM resulted optimal in terms of both classification accuracy and training costs. RF represented an optimal solution in multi-class problem in terms of performance and computational costs, despite requiring larger datasets (O'Reilly et al., 2017), whereas LDA was reported to perform well in simple drills classification, even allowing real-time applications, when considering a single sensor (Crema et al., 2019).

One of the main limitations of the presented work was the reduced number of involved subjects, compared to some validated machine learning approaches found in scientific literature (O'Reilly et al., 2017,a,b). Our study was exploratory, and the observed sample was relatively homogeneous in terms of sporting abilities. Having access to a larger and more varied group of participants would allow covering a wider spectrum of individual characteristics and, possibly, making the classifier more robust to inherent intra- and inter-subject variability (Preatoni et al., 2013) in movement execution. It could also allow to distinguish between expert and novice performance and/or between different level of movement intensity. Although our sample size was relatively small for typical machine learning studies, our method achieved a classification performance not inferior to equivalent approaches applied in different sports scenarios, including simple tasks, such as walking or running, and even more complex exercise including fitness training. Another potential limit lies in the labeling procedures, which relied in the use of footage from a single 50 Hz camera. A single plane of view for four distinguished movements could make establishing their exact start and finish time more difficult. Differences in sampling rates between different systems could also add minor discrepancy in time line reconstruction. Finally, a potential bias to the assessment of classification performance could be the disproportion between the periods of transition and of functional movement execution, with the former being an order of magnitude more numerous (>3,000 transition windows vs. ~100–200 windows per each functional movement). Arguably, in our application, transitions are not static, easily detectable situations, and rather contain a spectrum of movement features that are as or even more varied than the four movements of interest. Thus, high prevalence of transition intervals should not decrease the value of the solution proposed.

CONCLUSIONS

Our study addressed a novel issue in the area of automatic activity tracking. We used wearable sensor data of the same kind of what could be provided by modern smart technologies and obtained from body locations similar to where those technologies could be secured. Classifying functional fitness movements within a continuous workout is a non-trivial task that, to the best of our knowledge, no other research had investigated. Despite the relatively small dataset used to train the algorithm, the accuracy achieved in detecting and recognizing four popular training drills was encouraging, even considering a simpler 1-sensor or 2-sensor configuration. Reducing input data to accelerations and angular velocities provided by a single sensor did not degrade excessively the classification ability of the algorithm, which still generated an overall level of accuracy similar to what obtained from the whole dataset available. These findings are particularly interesting as commercially available devices such as smart watches and/or phones contain inertial sensors and are typically worn in similar locations (i.e., upper arm and wrist) to where IMUs were attached in our study. This work perfectly fits the current technological trend on the combined use of wearable devices and artificial intelligence to track human activities automatically (Attal et al., 2015) and support sports activities (Cust et al., 2019). In the longer perspective, the proposed approach could drive the development of software and applications to aid on-field coaching and judging and provide a more objective, quantitative way to evaluate movement technique and correct/safe execution of specific drills.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the study was approved by the Research Ethics Approval Committee for Health (REACH) of the University of Bath, with reference number EP17/18 247. The patients/participants provided their written informed consent to participate in this study.

CONSENT STATEMENT

Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Original conceptualization was designed by EP, SN, and NL. SN and EP defined the setup. SN performed the acquisition and data analysis. EP, SN, and NL contributed to the original draft preparation. EP and NL finalized the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Adelsberger, R., and Troster, G. (2013). "Experts lift differently: classification of weight-lifting athletes," in *2013 IEEE International Conference on Body Sensor Networks* (Cambridge, MA: IEEE), 1–6. doi: 10.1109/BSN.2013.6575458

Adesida, Y., Papi, E., and McGregor, A. H. (2019). Exploring the role of wearable technology in sport kinematics and kinetics: a systematic review. *Sensors (Switzerland)* 19:1597. doi: 10.3390/s19071597

Ahmad, M., Khan, A., Mazzara, M., and Distefano, S. (2017). Seeking optimum system settings for physical activity recognition on smartwatches. *arXiv1706.01720* 944, 220–233. doi: 10.1007/978-3-030-17798-0_19

Arifoglu, D., and Bouchachia, A. (2017). Activity recognition and abnormal behaviour detection with recurrent neural networks. *Proc. Comput. Sci.* 110, 86–93. doi: 10.1016/j.procs.2017.06.121

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors (Switzerland)* 15, 31314–31338. doi: 10.3390/s151229858

Bachmann, D., Weichert, F., and Rinkenauer, G. (2018). Review of three-dimensional human-computer interaction with focus on the leap motion controller. *Sensors* 18:2194. doi: 10.3390/s18072194

Barbieri, J. F., Figueiredo, G. T., da, C., Castano, L. A. A., Guimaraes, P. dos, S., et al. (2019). A comparison of cardiorespiratory responses between crossfit® practitioners and recreationally trained individual. *J. Phys. Educ. Sport* 19, 1606–1611. doi: 10.7752/jpes.2019.03233

Camomilla, V., Bergamini, E., Fantozzi, S., and Vannozzi, G. (2018). Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review. *Sensors (Switzerland)* 18:873. doi: 10.3390/s18030873

Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K., and Buckles, B. P. (2015). *Advances in Human Action Recognition: a Survey*. Available online at: http://arxiv.org/abs/1501.05964.

Chetty, G., and White, M. (2016). "Body sensor networks for human activity recognition," in *3rd International Conference on Signal Processing and Integrated Networks, SPIN 2016* (Noida, Delhi NCR: IEEE), 660–665. doi: 10.1109/SPIN.2016.7566779

Crema, C., Depari, A., Flammini, A., Sisinni, E., Haslwanter, T., and Salzmann, S. (2019). Characterization of a wearable system for automatic supervision of fitness exercises. *Measurement* 147:106810. doi: 10.1016/j.measurement.2019.07.038

CrossFit (2019). *Exercise Demos. Available online at: https://www.crossfit.com/exercisedemos/ (accessed March* 1, 2019).

Cust, E. E., Sweeting, A. J., Ball, K., and Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *J. Sports Sci.* 37, 568–600. doi: 10.1080/02640414.2018.1521769

Davila, J. C., Cretu, A. M., and Zaremba, M. (2017). Wearable sensor data classification for human activity recognition based on an iterative learning framework. *Sensors (Switzerland)* 17:1287. doi: 10.3390/s17061287

Endres, F., Hess, J., and Burgard, W. (2012). Graph-Based Action Models for Human Motion Classification. *Robot. 2012*, 1–6. Available online at: http://ais.informatik.uni-freiburg.de/publications/papers/endres12robotik.pdf; https://www.vde-verlag.de/proceedings-en/453418001.html

Fan, S., Jia, Y., and Jia, C. (2019). A feature selection and classification method for activity recognition based on an inertial sensing unit. *Information* 10:290. doi: 10.3390/info10100290

Ghazali, N. F., Shahar, N., Rahmad, N. A., Sufri, N. A. J., As'ari, M. A., and Latif, H. F. M. (2018). "Common sport activity recognition using inertial sensor," in *Proc. - 2018 IEEE 14th Int. Colloq. Signal Process. Its Appl. (CSPA)* (Batu Feringghi), 67–71. doi: 10.1109/CSPA.2018.8368687

Gianzina, E. A., and Kassotaki, O. A. (2019). The benefits and risks of the high-intensity crossfit training. *Sport Sci. Health* 15, 21–33. doi: 10.1007/s11332-018-0521-7

Hagenbuchner, M., Cliff, D. P., Trost, S. G., Van Tuc, N., and Peoples, G. E. (2015). Prediction of activity type in preschool children using machine learning techniques. *J. Sci. Med. Sport* 18, 426–431. doi: 10.1016/j.jsams.2014.06.003

Hausberger, P., Fernbach, A., and Kastner, W. (2016). "IMU-based smart fitness devices for weight training," in *IECON Proc. Industrial Electron. Conf.* (Florenz), 5182–5189. doi: 10.1109/IECON.2016.7793510

Hoettinger, H., Mally, F., and Sabo, A. (2016). Activity recognition in surfing-a comparative study between hidden markov model and support vector machine. *Procedia Eng.* 147, 912–917. doi: 10.1016/j.proeng.2016.06.279

Howell, D. R., Oldham, J. R., DiFabio, M., Vallabhajosula, S., Hall, E. E., Ketcham, C. J., et al. (2017). Single-task and dual-task gait among collegiate athletes of different sport classifications: implications for concussion management. *J. Appl. Biomech.* 33, 24–31. doi: 10.1123/jab.2015-0323

Hsu, Y. L., Yang, S. C., Chang, H. C., and Lai, H. C. (2018). Human daily and sport activity recognition using a wearable inertial sensor network. *IEEE Access* 6, 31715–31728. doi: 10.1109/ACCESS.2018.2839766

Jones, R. L., and Wallace, M. (2005). Another bad day at the training ground: coping with ambiguity in the coaching context. *Sport Educ. Soc.* 10, 119–134. doi: 10.1080/1357332052000308792

Liebenson, C. (2006). Functional training for performance enhancement—Part 1: the basics. *J. Bodyw. Mov. Ther.* 10, 154–158. doi: 10.1016/j.jbmt.2006.01.003

Mannini, A., and Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors* 10, 1154–1175. doi: 10.3390/s100201154

O'Reilly, M. A., Whelan, D. F., Ward, T. E., Delahunt, E., and Caulfield, B. (2017a). Classification of lunge biomechanics with multiple and individual inertial measurement units. *Sport. Biomech.* 16, 342–360. doi: 10.1080/14763141.2017.1314544

O'Reilly, M. A., Whelan, D. F., Ward, T. E., Delahunt, E., and Caulfield, B. (2017c). Technology in strength and conditioning tracking lower-limb exercises with wearable sensors. *J. Strength Cond. Res.* 31, 1726–1736. doi: 10.1519/JSC.0000000000001852

O'Reilly, M. A., Whelan, D. F., Ward, T. E., Delahunt, E., and Caulfield, B. M. (2017b). Classification of deadlift biomechanics with wearable inertial measurement units. *J. Biomech.* 58, 155–161. doi: 10.1016/j.jbiomech.2017.04.028

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Pernek, I., Kurillo, G., Stiglic, G., and Bajcsy, R. (2015). Recognizing the intensity of strength training exercises with wearable sensors. *J. Biomed. Inform.* 58, 145–155. doi: 10.1016/j.jbi.2015.09.020

Popoola, O. P., and Wang, K. (2012). Video-based abnormal human behavior recognition—a review. *IEEE Trans. Syst. Man, Cybern. C Appl. Rev.* 42, 865–878. doi: 10.1109/TSMCC.2011.2178594

Preatoni, E., Hamill, J., Harrison, A. J., Hayes, K., Van Emmerik, R. E. A., Wilson, C., et al. (2013). Movement variability and skills monitoring in sports. *Sport. Biomech.* 12, 69–92. doi: 10.1080/14763141.2012.738700

Quitadamo, L. R., Cavrini, F., Sbernini, L., Riillo, F., Bianchi, L., Seri, S., et al. (2017). Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: a review. *J. Neural Eng.* 14:011001. doi: 10.1088/1741-2552/14/1/011001

Rawashdeh, S. A., Rafeldt, D. A., and Uhl, T. L. (2016). Wearable IMU for shoulder injury prevention in overhead sports. *Sensors (Basel).* 16:1847. doi: 10.3390/s16111847

Sarig Bahat, H., Takasaki, H., Chen, X., Bet-Or, Y., and Treleaven, J. (2015). Cervical kinematic training with and without interactive VR training for chronic neck pain - a randomized clinical trial. *Man. Ther.* 20, 68–78. doi: 10.1016/j.math.2014.06.008

Singh, S., and Saini, H. K. (2019). Effect of ten weeks of crossfit training on 'Yo-Yo test' perfor-mance of cricketers. *Int. J. Recent Technol. Eng.* 8, 1458–1460. doi: 10.35940/ijrte.B1269.0782S319

Taha, Z., Musa, R. M., Abdul Majeed, A. P. P., Alim, M. M., and Abdullah, M. R. (2018). The identification of high potential archers based on fitness and motor ability variables: a Support vector machine approach. *Hum. Mov. Sci.* 57, 184–193. doi: 10.1016/j.humov.2017.12.008

van der Kruk, E., and Reijne, M. M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. *Eur. J. Sport Sci.* 18, 806–819. doi: 10.1080/17461391.2018.1463397

Wang, Y., Cang, S., and Yu, H. (2016). "Realization of wearable sensors-based human activity recognition with an augmented feature group," in *22nd*

International Conference on Automation and Computing (ICAC) (Colchester, UK), 473–478. doi: 10.1109/IConAC.2016.7604965

Willetts, M., Hollowell, S., and Aslett, L. (2018). Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci. Rep.* 8:7961. doi: 10.1038/s41598-018-2 6174-1

WODstar (2019). *Wodstar Video Movement Index.* Available online at: https:// wodstar.com/ (accessed March 1, 2019).

Zhang, M., and Sawchuk, A. A. (2013). Human daily activity recognition with sparse representation using wearable sensors. *IEEE J. Biomed. Heal. Informatics* 17, 553–560. doi: 10.1109/JBHI.2013.2253613

Hand Resting Tremor Assessment of Healthy and Patients With Parkinson's Disease: An Exploratory Machine Learning Study

Ana Camila Alves de Araújo[1], Enzo Gabriel da Rocha Santos[2], Karina Santos Guedes de Sá[3], Viviane Kharine Teixeira Furtado[4], Felipe Augusto Santos[3], Ramon Costa de Lima[5], Lane Viana Krejcová[6], Bruno Lopes Santos-Lobato[3], Gustavo Henrique Lima Pinto[2], André dos Santos Cabral[7], Anderson Belgamo[8], Bianca Callegari[3], Ana Francisca Rozin Kleiner[9,10], Anselmo de Athayde Costa e Silva[3] and Givago da Silva Souza[4,5]*

[1] Núcleo de Teoria e Pesquisa do Comportamento, Universidade Federal do Pará, Belém, Brazil, [2] Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, Brazil, [3] Instituto de Ciências da Saúde, Universidade Federal do Pará, Belém, Brazil, [4] Núcleo de Medicina Tropical, Universidade Federal do Pará, Belém, Brazil, [5] Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil, [6] Instituto de Ciências da Arte, Universidade Federal do Pará, Belém, Brazil, [7] Centro de Ciências Biológicas e da Saúde, Universidade do Estado do Pará, Belém, Brazil, [8] Departamento de Ciência da Computação, Instituto Federal de São Paulo, Piracicaba, Brazil, [9] Laboratório Rainha Sílvia de Análise do Movimento, Rio Claro, Brazil, [10] Departamento de Fisioterapia, Universidade Federal de São Carlos, São Carlos, Brazil

The aim of this study is comparing the accuracies of machine learning algorithms to classify data concerning healthy subjects and patients with Parkinson's Disease (PD), toward different time window lengths and a number of features. Thirty-two healthy subjects and eighteen patients with PD took part on this study. The study obtained inertial recordings by using an accelerometer and a gyroscope assessing both hands of the subjects during hand resting state. We extracted time and temporal frequency domain features to feed seven machine learning algorithms: k-nearest-neighbors (kNN); logistic regression; support vector classifier (SVC); linear discriminant analysis; random forest; decision tree; and gaussian Naïve Bayes. The accuracy of the classifiers was compared using different numbers of extracted features (i.e., 272, 190, 136, 82, and 27) from different time window lengths (i.e., 1, 5, 10, and 15 s). The inertial recordings were characterized by oscillatory waveforms that, especially in patients with PD, peaked in a frequency range between 3 and 8 Hz. Outcomes showed that the most important features were the mean frequency, linear prediction coefficients, power ratio, power density skew, and kurtosis. We observed that accuracies calculated in the testing phase were higher than in the training phase. Comparing the testing accuracies, we found significant interactions among time window length and the type of classifier ($p < 0.05$). The study found significant effects on estimated accuracies, according to their type of algorithm, time window length, and their interaction. kNN presented the highest accuracy, while SVC showed the worst results. kNN feeding by features extracted from 1 and 5 s were the combination with more frequently highest accuracies. Classification

using few features led to similar decision of the algorithms. Moreover, performance increased significantly according to the number of features used, reaching a plateau around 136. Finally, the results of this study suggested that *k*NN was the best algorithm to classify hand resting tremor in patients with PD.

INTRODUCTION

More than 6.1 million people worldwide are affected by Parkinson's disease (PD) (Gbd 2016 Parkinson's Disease Collaborators, 2018) – this number is expected to rise with the increasing of the population life expectancy (Wanneveich et al., 2018). PD has very heterogeneous clinical features, but tremor at rest, akinesia, and rigidity are considered the clinical cardinal motor signatures of this disease (Kalia and Lang, 2015; Poewe et al., 2017). It is hard to diagnose PD, both in its early stages and during its progression. Its diagnosis is usually carried out by clinical observation or by using scales such as the Unified Parkinson's Disease Rating Scale (UPDRS) or the Hoehn and Yahr scale (H-Y) (Hoehn and Yahr, 1967; Rizek et al., 2016; Holden et al., 2018).

Literature has proposed alternative ways to quantify PD symptoms in order to assist its diagnosis and progression (Jilbab et al., 2017). Inertial measures of the hand resting tremor associated to machine learning algorithms have been extensively investigated to distinct data from healthy people and patients with PD (Jeon et al., 2017a,b), to quantify the progression of the disease (Pedrosa et al., 2018), and to evaluate the effect of therapeutics on hands' tremor (LeMoyne et al., 2019).

Although many investigations have evaluated the machine learning classifier performance to precisely categorize the inertial measurements from patients with PD, there are few methodological studies concerning the influence of the technical parameters of this kind of approach. Parameters like the time interval of the inertial sensor readings, type of features extracted from the inertial sensor readings, the number of features used, the type of machine learning classifier, and the type of inertial sensor used have potential to increase or decrease the accuracy of the algorithm (Jeon et al., 2017a; Rovini et al., 2017; Ramdhani et al., 2018; Wang et al., 2018; Nurwulan and Jiang, 2020). **Table 1** lists examples of studies that associated inertial measurements with machine learning approaches and their methodological choices. It displays a large variability of methodological settings and few explanations justifying such choices.

Several investigations have used a number of machine learning algorithms to classify and/or to quantify the resting hand tremor of patients with PD, obtaining high accuracy levels (Kostikis et al., 2015: 78–94%; Jeon et al., 2017a: 80–85%; Pedrosa et al., 2018: 92.8%). There is no consensus about what machine learning algorithms are preferable to classify features of inertial readings or what are the optimal conditions to use any of the algorithms.

Several studies have segmented inertial recordings in different window size durations to extract dozens or hundreds of features

that fed a machine learning algorithm (Jeon et al., 2017a). Short-term inertial readings could be good to get a fast evaluation, but they lead to high false positive detection. On the other hand, long-term recordings may potentially prolong the recording process, adding redundant information (Nurwulan and Jiang, 2020). In the same way, using a few features may not be enough to bring clear information about the differences among patients with PD; and an excessive number of features may overload the computing process. It is important to select the best set of features in order to potentialize algorithm classification and to avoid collinearity among data.

The present study aimed to compare the performance of machine learning algorithms to classify recordings of inertial sensors as healthy people or patients with PD considering different numbers of features extracted from a variety of window length duration of inertial recordings. Those results may contribute in the decision making of the best parameter for the classification of inertial sensor measures analyzed by machine learning algorithms.

MATERIALS AND METHODS

Ethical Considerations

All individual participants included in this study gave us their informed and written consent. Every procedure carried out in the present study was in accordance with the ethical standards of the Ethics Committee in Research with Humans from the University Hospital João de Barros Barreto (report #1.338.241) and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Subjects

Our sample comprised of 50 right-handed participants grouped into healthy control participants (*n* = 32 individuals, 16 females and 16 males) and participants with PD (*n* = 18 individuals, 8 females and 10 males). Participants' handedness was established according to the hand they use to handwrite. Healthy participants ranged from 41 to 79 years (mean ± standard deviation: 64.3 ± 11.1 years), while patients with PD ranged from 48 to 73 years (mean ± standard deviation: 60.2 ± 8.4 years). Control participants were recruited by convenience. They had no history of neurological or systemic diseases, no self-reported tremor of the hands nor difficulties in carrying out daily activities. All patients with PD were diagnosed by a neurologist in the Neurology Department of the University Hospital João de Barros Barreto, Brazil, according to the clinical diagnostic criteria of the United Kingdom Parkinson's Disease Society Brain Bank

TABLE 1 | References that used inertial sensors features to feed machine learning to evaluate the hand tremor of PD patients.

References	Hand activity	Sensor (AR)	Recording duration	Methods of classification	Accuracy
Alam et al. (2016)	Resting tremor	Acc and gyros (200 Hz)	25–30 s	Support vector machine	59–88.9%
LeMoyne et al. (2015)	Kinetic tremor	Acc (100 Hz)	5 s	Support vector machine	100%
Butt et al. (2017)	Kinetic tremor	Gyros (100 Hz)	10 s	Support vector machine, logistic regression, neural network classifier	76.2–83.1%
Stamatakis et al. (2013)	Finger tapping	Acc (167 Hz)	Free	Ordinal logistic regression	87.2–96.5%
Jeon et al. (2017a)	Resting tremor	Acc (125 Hz)	10 s	SVM, decision tree, random forest, discriminant analysis	80.9–85.6

(Hughes et al., 1992). For each patient, the severity of PD was scored by using the Hoehn and Yahr (H-Y) scale. All patients with PD had disease diagnosed within the less 6 years; except by one subject (H-Y 3), all other patients were staged as functionally independent (H-Y 1 or 2). All patients were using levodopa or dopamine agonist therapy for over a year.

Inertial Measurement Unit Recordings

We used a wearable device MetaMotionC (mbientlab, San Francisco, United States), with on-board sensors, such as a triple-axis gyroscope and an accelerometer (16 bits, $\pm$ 2000°/s, $\pm$ 16 g). Researchers positioned a wearable device over each patient's third metacarpal bone at their midway between the carpal and the digital extremities of their metacarpal (**Figure 1**) – with their forearm supported on a table, and their hand relaxed over its edge. Researchers recorded the patients in resting state with the acquisition rate at 100 Hz and 16-bit analog to digital conversion resolution. An Android app (MetaBase, mbientlab, United States) controlled the sensors via Bluetooth. Bluetooth also transmitted their signals to an ordinary computer. The study delivered 2-min recordings. One trial was carried out for each one of the hands of all participants.

Data Analysis

To carry out data analysis, researchers programmed Python scripts (Python v3.7.4) by using SciPy (version 1.3.1), NumPy (version 1.17.2), PyWavelets (version 1.0.3), and LibROSA (version 0.7.2) tools. SciPy is a Python-based ecosystem of open-source software for mathematics, science, and engineering; NumPy is a library for the Python programming used to operate on arrays; LibROSA is a Python package that provides the building blocks necessary to create music information retrieval systems; and PyWavelets is an open source wavelet that transforms software for Python.

Our sequence of analysis consisted of: (1) inertial recordings; (2) raw data filtering; (3) segmentation of the time series in different sets of waveform lengths; (4) data normalization; (5) extraction of features; (6) selection of the best features; (7– 8) performing machine learning algorithms with training and test phases; and (9) measuring machine learning performance. **Figure 2** illustrates data analysis summary.

Raw Data Filtering

We computed a magnitude vector from each sensor dimension (x, y, and z) using Eq. (1), which is less sensitive to orientation

changes (Janidarmian et al., 2017). The recordings were filtered by a fourth-order bandpass digital Butterworth filter between 1 and 30 Hz to exclude low and high frequency artifacts.

$$v = \sqrt{x^2 + y^2 + z^2} \tag{1}$$

where v is the magnitude vector, x, y, and z represented the 3-D readings of the inertial sensor.

After this, we applied the *scipy.signal.detrend* function using its linear list squared fit to detrend the inertial readings.

Segmentation of the Time Series

We segmented the inertial recordings in fixed sized windows, with no inter-window gaps and non-overlapping between adjacent windows. We also segmented these time series in sets of waveforms with 1, 5, 10, and 15 s window sizes.

Extraction of Features

We extracted features from time and temporal domains for each sensor dimension. **Table 2** presents a list of features extracted from inertial data, as well as Python main codes related to them.

The study extracted 272 features from each one of our participants, considering data extracted: (a) from each one of their hands (dominant and non-dominant); (b) from each inertial sensor parameter (accelerometer and gyroscope); and, (c) from the four dimensions of each sensor (x, y, z, and magnitude).

Data Normalization

The study applied *sklearn.preprocessing* package and its *StandardScaler* function to standardize features by removing their mean and scaling them to unit variance, as shown in Eq. (2).

$$z\_score = \frac{(x - \mu)}{s} \tag{2}$$

Selection of Features

The study used algorithm *SelectKBest* to select the k most important features based in a score which was the ANOVA *F*-value. The chosen selection of the most important features to feed the machine learning algorithms in this study where: 272 features (100%), 190 features (70%), 136 features (50%), 82 features (30%), and 27 features (10%).

Splitting Data

To validate the predictive models, we applied the tenfold cross-validation method by using the *Scikit-learn* library (version
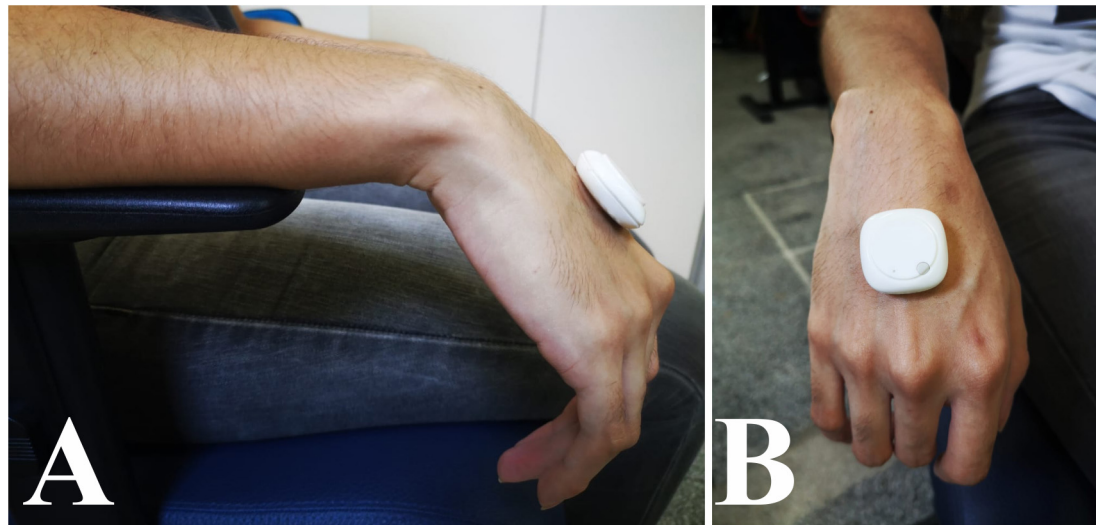
FIGURE 1 | IMU Positioning in the hand of the participant. **(A)** Lateral view. **(B)** Frontal view. The patient was instructed to keep the hand in rest for 120 s, while the experimenter controlled the recording using a mobile app.
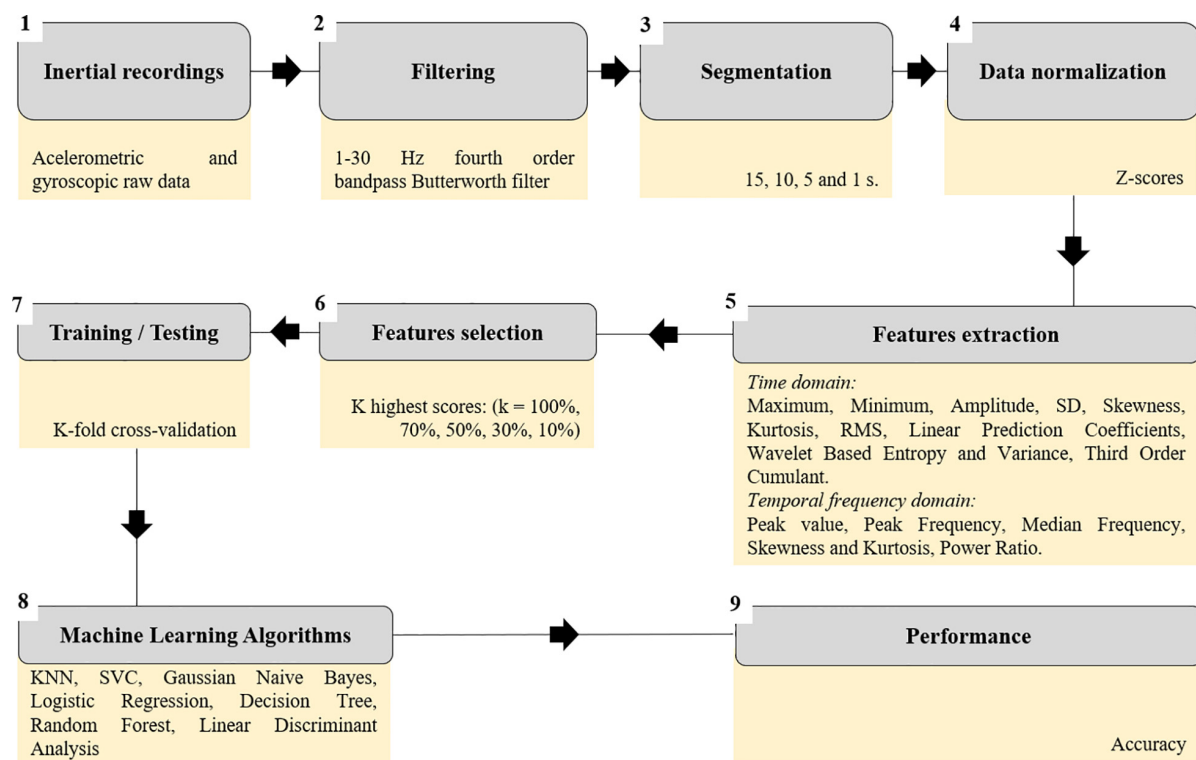
FIGURE 2 | Flow chart of the data analysis steps.

0.21.3) and *ShuffleSplit* function. The study randomly split data into 80% for model training and 20% for model testing.

Machine Learning Algorithms

We applied seven types of machine learning algorithms to classify the data from both healthy and PD groups. The algorithms were: *k*-nearest-neighbor (*k*NN); support vector classifier (SVC); logistic regression (LR); linear discriminant analysis (LDA); random forest (RF); decision tree (DT); and Gaussian Naïve Bayes (GNB).

The next sentences describe the Python functions used to proceed the machine learning algorithms, as well as the

TABLE 2 | Features extracted from the inertial readings.

Features	Python code
Time domain	
Range	range = values.max() - values.min()
Standard deviation	std = values.std()
Root mean square	rms = numpy.sqrt(numpy.mean(values**2))
Skewness	sk = scipy.stats.skew(values)
Kurtosis	kt = scipy.stats.kurtosis(values)
Linear prediction coefficients	lp_coefs = librosa.lpc(values, 3)
Wavelet transform detail coefficients (cD)	_, cD = pywt.dwt(values, 'db3')
cD variance	variance = numpy.var(cD)
cD entropy	def approximate_entropy(U, m = 2, r = 3):
	U = numpy.array(U)
	N = U.shape[0]
	def phi(m):
	z = N - m + 1.0
	x = numpy.array([U[i:i + m] \
	for i in range(int(z))])
	x_ = numpy.repeat(x[:, \
	numpy.newaxis], 1, axis = 2)
	C = numpy.sum(numpy.absolute(x - \
	x_).max(axis = 2) < = r, \
	axis = 0)/z
	return numpy.log(C).sum()/z
	entropy = abs(phi(m + 1) - phi(m))
Third order cumulant	third_order_cum = scipy.stats.moment(values, moment = 3)
Temporal frequency (tf) domain	
Peak of energy	p_tf = frequency_values.max()
Frequency at the peak energy	xf = numpy.linspace(0, af/2,
	frequency_values.size)
	tf_p = xf[numpy.argmax(frequency_values)]
Skewness_tf	sk_tf = scipy.stats.skew(frequency_values)
Kurtosis_tf	kt_tf = scipy.stats.kurtosis(frequency_values)
Mean frequency	def mean_frequency(frequency_values):
	xf = numpy.linspace(0, af/2,
	frequency_values.size)
	xf = xf[xf > = 1]
	total_area = numpy.trapz(frequency_values, xf)
	for i, x in enumerate(xf):
	partial_area = numpy.trapz(frequency_values[:i], xf[:i])
	if partial_area > total_area/2:
	mean_freq = xf[i-1]
Power ratio (1–6 Hz/6–12 Hz)	xf = numpy.linspace(0, af/2,
	frequency_values.size)
	num = frequency_values[(xf > = 1) &
	(xf < = 6)]
	den = frequency_values[(xf > = 6) &
	(xf < = 12)]
	power_ratio = num.mean()/den.mean()

values, inertial measures in the time domain vector; frequency_values, inertial measures in the temporal frequency domain vector; af, the acquisition frequency; and, xf, frequency values vector.

parameters that differed from default values. These parameters were changed to protect the model from overfitting.

(a) *k*-Nearest-Neighbor (*k*NN): the function *sklearn.neighbors.KNeighborsClassifier* was applied to proceed an *k*NN algorithm considering the Minkowski distance metrics, *k*-value ranging from 5 to 10. We applied a grid search using the *GridSearchCV* function to find which *k*-nearest-neighbor would deliver the best accuracy, then chosen as the best *k*-value.

(b) Support Vector Classifier (SVC): were applied an SVC algorithm (*sklearn.svm.SVC* function) with radial basis function kernel with *gamma* parameter equal to 1 and the *C penalty* parameter equal to 10.

(c) Logistic Regression (LR): a binary logistic regression algorithm *sklearn.linear_model.LogisticRegression* function was used considering the parameter *penalty* equal to "l1," and *solver* equal to "liblinear."

(d) Linear Discriminant Analysis (LDA): the study applied the function *sklearn.discriminant_analysis.LinearDiscriminantAnalysis* to proceed the LDA algorithm considering the parameter *solver* equal to "svd," and *store_covariance* as true.

(e) Random Forest (RF): we used the function *sklearn.ensemble.RandomForestClassifier* to implement random forest algorithm considering the parameter "*criterion*" the value "*gini impurity*" as a measure of the split quality, the parameters *n_estimators* equal to 50, and *max_depth* equal to 6.

(f) Decision Tree (DT): similarly to the random forest classifiers, the tree algorithm was proceed using the *sklearn.tree.DecisionTreeClassifier* function considering "*gini impurity*" to the parameter "*criterion*," and the parameters *n_estimators* were set to 50, and *max_depth* equal to 6.

(g) Gaussian Naïve Bayes (GNB): the function to proceed a Gaussian Naïve Bayes algorithm was the *sklearn.naive_bayes.GaussianNB*.

Measuring Machine Learning Performances

Equation (3) calculated accuracy in order to measure the success levels of the classifiers, as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (3)$$

where TP is the true positive value; TN is the true negative value; FP is the false positive value; and, FN is the false negative value.

Statistics

The study applied the unpaired *t*-test with Welch's correction to compare the accuracies obtained from training and testing phases for each classifier using features extracted from different time window lengths. For each percentage of features feeding the algorithms, we conducted a two-way ANOVA on the influence of the classifier type and the time window length of the accuracy of such classifier. The classifier type includes seven levels (SVC, GNB, RF, *k*NN, LR, LDA, and DT) and the time

window length consisted of five levels (1, 5, 10, and 15 s). As the two-way ANOVA test was significant, we computed the Tukey HSD for performing multiple pairwise-comparison between mean accuracies of both groups. We counted the number of times in which an algorithm presented a better performance when compared to the others (here named victory), by means of significant multiple comparisons at the different time window lengths and number of features. Thus, we used the chi-square goodness of fit (equal proportions) to compare the observed distribution of significant comparisons to the expected distribution considering the number of algorithms or of time window length. All the statistical tests were carried out by using R software (version 3.6) and considering the level of significance of 5%.

RESULTS

Selection of Recordings and Features

Figure 3 shows examples of the accelerometric and gyroscopic recordings for the 5-s time windows as a function of time and temporal frequency from representative subjects from both groups. The results for the 5 s time windows were qualitatively similar to the other time windows the study investigated. We characterized the inertial recordings by oscillatory waveforms that, especially in participants with PD, defined their peak in frequencies ranging between 3 and 8 Hz.

Regardless time window length, the most important features detected were mean frequency, linear prediction coefficients, power ratio, and the power density skew and kurtosis. **Figure 4** shows the 15 most important features selected from extracted data concerning time windows of 15 s (**Figure 4A**), 10 s (**Figure 4B**), 5 s (**Figure 4C**), and 1 s (**Figure 4D**).

Machine Learning Classifiers
Comparison Between Training and Testing Accuracies

Most of the comparisons had significant differences between training and testing phases. Whenever statistical significance ($p < 0.05$) was reached, testing accuracy was higher than training accuracy – except in two comparisons (random forest and kNN algorithms) – when using 30% of the features in the 1 s time window. **Supplementary Files 1–5** present tables with the training and testing phases of the machine learning.

The comparisons with no statistical significance were in time windows of:

(i) 1 s: random forest algorithm using all features and 70% of them, GNB using 50 and 10%;
(ii) 5 s: GNB with all features, 70 and 50% of them, kNN and LR using 30% of the features;
(iii) 10 s: GNB using 30 and 10% of the features;
(iv) 15 s: GNB using all features, 70, 50, and 10% of them, SVC using all features, 70 and 50% of them, LDA using all features and 70% of them, LR using 50% of the features, and RF using 30% of the features.

Figure 5 illustrates the comparisons between the accuracies obtained by the different classifiers using extracted features in different time windows considering 70, 50, 30, and 10% of the features, respectively.
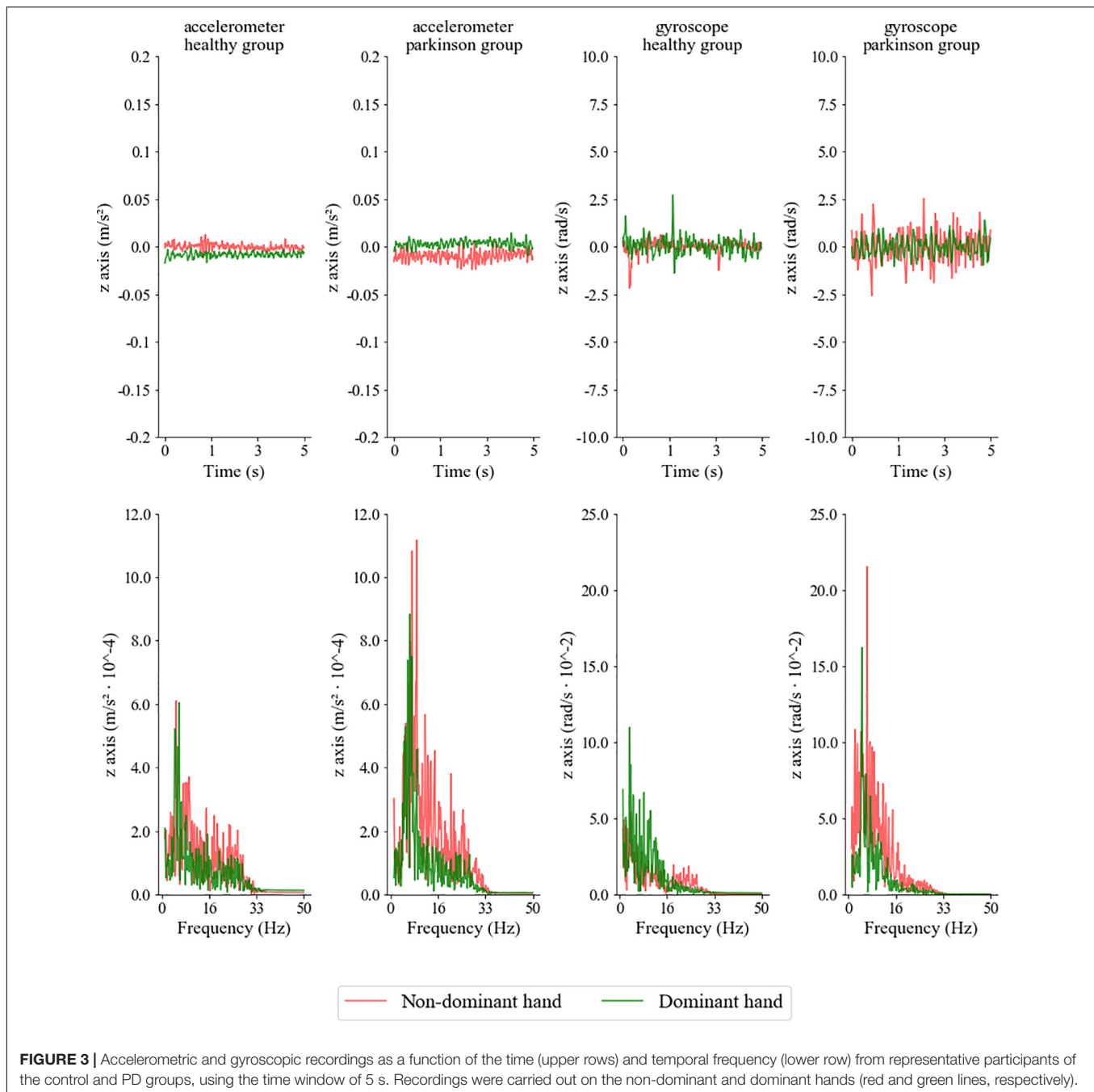
Comparing Test Accuracies Obtained From the Different Supervised Machine Learning Algorithms

In general, the effects of the machine learning phases on the accuracies were statistically significant. The main effect for classifier type yielded an F ratio of $F(6, 252) = 639.14$, $p < 0.0001$ for all the features; $F(6, 252) = 727.74$, $p < 0.0001$ for 70% of the features; $F(6, 252) = 478.15$, $p < 0.0001$ for 50% of the features; $F(6, 252) = 171.41$, $p < 0.0001$ for 30% of the features; and $F(6, 252) = 36.8$, $p < 0.0001$ for 10% of the features. The proportion of victories in the multiple comparisons significantly differed by algorithm for all numbers of features conditions. kNN was the algorithm that more frequently delivered high accuracy when compared to the others algorithms. SVC delivered the lowest frequency of victories among all tested algorithms. **Table 3** shows the number of "victories" of each algorithm in the significant multiple comparisons for each number of feature condition.

The main effect for time window length yielded an F ratio of $F(3, 252) = 51.7$, $p < 0.0001$ for all the features; $F(3, 252) = 47.4$, $p < 0.0001$ for 70% of the features; $F(3, 252) = 25.5$, $p < 0.0001$ for 50% of the features; $F(3, 252) = 5.5$, $p < 0.0001$ for 30% of the features; and $F(3, 252) = 14.8$, $p < 0.0001$ for 10% of the features. The proportion of victories in the multiple comparisons was similar by time window length for all numbers of feature conditions, except for 10% of the features. **Table 4** displays the number of "victories" from time window length in the significant multiple comparisons for each number of feature condition.

The interaction effect was significant for all numbers of features conditions (for all the features: $F(18, 252) = 19.04$, $p < 0.001$; for 70% of the features: $F(18, 252) = 15.23$, $p < 0.001$; For 50% of the features: $F(18, 252) = 7.61$, $p < 0.001$; and for 10% of the features: $F(18, 252) = 2.959$, $p < 0.001$), except for 30% of the features condition that yielded in a F ratio of $F(18, 252) = 2.959$, and $p = 0.29$.

Figures 6A–E shows tile plots representing the statistical significance of the *post hoc* multiple comparisons between the testing accuracies from any two classifiers. White tiles represent comparisons with significant differences, while dark tiles represent non-significant differences. The red line indicates the orientation of the significant difference. Horizontal lines represent higher accuracies for the classifiers in the row when compared to the classifiers in the column, while vertical lines represent the opposite situation. We observed that the number of significant differences between two classifiers (number of white tiles) was dependent of the number of features. For a low number of features (10% of the features we extracted, 27 features) the number of significant differences between two classifiers was also low and increased linearly up to reach a plateau level of 70% of the features (136 features). The combinations between classifier and time window length with highest accuracies were kNN and time windows of 1 and 5s.

FIGURE 3 | Accelerometric and gyroscopic recordings as a function of the time (upper rows) and temporal frequency (lower row) from representative participants of the control and PD groups, using the time window of 5 s. Recordings were carried out on the non-dominant and dominant hands (red and green lines, respectively).

DISCUSSION

This paper assessed the hand tremor in individuals with PD and healthy controls by using machine learning algorithms based on inertial sensor recordings. Our objectives were: (i) identifying the best machine learning algorithms to classify hand tremor by using inertial data; (ii) describing the best recording duration to be used by classification methods; (iii) stablishing the number of features necessary to the best performance of the algorithms.

Concerning these objectives, the results of this study showed that the *k*NN algorithm as the best classifier, followed by LR, and

RF algorithms respectively. On the other hand, research pointed out that SVC and GNB delivered the worst performances among all classifiers. Also, some classifiers had better performances with short time windows, while others needed long recordings to deliver more accurate performances. Our results also showed that the performance of the classifiers became more similar when using less features; and, with more features, differences between classifiers increased linearly until a maximum value (using around 136 features), reaching a plateau. Regardless the most important feature selected, the time window length was similar across tested conditions. Whereas, the more common
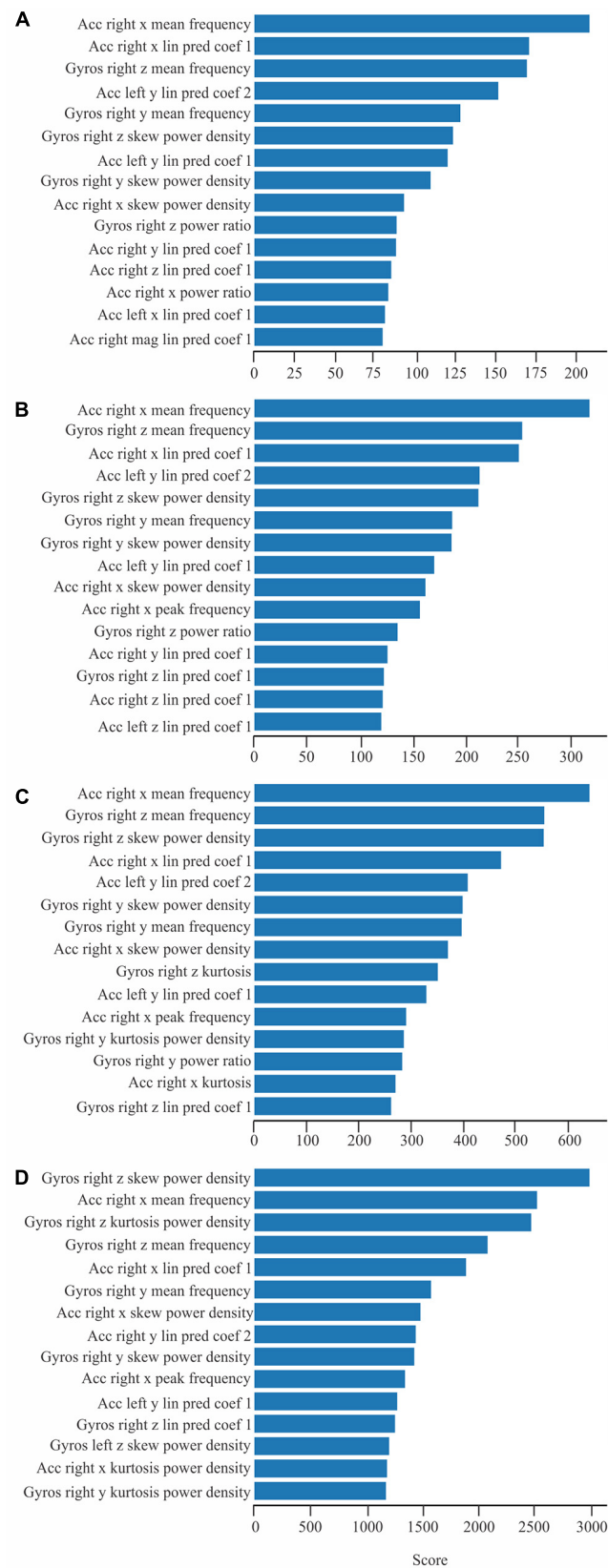
FIGURE 4 | Most important features extracted from recordings lasting 1 s **(A)**, 5 s **(B)**, 10 s **(C)**, and 15 s **(D)**.

FIGURE 5 | Comparison of the classifiers' performance in the training (solid bars) and testing (empty bars) phase according the number of features and time window length.

features selected were mean frequency for both accelerometer and gyroscope sensors; linear prediction coefficients for the accelerometer; skewness, power ratio, and the power density skewness and kurtosis for the gyroscope.

Many types of machine learning classifiers have been used to analyze PD tremor (Bind et al., 2015). We used 7 out of the most common algorithms used in the field. kNN was the best classifier across multiple comparisons, together with LR and RF algorithms, which had accuracy level above 90%.

The kNN algorithm groups similar classes of data based in the value of k nearest neighbors. Low values of k increase the accuracy of the classifier in the training phase, but difficult the generalization of the model for a new data (Li and Zhang, 2011). The k was used between 5 and 10 to facilitate the generalization of the model during test phase. Previous investigations – such as Jeon et al. (2017b) – have also found high accuracies using kNN algorithms. They assessed 85 PD patients to predict UPDRS results by using a wrist-watch-type wearable device for measuring

TABLE 3 | Number of victories of each classifier in the significant multiple comparisons for each number of feature condition.

	Number of features				
Algorithm	100%	70%	50%	30%	10%
SVC	5	5	3	0	4
GNB	12	16	16	13	2
RF	40	40	39	31	27
kNN	54	58	61	50	50
LR	53	48	41	31	6
LDA	34	38	35	27	3
DT	36	37	34	28	5
Number of significant multiple comparisons	234	242	229	180	97
X^2	63.53	57.72	63.50	57.38	142.51
P	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

TABLE 4 | Number of victories per time window length in the significant multiple comparisons for each number of feature condition.

	Number of features				
Time window length	100%	70%	50%	30%	10%
1 s	58	61	54	39	12
5 s	64	68	66	52	35
10 s	60	62	60	47	27
15 s	52	51	49	42	23
Number of significant multiple comparisons	234	242	229	180	97
X^2	1.28	2.46	2.84	2.17	11.33
P	0.73	0.48	0.51	0.53	<0.01

tremors and found an accuracy level close to 84% for kNN and RF algorithms. Also, kNN algorithm delivered performance improvement as we decreased the number of features, while other algorithms delivered impaired outcomes.

Random Forest is a combination of multiple tree predictors that make decisions based in random vectors of features. The RF decision is the more common decision of the collection of tree classifiers (Breiman, 2001). Previous studies have demonstrated the ability of RF models to detect freezing in the gait of patients with PD or the switching on and off state of deep brain stimulation in these patients (Tripoliti et al., 2013; Kuhner et al., 2017).

Logistic Regression is a classification algorithm that uses a logistic sigmoid function to transform observations in two or more classes. LeMoyne et al. (2019) used LR algorithms to distinguish inertial readings associated with on and off modes from deep brain stimulation in PD patients, getting an accuracy level of 95%.

Both GNB and SVC with the worst outcomes. When compared with other algorithms, the GNB classifier delivered lower (Susi et al., 2011) and higher (Bazgir et al., 2018) accuracies to detect human motion. GNB is an algorithm that evaluates the probability of events within different classes (Theodoridis et al., 2010; Bazgir et al., 2018). SVC aims to find an optimal separation hyperplane in order to minimize misclassifications (Vapnik, 1979). SVC has been widely used to detect tremor in PD patients. The accuracy level of its classifiers has ranged between 80 and 90% to quantify PD tremor (Alam et al., 2016; Jeon et al., 2017b). We used a radial compared to the best SVC used by Jeon et al. (2017b) finding similar results.

It is important to highlight that directly comparing the performance of the classifiers in different studies must be careful. Each study implements different parameters in the algorithms, which are not always fully described. Furthermore, the number and type of features may influence the classifier accuracies. The present study observed that few features make classifiers' decisions more similar, while an increased number of features enable the classifiers' performance to be distinguished, reaching a plateau around 176 features. One must find a trade-off between the number of features and the cost of computational processing for each algorithm especially when trying to implement such method with wearable or mobile devices.

The use of machine learning algorithms to recognize patterns of human motion requires the segmentation of motion recording time series. Previous studies have segmented time series in different lengths for pattern recognition tasks (Bussmann et al., 2001; Dehghani et al., 2019). Although, short lengths accelerate the duration of the recordings, their random nature can present negative influence on the classifiers' performance (Mannini et al., 2013). Short duration recordings in the scale of 100 ms have been successfully used to recognize human motion. At the same time, long-term recordings also returned high accuracy when detecting PD tremor as we can observe in **Table 1**.

This study evaluated the accuracy of classifiers by using different time window lengths. We observed that recordings lasting 5s or 1s delivered the highest accuracy levels. The study also noticed some interaction between the window time length and classifiers, indicating that some classifiers were better to analyze short recordings (i.e., kNN algorithm), while others showed higher accuracies when using long recordings (i.e., GNB). There is no rule concerning the length of inertial readings for the predictive modeling problem. Banos et al. (2014) investigated the effects of the windowing procedures on the activity recognition process using inertial data. They observed that intervals between 1 and 2 s offered the best trade-off between recognition speed and accuracy.

The more common features extracted from inertial readings express amplitude of oscillatory series, their spectral content, regularity, and coherence (Meigal et al., 2012; Twomey et al., 2018). The present study observed that mean frequency for both accelerometer and gyroscope sensors, linear prediction coefficients for the accelerometer, and skew power ratio, and the power density skew and kurtosis for the gyroscope frequently figure among the fifteen top features. Frequency domain features have been successfully employed in the machine learning algorithms by other researchers (Bazgir et al., 2018; Pedrosa et al., 2018).

We based our approach exclusively on accelerometer and gyroscope sensors, though other sensors are reported in the literature to quantify PD hand tremor using machine
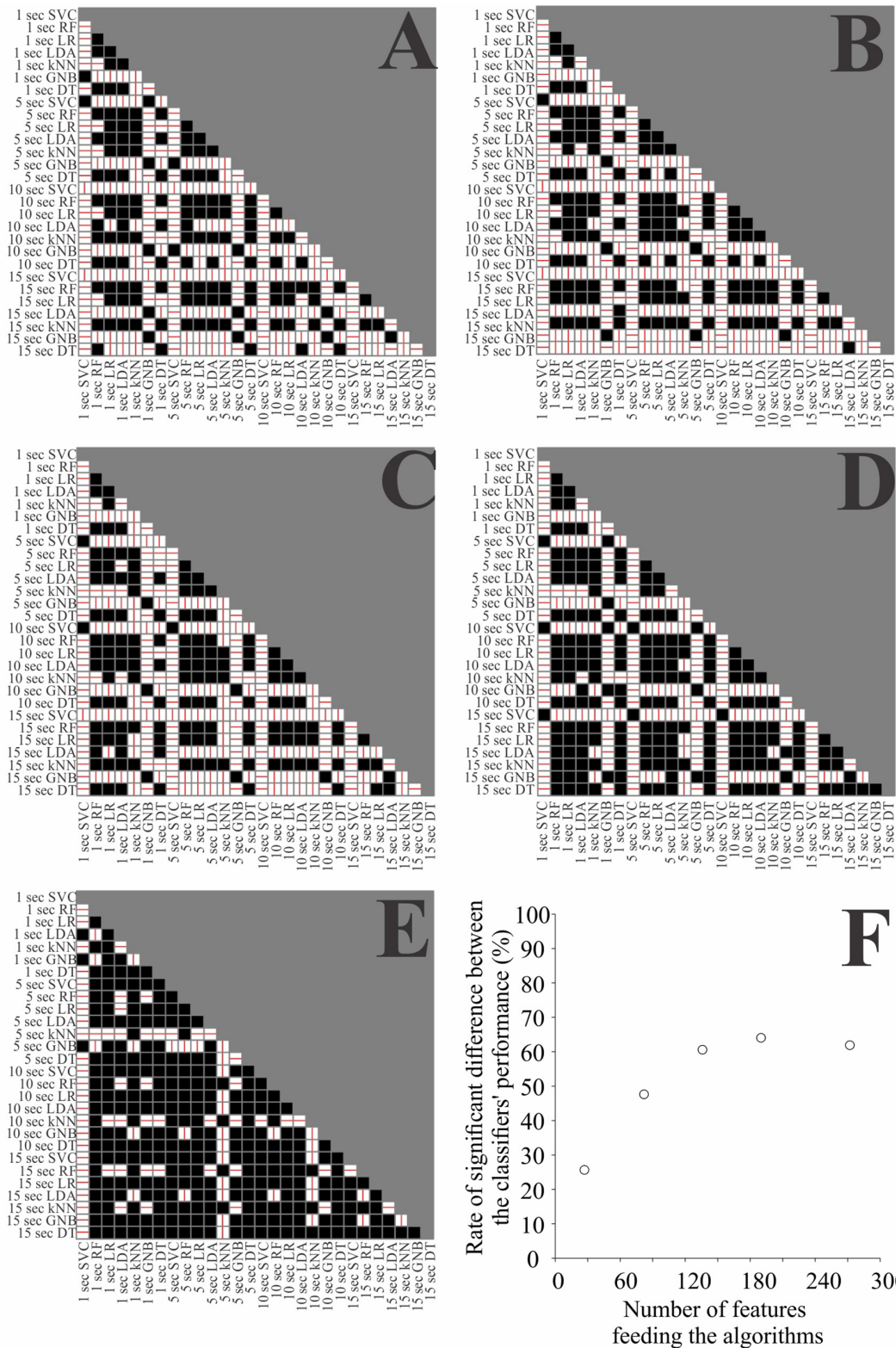
FIGURE 6 | Comparison of the classifier's performance in the testing phase when using all the features (A), 70% (B), 50% (C), 30% (D), and 10% (E) of the features. White squares represent the significant difference between the classifiers on the respective row and column, while black squares represent non-significance for the comparison. The line in the white squares represent the direction of the difference, horizontal lines indicates that the classifier on the row had higher accuracy than the classifier on the column, and vertical lines represent the opposite. (F) Number of significant differences between two classifiers as a function of number of features.

learning algorithms. For example, Lonini et al. (2018) used the MC10 BioStampRC sensor, a sensor tape that records electromyographic signals to accelerometers and gyroscopes in 6 body positions. Even considering that additional sensors can contribute to increase the accuracy of a classifier, there is a high cost in its implementation that can reduce the applicability of the proposal. Inertial sensors are inexpensive instruments that are available in a wide variety of wearable equipment.

This study has some potential limitations that deserve further comments. To date, research on this topic has been exploratory. There are no guidelines regarding the use of machine learning approach to quantify hand tremor in PD patients, as well as no established parameters for the choice of inertial sensors. A larger sample size and longitudinal follow-up could reinforce the present interpretations.

CONCLUSION

The present study suggested *k*NN using hundreds of features extracted from short-term inertial recordings as the best settings for machine learning configuration to classify hand tremor in PD patients. Our results can be used to assist the diagnosis and follow up of PD patients. We consider that our results are robust, because (i) of the high accuracy level obtained with the classifiers, (ii) the study could separate patients in the early stage of the PD (low H-Y score) from healthy people.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee in Research with Humans from the University Hospital João de Barros Barreto. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GS, AK, GP, and AAC conceived of the presented idea. ES and GP performed the computations. AA, KS, VF, FS, and RL collected the inertial recordings. LK and BS-L collected the clinical data. AA, ES, GS, AAC, and BC verified the analytical methods. ASC and AB contributed to the interpretation of the results. GS and AAC drafted the manuscript. All authors discussed the results and contributed to the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00778/full#supplementary-material

REFERENCES

Alam, M. N., Johnson, B., Gendreau, J., Tavakolian, K., Combs, C., and Fazel-Rezai, R. (2016). "Tremor quantification of Parkinson's disease - a pilot study," in *proceedings of the 2016 IEEE International Conference on Electro Information Technology (EIT)*. (Piscataway, NJ: IEEE) doi: 10.1109/eit.2016.7535334

Banos, O., Galvez, J. M., Damas, M., Pomares, H., and Rojas, I. (2014). Window size impact in human activity recognition. *Sensors (Basel)* 14, 6474-6499. doi: 10.3390/s140406474

Bazgir, O., Habibi, S., Palma, L., Pierleoni, P., and Nafees, S. (2018). A classification system for assessment and home monitoring of tremor in patients with Parkinson's Disease. *J. Med. Sci. Sensors* 8, 65–72.

Bind, S., Tiwari, A. K., and Sahani, A. K. (2015). A survey of machine learning based approaches for Parkinson disease prediction. *Int. J. Comp. Sci. Info. Techn.* 6, 1648–1655.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bussmann, J. B., Martens, W. L., Tulen, J. H., Schasfoort, F. C., van den Berg-Emons, H. J., and Stam, H. J. (2001). Measuring daily behavior using ambulatory accelerometry: the activity monitor. *Behav. Res. Methods Instrum. Comput.* 33, 349-356. doi: 10.3758/bf03195388

Butt, A. H., Rovini, E., Esposito, D., Rossi, G., Maremmani, C., and Cavallo, F. (2017). Biomechanical parameter assessment for classification of Parkinson's disease on clinical scale. *Int. J. Dist. Sensor Netw.* 13, 1-15. doi: 10.1177/1550147717707417

Dehghani, A., Sarbishei, O., Glatard, T., and Shihab, E. (2019). A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors* 19:5026. doi: 10.3390/s19225026

Gbd 2016 Parkinson's Disease Collaborators (2018). Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 17, 939-953. doi: 10.1016/S1474-4422(18)30295-3

Hoehn, M. M., and Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology* 17, 427-442. doi: 10.1212/wnl.17.5.427

Holden, S. K., Finseth, T., Sillau, S. H., and Berman, B. D. (2018). Progression of MDS-UPDRS scores over five years in de novo Parkinson disease from the Parkinson's progression markers initiative cohort. *Mov. Disord. Clin. Pract.* 5, 47-53. doi: 10.1002/mdc3.12553

Hughes, A. J., Daniel, S. E., Kilford, L., and Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinic-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiat.* 55, 181–184. doi: 10.1136/jnnp.55.3.181

Mannini, A., Intille, S. S., Rosenberger, M., Sabatini, A. M., and Haskell, W. (2013). Activity recognition using a single accelerometer placed at the wrist or ankle. *Med. Sci. Sports Exerc.* 45, 2193–2203. doi: 10.1249/MSS.0b013e31829736d6

Janidarmian, M., Roshan Fekr, A., Radecka, K., and Zilic, Z. (2017). A comprehensive analysis on wearable acceleration sensors in human activity recognition. *Sensors (Basel)* 17:529. doi: 10.3390/s17030529

Jeon, H., Lee, W., Park, H., Lee, H. J., Kim, S. K., Kim, H. B., et al. (2017a). Automatic classification of tremor severity in Parkinson's disease using a wearable device. *Sensors (Basel)* 17:2067. doi: 10.3390/s17092067

Jeon, H., Lee, W., Park, H., Lee, H. J., Kim, S. K., Kim, H. B., et al. (2017b). High-accuracy automatic classification of Parkinsonian tremor severity using machine learning method. *Physiol. Meas.* 38, 1980-1999. doi: 10.1088/1361-6579/aa8e1f

Jilbab, A., Benba, A., and Hammouch, A. (2017). Quantification system of Parkinson's disease. *Int. J. Speech Technol.* 20, 143–150. doi: 10.1007/s10772-016-9394-9

Kalia, L. V., and Lang, A. E. (2015). Parkinson's disease. *Lancet* 386, 896-912. doi: 10.1016/S0140-6736(14)61393-3

Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M., and Kotsavasiloglou, C. A. (2015). Smartphone-based tool for assessing parkinsonian hand tremor. *IEEE J. Biomed. Health. Inform.* 19, 1835-1842. doi: 10.1109/JBHI.2015.2471093

Kuhner, A., Schubert, T., Cenciarini, M., Wiesmeier, I. K., Coenen, V. A., Burgard, W., et al. (2017). Correlations between motor symptoms across different motor tasks, quantified via random forest feature classification in Parkinson's disease. *Front. Neurol.* 8:607. doi: 10.3389/fneur.2017.00607

LeMoyne, R., Tomycz, N., Mastroianni, T., McCandless, C., Cozza, M., and Peduto, D. (2015). Implementation of a smartphone wireless accelerometer platform for establishing deep brain stimulation treatment efficacy of essential tremor with machine learning. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2015, 6772-6775. doi: 10.1109/EMBC.2015.7319948

LeMoyne, R., Mastroianni, T., Whiting, D., and Tomycz, N. (2019). "Assessment of machine learning classification strategies for the differentiation of deep brain stimulation "On" and "Off" status for Parkinson's disease using a smartphone as a wearable and wireless inertial sensor for quantified feedback," in *Wearable Wireless Sys. Healthcare II* (New York, NY: Springer) 113–126. doi: 10.1007/978-981-13-5808-1-9

Li, Y., and Zhang, X. (2011). "Improving *k* nearest neighbor with exemplar generalization for imbalanced classification," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2011. Lecture Notes in Computer Science*, Vol. 6635, eds J. Z. Huang, L. Cao, and J. Srivastava (Heidelberg: Springer), 321–332.

Lonini, L., Dai, A., Shawen, N., Simuni, T., Poon, C., Shimanovic, L., et al. (2018). Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *NPJ Digit. Med.* 1:64. doi: 10.1038/s41746-018-0071-z

Meigal, A. Y., Rissanen, S. M., Tarvainen, M. P., Georgiadis, S. D., Karjalainen, P. A., Airakisinen, O., et al. (2012). Linear and nonlinear tremor acceleration characteristics in patients with Parkinson's disease. *Physiol. Meas.* 33, 395-412. doi: 10.1088/0967-3334/33/3/395

Nurwulan, N. R., and Jiang, B. C. (2020). Window selection impact in human activity recognition. *Int. J. Innov. Techn Interdisc. Sci.* 3, 381–394. doi: 10.15157/IJITIS.2020.3.1.381-394

Pedrosa, T. I., Vasconcelos, F. F., Medeiros, L., and Silva, L. D. (2018). Machine learning application to quantify the tremor level for parkinson's disease patients. *Procedia Comp. Sci.* 138, 215–220. doi: 10.1016/j.procs.2018.10.031

Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., et al. (2017). Parkinson disease. *Nat. Rev. Dis. Primers* 3:17013. doi: 10.1038/nrdp.2017.13

Ramdhani, R. A., Khojandi, A., Shylo, O., and Kopell, B. H. (2018). Optimizing clinical assessments in Parkinson's disease through the use of wearable sensors and data driven modeling. *Front. Comput. Neurosci.* 12:72. doi: 10.3389/fncom.2018.00072

Rizek, P., Kumar, N., and Jog, M. S. (2016). An update on the diagnosis and treatment of Parkinson disease. *CMAJ* 188, 1157-1165. doi: 10.1503/cmaj.151179

Rovini, E., Maremmani, C., and Cavallo, F. (2017). How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. *Front. Neurosci.* 11:555. doi: 10.3389/fnins.2017.00555

Stamatakis, J., Ambroise, J., Crémers, J., Sharei, H., Delvaux, V., Macq, B., et al. (2013). Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers. *Comput. Intell. Neurosci.* 2013, 717853. doi: 10.1155/2013/717853

Susi, M., Borio, D., and Lachapelle, G. (2011). "Accelerometer signal features and classification algorithms for positioning applications," *Proceedings of the 2011 International Technical Meeting of The Institute of Navigation*, San Diego, CA, 158-169.

Theodoridis, S., Pikrakis, A., Koutroumbas, A., and Cavouras, D. (2010). *Introduction to pattern recognition. A Matlab Approach*. Cambridge, MA: Academic Press, 231.

Tripoliti, E. E., Tzallas, A. T., Tsipouras, M. G., Rigas, G., Bougia, P., Leontiou, M., et al. (2013). Automatic detection of freezing of gait events in patients with Parkinson's disease. *Comput. Methods Programs Biomed.* 110, 12–26. doi: 10.1016/j.cmpb.2012.10.016

Twomey, N., Diethe, T., Fafoutis, X., Elsts, A., McConville, R., Flach, P., et al. (2018). A comprehensive study of activity recognition using accelerometers. *Informatics* 5:27. doi: 10.3390/informatics5020027

Vapnik, V. (1979). *Estimation of Dependences Based on Empirical Data [in Russian]*. Moscow: Nauka (English translation: Springer Verlag, New York, 1982).

Wang, G., Li, Q., Wang, L., Wang, W., Wu, M., and Liu, T. (2018). Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors. *Sensors (Basel)* 18:1965. doi: 10.3390/s18061965

Wanneveich, M., Moisan, F., Jacqmin-Gadda, H., Elbaz, A., and Joly, P. (2018). Projections of prevalence, lifetime risk, and life expectancy of Parkinson's disease (2010-2030) in France. *Mov. Disord.* 33, 1449-1455. doi: 10.1002/mds.27447

Classifying Elite From Novice Athletes Using Simulated Wearable Sensor Data

Gwyneth B. Ross[1], Brittany Dowling[2], Nikolaus F. Troje[3], Steven L. Fischer[4] and Ryan B. Graham[1,4]*

[1] School of Human Kinetics, Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada, [2] Motus Global, Rockville Centre, New York, NY, United States, [3] Centre for Vision Research, York University, Toronto, ON, Canada, [4] Department of Kinesiology, University of Waterloo, Waterloo, ON, Canada

Movement screens are frequently used to identify differences in movement patterns such as pathological abnormalities or skill related differences in sport; however, abnormalities are often visually detected by a human assessor resulting in poor reliability. Therefore, our previous research has focused on the development of an objective movement assessment tool to classify elite and novice athletes' kinematic data using machine learning algorithms. Classifying elite and novice athletes can be beneficial to objectively detect differences in movement patterns between the athletes, which can then be used to provide higher quality feedback to athletes and their coaches. Currently, the method requires optical motion capture, which is expensive and time-consuming to use, creating a barrier for adoption within industry. Therefore, the purpose of this study was to assess whether machine learning could classify athletes as elite or novice using data that can be collected easily and inexpensively in the field using inertial measurement units (IMUs). A secondary purpose of this study was to refine the architecture of the tool to optimize classification rates. Motion capture data from 542 athletes performing seven dynamic screening movements were analyzed. A principal component analysis (PCA)-based pattern recognition technique and machine learning algorithms with the Euclidean norm of the segment linear accelerations and angular velocities as inputs were used to classify athletes based on skill level. Depending on the movement, using metrics achievable with IMUs and a linear discriminant analysis (LDA), 75.1–84.7% of athletes were accurately classified as elite or novice. We have provided evidence that suggests our objective, data-driven method can detect meaningful differences during a movement screening battery when using data that can be collected using IMUs, thus providing a large methodological advance as these can be collected in the field using sensors. This method offers an objective, inexpensive tool that can be easily implemented in the field to potentially enhance screening, assessment, and rehabilitation in sport and clinical settings.

Keywords: inertial measurement units, machine learning, artificial intelligence, principal component analysis, pattern recognition, athletics, movement screening

INTRODUCTION

Movement screens are widely used across many disciplines including in ergonomic, clinical, and athletic settings to identify aberrant movement patterns in hopes of decreasing risk of injury and/or improving performance (Donà et al., 2009; Kritz et al., 2009; Padua et al., 2009; Cook et al., 2014; McCall et al., 2014; McCunn et al., 2016). Most commonly, during a movement screen, an individual's movement is evaluated based on visual appraisal (McCunn et al., 2016); however, there is agreement within the literature that inter-rater and inter-session (participants tested during two separate sessions) reliability of these subjective movement screens are poor (Onate et al., 2012; Smith et al., 2013; Gulgin and Hoogenboom, 2014). Therefore, our previous research focused on the development and application of an objective framework as a data-driven alternative to objectively classify movement strategies and quality during a movement screen (Ross et al., 2018), known as the Objective Movement Assessment Tool (OMAT).

The previously published technique with optical motion capture, herein referred to as OMAT-OPT, uses principal component analysis (PCA) (Troje, 2002; Federolf et al., 2014; Young and Reinkensmeyer, 2014) in conjunction with linear discriminant analysis (LDA) to objectively differentiate and score whole-body movement patterns between desired binary classifiers (Ross et al., 2018). For OMAT-OPT, the data input into the PCA are time-series trajectories of joint centers and select anatomical markers, representing the whole-body, captured using an optical motion capture system. During a non-sport-specific movement screening battery consisting of seven unique dynamic movements that challenge stability and mobility across all major joints, between 70.7% and 82.9% of athletes were appropriately classified as either elite or novice depending on the movement (Ross et al., 2018). Although OMAT-OPT provides an objective, data-driven method that can detect meaningful movement pattern differences during a movement screening battery for binary classification, it requires optical motion capture technology, which is expensive and time-consuming to set up, capture and post-process data, reducing the accessibility and feasibility of the current technique in clinical, ergonomic, and sport settings (Hadjidj et al., 2013).

The use of wearable systems are increasing in popularity in clinical, sport, and ergonomic settings (Patel et al., 2012; Hadjidj et al., 2013), offering an inexpensive alternative to optical motion capture systems. The wearable systems are easily transportable, require minimal post-processing, are able to collect data in larger capture volumes compared to optical systems, and are immune to problems associated with optical systems such as occlusion and line-of-sight problems (Zhou and Hu, 2008). A common type of sensor used is the inertial measurement unit (IMU). IMUs contain an accelerometer, gyroscope, and magnetometer, allowing measurement of linear accelerations and angular velocities in three axes and the triaxial magnetic fields of the earth. IMUs are susceptible to drift, especially when close to metal, although more robust algorithms are continuously being developed to mitigate these effects (Madgwick et al.,

2011; Wittmann et al., 2019), making them more suitable for use in the field.

IMU data have been used to objectively classify movement based on different classifiers during non-sport specific tasks (Sgro et al., 2017; Johnston et al., 2016, 2019; Zago et al., 2019). Machine learning with IMU data as the input has been able to objectively identify children of different motor development levels during a standing long jump (Sgro et al., 2017), rugby players at a higher risk of a sport-related concussion based on a Y-balance test (Johnston et al., 2019), Australian football players at different levels of fatigue during a Y-balance test (Johnston et al., 2016), and to predict change of direction, speed, and mechanical work during cutting maneuvers (Zago et al., 2019), to name a few. Although these studies only looked at a single IMU placed on the low-back of the participant, these findings suggest that IMUs can be used as an inexpensive alternative to optical motion capture to characterize and classify motion.

Although research using machine learning to classify elite and novice athletes is limited, discriminant analysis has been previously used to classify novice, good, and elite rowers during ergometer testing (Smith and Spinks, 1995). The ability to objectively differentiate movement patterns between novice and elite athletes is useful to highlight emergent differences in movement performance. Guided by those differences, coaches can improve quality of feedback to their athletes (Smith and Spinks, 1995). We chose skill level as the dichotomous factor to initially assess due to its likelihood to influence movement quality and performance, with the intention of in the future expanding to sex, sport played, and injury history or risk.

Feature selection approaches and machine learning algorithms may also influence the accuracy of classification between elite and novice athletes using IMU data and are therefore important secondary considerations. Previously, the OMAT-OPT used the first 35 principal component (PC) scores as the input data for the LDA; however, alternative feature selection approaches could provide an objective method to best decide which PC scores to use as input data to maximize classification. Ensemble feature selection, which is based on the same ideology of ensemble supervised classifiers, is a useful approach to evaluate. Ensemble feature selection includes the use of multiple feature selection algorithms to select features and has been found to have greater stability (i.e., less likelihood of features changing if data are added or removed) and better generalizability than using a single feature selection technique (Saeys et al., 2008). In addition, the OMAT currently uses LDA, which was selected due to superior performance during testing. However, it is unknown whether LDA would still garner the highest classification rates when using PC scores selected by an ensemble feature selection approach, rather than the first 35 PC scores and/or when using IMU data. Alternative machine learning algorithms including binary logistic regression (BLR), decision trees (DT), K-nearest neighbors (kNN), naïve Bayes (NB), support vector machine with a linear kernel (SVM), and support vector machine with a radial basis function kernel (RBF) may strengthen classification accuracy relative to our existing LDA approach. As a result, while investigating the utility of IMUs to classify movements between novice and elite athletes, it

remains important to concurrently evaluate the underlying machine learning model architecture required to generate the best possible classification.

Therefore, the purpose of this study was to assess the ability of the previously developed framework to differentiate whole-body movement patterns between novice and elite athletes performing a non-sport-specfic movement screening battery when using data extractable from an IMU (i.e., simulated IMU data; OMAT-sIMU), which can be collected easily and inexpensively in the field. Although data in the current study are simulated IMU data based on optical motion capture, this study provides proof-of-concept that IMU-based data can provide enough information to successfully classify athletes' movement patterns based on skill level. A secondary purpose of this study was to refine the architecture of the OMAT to optimize classification rates by incorporating feature selection and multiple machine learning algorithms (i.e., BLR, DT, KNN, LDA, NB, SVM, and RBF) for both the OMAT-OPT and OMAT-sIMU.

MATERIALS AND METHODS

Participants
Kinematic data were collected on 542 athletes by Motus Global (Rockville Centre, NY, United States). The sample included athletes competing in 11 different sports (i.e., baseball, basketball, soccer, golf, tennis, track and field, squash, cricket, lacrosse, football, or volleyball) and ranging in skill level from recreational to professional (e.g., NBA, MLB, NFL, PGA, FIFA). The athletes were assigned to either the novice or elite group based on previous research that found that those athletes accruing over 10,000 h of deliberate practice are experts in their sport (Helsen et al., 1998; Baker et al., 2003). Therefore, athletes competing at the inter-collegiate, semi-professional, and professional level were considered elite athletes and those competing at less competitive levels (e.g., high-school, youth, recreational, etc.) were considered novices. Before data collection, each athlete read and signed an informed consent form permitting Motus Global to use the data for future research. The Health Sciences Research Ethics Board at the University of Ottawa approved the secondary use of the data for research purposes (file no: H-08-18-1085).

Protocol
Upon arrival to the Motus Global laboratory, each athlete read and signed an informed consent form, provided information on injury history for the previous 10 years, and had their height (with shoes on) and weight recorded. The athlete was then outfitted with 45 passive, reflective markers (B&L Engineering, Santa Ana, CA) to capture whole-body motion (Mcpherson et al., 2016; Ross et al., 2018). After being outfitted with the markers, the athlete completed a static and dynamic calibration trial (Ross et al., 2018). The static calibration trial was used to develop a whole-body biomechanical model for each athlete.

After the calibration trials, each athlete completed a movement battery consisting of 21 unique movements testing athletes' range of motion at each joint, stability, power and balance. However, only seven movements were used in the analysis due to their dynamic nature and ability to challenge the athletes' coordination, stability, and mobility across all major joints. The seven tasks included: drop jump, bird-dog, hop-down, lunge, step-down, L-hop, and T-balance (**Figure 1**). Each movement was performed bilaterally on the left and right side except for the drop jump which was performed symmetrically, resulting in a total of 13 movement trials (Ross et al., 2018). The athlete performed each task until they believed they did it to the best of their ability with only the trial that was deemed the best being retained for each athlete. Full-body motion data were captured at 120 Hz using an 8-camera Raptor-E (Motion Analysis Corporation, Santa Rosa, CA) motion capture system.

Data Analysis
Pre-processing
Motion capture data were collected, labeled, and gap-filled using Cortex (Motion Analysis Corporation, Santa Rosa, CA). Data from anatomical landmarks and the tracking markers during the calibration trial were used to develop a whole-body 3D kinematic model in Visual3D v6 (C-Motion, Inc., Germantown, MD). The model was then applied to all motion trials outputting joint centers bilaterally for the wrist, elbow, shoulder, foot, ankle, knee, and hip; centers of gravity for the trunk, head, and pelvis; marker positional data for the left and right heel, T_2, T_8, sternum, and the back, front and sides of the head for the OMAT-OPT model and segment angular velocities and center of gravity (CoG) linear velocities of the head, trunk, pelvis, upper arms, forearms, thighs, shanks, and feet for the OMAT-sIMU model. Data were then exported and analyzed using Python 3.0. All trials were trimmed to specific start and end-point criteria (Ross et al., 2018), and filtered using a dual-pass, low-pass Butterworth filter with a cutoff of 15 Hz. Since elite athletes were significantly taller than novices ($F = 138.25$, $p < 0.001$), all data for each movement were normalized by each athlete's individual height by dividing each raw data point by their own height. Normalization ensured that differences in PC scores between groups were not strictly due to variation in size.

OMAT-OPT data
The 3D positional data of the joint centers and markers retained in the OPT model for each participant were rotated so that the local coordinate system of the trunk was aligned with the global coordinate system. The data were then translated so that the midpoint between the left and right hip of the first frame of data was aligned with the global origin (i.e., midpoint of left and right hip equaled 0,0,0 for x, y, and z coordinates, respectively). The rotated 3D data were then time normalized to 500 frames using Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation to control for differences in absolute movement time for each participant. An [$n \times 39,000$] matrix for each movement was then constructed, where n was the number of subjects and 39,000 was the time-normalized x, y, and z data for each joint center, center of gravity, and retained markers mentioned above (26 positions $\times$ 3 axes $\times$ 500 time points). Due to marker occlusion and some athletes not performing all tasks, n was dependent on the movement task (**Table 1**).

FIGURE 1 | Schematic drawings of the seven unique movements performed by the athletes: drop-jump, hop-down, L-hop, bird-dog, lunge, step-down, and T-balance.
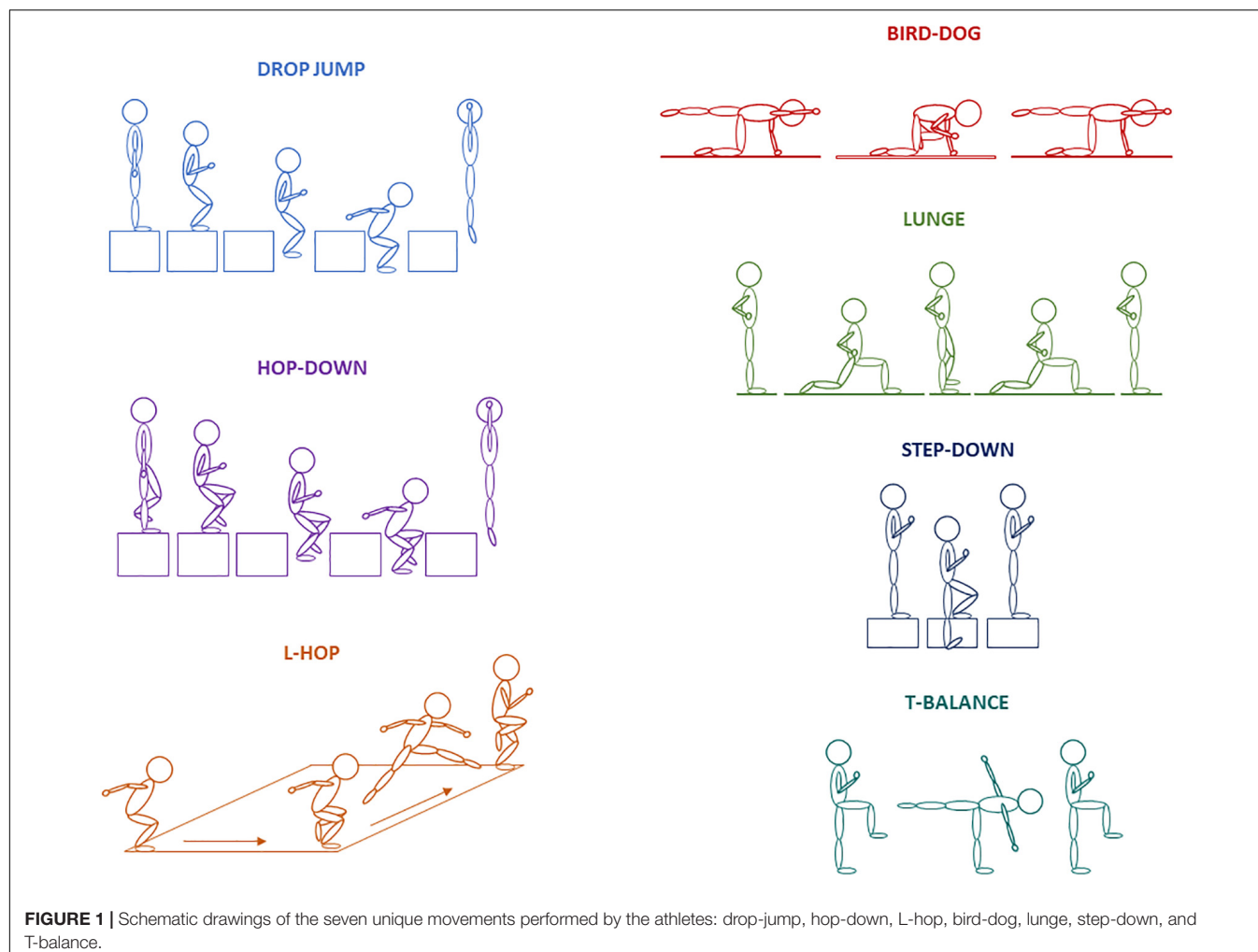
TABLE 1 | OMAT-OPT: The number of athletes broken down by sex and skill level and the overall classification accuracy, hit rate, miss rate, false alarm (FA) rate, correct rejection (CR) rate, D' and C when the optimal number of PCs were retained for each movement task.

Movement	n	Male		Female		# of PCs	Accuracy (%)	Hit	Miss	FA	CR	D'	C
		Elite	Novice	Elite	Novice								
Bird-Dog Left	380	242	83	12	43	10	82.63	0.91	0.09	0.34	0.66	1.75	−0.46
Bird-Dog Right	387	244	88	11	44	18	80.88	0.90	0.10	0.36	0.64	1.62	−0.46
Drop Jump	275	168	64	7	36	12	80.36	0.88	0.12	0.33	0.67	1.61	−0.37
Hop-Down Left	396	242	99	10	45	14	77.27	0.87	0.13	0.40	0.60	1.39	−0.45
Hop-Down Right	396	242	97	11	46	9	74.24	0.84	0.16	0.43	0.57	1.17	−0.40
L-Hop Left	266	159	67	6	34	15	83.83	0.89	0.11	0.25	0.75	1.91	−0.27
L-Hop Right	267	160	67	6	34	14	79.03	0.87	0.13	0.34	0.66	1.54	−0.35
Lunge Left	399	246	97	12	44	11	78.20	0.87	0.13	0.38	0.62	1.43	−0.40
Lunge Right	401	248	97	12	44	17	78.30	0.88	0.12	0.39	0.61	1.44	−0.44
Step-Down Left	399	246	98	12	43	17	75.94	0.84	0.16	0.40	0.60	1.28	−0.38
Step-Down Right	399	247	96	11	45	16	74.19	0.83	0.17	0.42	0.58	1.16	−0.37
T-Balance Left	392	244	92	11	45	13	77.30	0.89	0.11	0.45	0.55	1.37	−0.56
T-Balance Right	395	244	94	12	45	18	73.16	0.83	0.17	0.45	0.55	1.08	−0.41
Average	**365.54**	**225.54**	**87.62**	**10.23**	**42.15**	**14.15**	**78.10**	**0.87**	**0.13**	**0.38**	**0.62**	**1.44**	**−0.41**
STD	**55.17**	**36.14**	**13.11**	**2.31**	**4.38**	**3.02**	**3.26**	**0.03**	**0.03**	**0.06**	**0.06**	**0.25**	**0.07**

Bold values represent the average and standard deviation across all movement tasks.

OMAT-sIMU data

In order to simulate IMU accelerometer data, we extracted the segment CoG linear velocities of each segment retained in the model and then differentiated the data once to calculate segment linear accelerations, and to simulate the IMU gyroscope data, we extracted the segment angular velocities of the same segments. Once all data were extracted and calculated, the data were time normalized to 500 frames as per the OMAT-OPT. In anticipation of the future implementation of the method in clinic or industry, the Euclidean norm (i.e., square root of the sum of squares) of the x, y, and z axes of the linear segment accelerations and segment angular velocities were taken to minimize the effect of sensor brand or orientation (Clouthier et al., 2020) and to reduce the dimensionality of the data (Bergmann et al., 2014).

A matrix for each movement was then constructed with the Euclidean norm of the linear segment accelerations and segment angular velocities for each segment and each participant. Segment linear accelerations and angular velocities were chosen to mimic outputs collected via IMUs. Each matrix was n (number of participants; **Table 2**) $\times$ 13000 (Euclidean norm $\times$ 2 data features $\times$ 13 body segments $\times$ 500 time points). Because the units were different between the data features (i.e., linear accelerations in m/s^2, angular velocities in rad/s), the scale of the data between the two data features varied widely, which would lead to classification being driven primarily by the data feature with the larger scale. Therefore, the data were feature scaled to be between 0 and 1 for each movement using scikit-learn Robust Scaler (Pedregosa et al., 2011), which removes the median from each feature and scales the data according to the 1st and 3rd quartile of the data, mitigating the effect of outliers during scaling.

Feature Selection

For both OMAT-sIMU and OMAT-OPT, PCA was applied to each matrix, resulting in a unique model per task per data type (i.e., OMAT-sIMU, OMAT-OPT). Using the PC scores as features, ensemble feature selection, consisting of six common feature selection techniques (Pearson correlation, chi-squared, recursive feature elimination, lasso, random forest, and LightGBM), was used to rank the PCs based on contribution to the model for each movement task and data type. Ensemble feature selection has been found to improve the robustness of feature ranking and feature subset selection as well as increase the generalizability of the features selected (Saeys et al., 2008). The scikit-learn library was used for the chi-squared, recursive feature elimination, lasso, and random forest (Pedregosa et al., 2011). The top 25 features per data type were retained for each technique. The features were then sorted based on the number of techniques where they ranked in the top 25 features. PC scores that ranked in the top 25 for at least 50% of the techniques (i.e., 3) were retained for the classifier (**Table 1**; OMAT-OPT and **2**; OMAT-sIMU). To minimize overfitting of the models, the maximum number of features retained was the square root of the number of samples for each movement task (Hua et al., 2005) (e.g., lunge right had 401 samples, therefore a maximum of 20 PC scores could be retained for that task).

Classification

To refine the architecture of the OMAT, seven different kinds of classifiers were used: BLR, DT, kNN, LDA, NB, SVM, and RBF to classify athletes based on skill level (elite vs. novice). All classifiers were employed using the scikit-learn library (Pedregosa et al., 2011). For all classifiers, PC scores retained from feature selection were used as predictors and leave-one-out cross-validation was used for validation. Each model was rerun to use between 1 and the total number of PCs retained to determine the optimal number of PCs to retain for each classifier for each movement task. The model with the highest classification rate was deemed the optimal model. Due to a lack of a testing dataset, leave-one-out validation was used where one of the athletes' data were taken out (test athlete) and the PCA, feature selection, and classifier models were computed on the remaining athletes (training athletes). After computing the new PCA, feature selection, and classifier models, the test athlete was projected into the PCA, feature selection, and classifier model spaces computed on the training athletes. The procedure was repeated until all athletes had been left out and projected back into the PCA, feature selection, and classifier models (Troje, 2002; Ross et al., 2018).

Signal Detection Theory

For the best classifier for each data type, to test the separation between the signal and the noise and to determine the strategy used by the frameworks, a signal detection theory (SDT) model was used for each optimal model retained. In SDT, there are four types of classification: (1) Hit, (2) Miss, (3) False alarm (FA), and (4) Correct rejection (CR) (Abdi, 2007). For this study, a hit was when an elite athlete was correctly classified as an elite (equivalent to sensitivity), a miss was when an elite athlete was misclassified as a novice, a FA was when a novice athlete was misclassified as an elite, and a CR was when a novice athlete was correctly classified as a novice (equivalent to specificity). Parameter D' is calculated subtracting the probability (z-score) of a false alarm from the probability (z-score) of a hit and tells the distance between the two peaks (e.g., elite and novice) in standard deviations; the higher the score the more separable the two groups are with a score of 0 representing chance (Abdi, 2007). Parameter C is calculated by taking the average probability (z-score) of a hit and false alarm and represents the strategy used by the framework. A positive value represents the framework being conservative (e.g., more likely to classify an athlete as novice), where as a negative value represents the framework being liberal (e.g., more likely to classify an athlete as elite) (Abdi, 2007). The closer the value is to 0, the closer the framework is to being the ideal observer (e.g., not more likely to classify as either elite or novice) (Abdi, 2007).

RESULTS

OMAT-OPT

For all tasks the linear classifiers (i.e., BLR, LDA, and SVM) outperformed DT, kNN, and RBF, except RBF performed as well as the linear classifiers for the lunge left and step-down left (**Figure 2**). For the drop-jump, hop-down left, L-hop left, and

TABLE 2 | OMAT-sIMU: The number of athletes broken down by sex and skill level and overall classification accuracy, hit rate, miss rate, false alarm rate, correct rejection rate, D' and C when the optimal number of PCs were retained for each movement task.

| Movement | n | Male | | Female | | # of PCs | Accuracy (%) | Hit | Miss | FA | CR | D' | C |
		Elite	Novice	Elite	Novice								
Bird-Dog Left	380	242	83	12	43	6	81.22	0.92	0.08	0.45	0.55	1.56	−0.66
Bird-Dog Right	387	244	88	11	44	7	81.98	0.90	0.10	0.49	0.51	1.34	−0.64
Drop Jump	275	168	64	7	36	14	84.67	0.89	0.11	0.27	0.73	1.81	−0.29
Hop-Down Left	396	242	99	10	45	10	81.75	0.88	0.13	0.36	0.64	1.51	−0.40
Hop-Down Right	396	242	97	11	46	13	79.70	0.89	0.11	0.40	0.60	1.49	−0.50
L-Hop Left	266	159	67	6	34	12	83.15	0.87	0.13	0.29	0.71	1.70	−0.29
L-Hop Right	267	160	67	6	34	15	82.71	0.90	0.10	0.33	0.67	1.71	−0.41
Lunge Left	399	246	97	12	44	18	80.70	0.93	0.07	0.50	0.50	1.47	−0.74
Lunge Right	401	248	97	12	44	18	81.25	0.91	0.09	0.38	0.62	1.62	−0.51
Step-Down Left	399	246	98	12	43	12	75.19	0.87	0.13	0.52	0.48	1.07	−0.60
Step-Down Right	399	247	96	11	45	6	76.37	0.88	0.12	0.50	0.50	1.17	−0.60
T-Balance Left	392	244	92	11	45	14	76.47	0.90	0.10	0.55	0.45	1.15	−0.71
T-Balance Right	395	244	94	12	45	10	75.13	0.87	0.13	0.50	0.50	1.14	−0.58
Average	**365.54**	**225.54**	**87.62**	**10.23**	**42.15**	**11.92**	**80.02**	**0.89**	**0.11**	**0.43**	**0.57**	**1.44**	**−0.53**
STD	**55.17**	**36.14**	**13.11**	**2.31**	**4.38**	**4.03**	**3.19**	**0.02**	**0.02**	**0.10**	**0.10**	**0.25**	**0.15**

Bold values represent the average and standard deviation across all movement tasks.
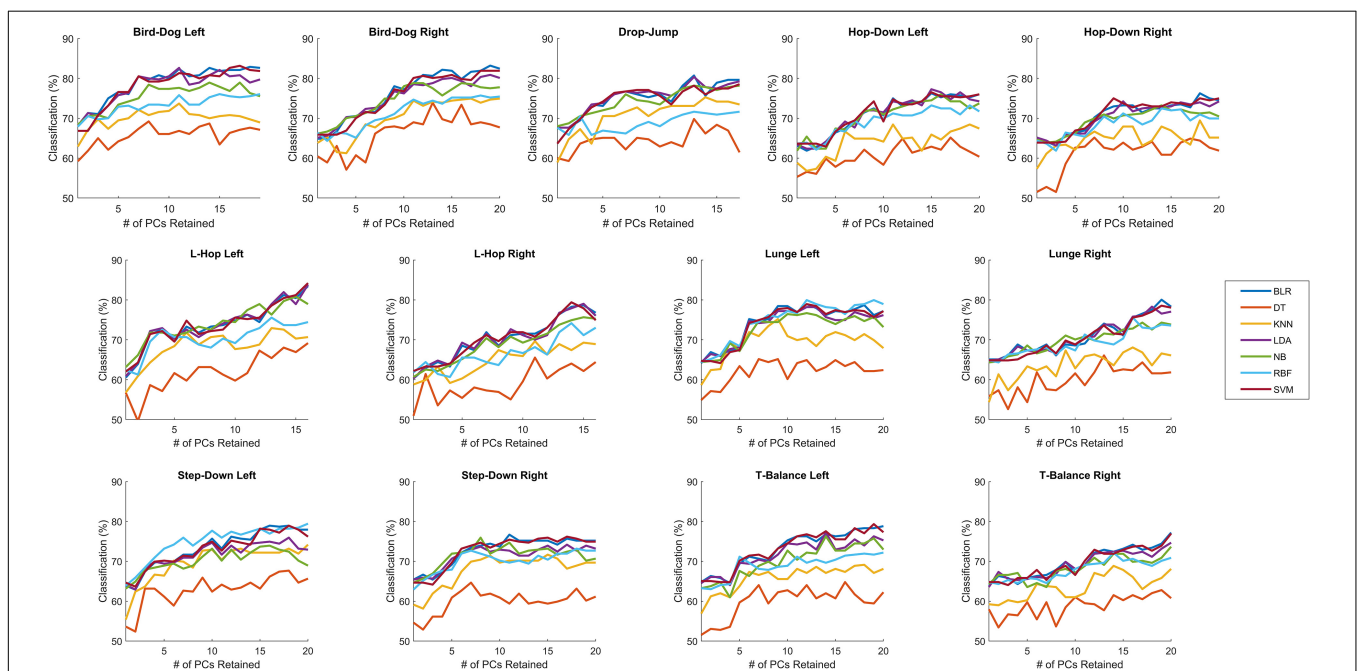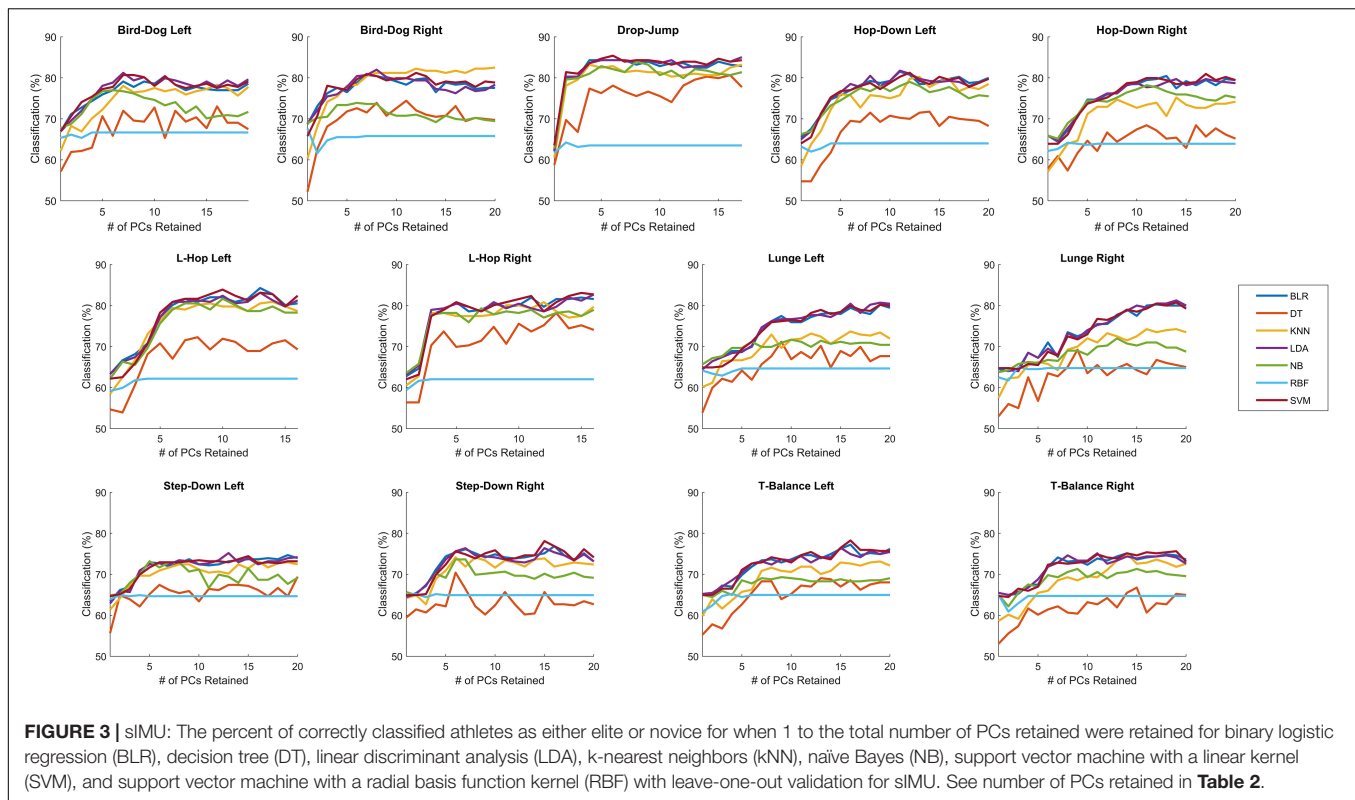

FIGURE 2 | OPT: The percent of correctly classified athletes as either elite or novice for when 1 to the total number of PCs retained were retained for binary logistic regression (BLR), decision tree (DT), linear discriminant analysis (LDA), k-nearest neighbors (kNN), naïve Bayes (NB), support vector machine with a linear kernel (SVM), and support vector machine with a radial basis function kernel (RBF) with leave-one-out validation for OPT. See number of PCs retained in **Table 1**.

lunge left, NB performed as well as the linear classifiers, however, for all other tasks, they performed in between the linear classifiers and DT, kNN, and RBF. Since there were minimal differences (<0.5%) on the average classification rates for all tasks between BLR, LDA, and SVM, and to be able to compare the current results to previous results (Ross et al., 2018), LDA was selected for further analysis. When using LDA, the optimal number of PCs

retained ranged from 9 (hop-down right) to 18 (bird-dog left, T-balance right) with an average of 14.15 ± 3.02 PCs retained (**Table 1**). The OMAT-OPT accurately classified between 73.1% (T-balance right) to 83.8% (L-hop left) of athletes as either elite or novice (**Table 1**). The average classification rate across all tasks was 78.1% ± 3.26%. For SDT, on average, OMAT-OPT had a hit, miss, FA, and CR rate of 0.87 ± 0.03, 0.13 ± 0.03, 0.38 ± 0.06, and

FIGURE 3 | sIMU: The percent of correctly classified athletes as either elite or novice for when 1 to the total number of PCs retained were retained for binary logistic regression (BLR), decision tree (DT), linear discriminant analysis (LDA), k-nearest neighbors (kNN), naïve Bayes (NB), support vector machine with a linear kernel (SVM), and support vector machine with a radial basis function kernel (RBF) with leave-one-out validation for sIMU. See number of PCs retained in **Table 2**.

0.62 ± 0.06, respectively (**Table 1**). The average D' was 1.44 ± 0.25 and the average C was -0.41 ± 0.07 (**Table 1**).

OMAT-sIMU

Similar to OMAT-OPT, for all tasks the linear classifiers (i.e., BLR, LDA, and SVM) outperformed all the other classifiers (i.e., DT, kNN, NB, and RBF), except KNN performed as well as the linear classifiers in the bird-dog left and right, hop-down left, and step-down left (**Figure 3**). Since there were again minimal differences between the average classification rates for all movement tasks between BLR, LDA, and SVM, LDA was selected for further analysis. When using segment linear accelerations and angular velocities, data available from an IMU system, the optimal number of PCs retained ranged from 6 (bird-dog left, step-down right) to 18 (lunge left and right) with an average of 11.92 ± 4.03 PCs retained (**Table 2**). The OMAT-sIMU accurately classified between 75.1% (T-balance right) to 84.7% (drop-jump) of athletes as either elite or novice (**Table 2**). The average classification rate across all tasks was $80.0\% \pm 3.19\%$. For SDT, on average, OMAT-sIMU had a hit, miss, FA, and CR rate of 0.89 ± 0.02, 0.11 ± 0.02, 0.43 ± 0.1, and 0.57 ± 0.1, respectively (**Table 2**). The average D' was 1.44 ± 0.25 and the average C was -0.53 ± 0.15 (**Table 2**).

When comparing the OMAT-OPT and OMAT-sIMU classification rates on average, OMAT-sIMU had higher classification rates than OMAT-OPT by 1.92%. OMAT-sIMU outperformed OMAT-OPT in the bird-dog right (1.1%), drop-jump (4.31%), hop-down left (4.48%) and right (5.46%), L-hop right (3.68%), lunge left (2.50%) and right (2.95%), step-down right (2.18%), and T-balance right (1.97%), whereas OMAT-OPT

had a higher classification rate than OMAT-sIMU for bird-dog left (1.41%) (**Figure 4**). The two models performed relatively the same (< 1% difference) for the L-hop left, step-down left, and T-balance left.

DISCUSSION

The primary purpose of this study was to assess the ability of the OMAT to differentiate whole-body movement patterns between novice and elite athletes performing a non-sport-specific movement screening battery using data able to be collected via an IMU. The secondary purpose of this study was to refine the architecture of the OMAT by incorporating feature selection and testing multiple classifiers. For both the OMAT-OPT and OMAT-sIMU, BLR, LDA, and SVM, on average, outperformed all other classifiers tested. These findings suggest that the data can be separated using a linear plane; and therefore, the use of more complicated, computationally expensive non-linear classifiers is not only not required, but can be detrimental. There were minimal differences between BLR, LDA, and SVM, so therefore in order to easily compare the current results with previous work, LDA was chosen as the classifier to report the results.

OMAT-OPT with feature selection outperformed the previously published results on 7 of the 13 tasks (i.e., bird-dog left and right, hop-down left, L-hop left, lunge left and right, and T-balance left) and OMAT-sIMU was able to outperform the previously published results of the OMAT-OPT in all tasks except the step-down left (Ross et al., 2018). This in part is due
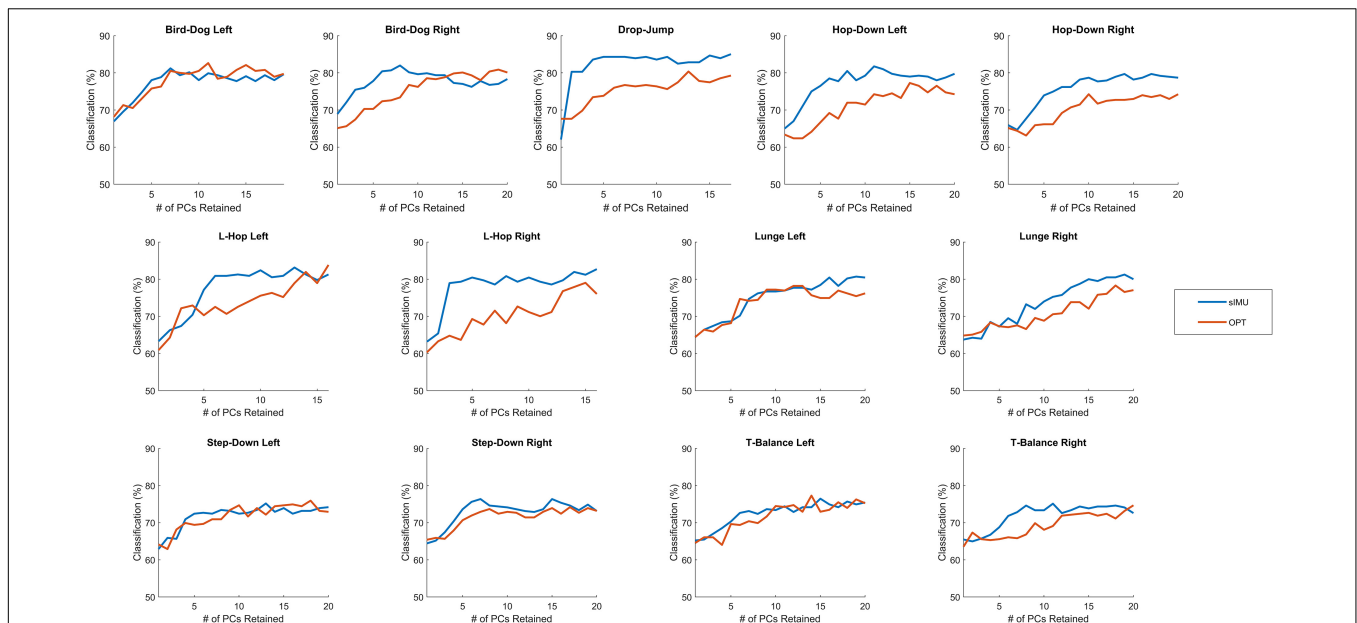
FIGURE 4 | The percent of correctly classified athletes as either elite or novice for when 1 to the total number of PCs retained were retained for the linear discriminant analysis with leave-one-out validation when using OPT and sIMU data.

to the introduction of feature selection into the methodology, reinforcing the value of this approach for future work aiming to objectively classify movement patterns. Compared to the previous study, where PCs 1–35 were retained (Ross et al., 2018), using feature selection, we are now able to have greater classification rates using fewer PCs, which requires fewer computational resources and decreases the risk of overfitting for 7 of the 13 tasks. OMAT-sIMU outperformed or performed equally to OMAT-OPT in all movements except the bird-dog left. These findings suggest that the OMAT-sIMU approach better captures movement pattern differences between novice and elite athletes compared to OMAT-OPT data. This is thought to be due to the different types of data analyzed for OMAT-OPT and OMAT-sIMU. OMAT-OPT uses joint center trajectories, which due to the constrained nature of the tasks, may be capturing more gross motor patterns that are unrelated to skill. In contrast, the OMAT-sIMU uses linear acceleration and angular velocity that are more likely to capture the smoothness of the movement, which may be a better indicator of skill level than gross motor patterns. However, for both the OMAT-OPT and OMAT-sIMU, when looking at trends in individual athlete data across tasks, if there were differences in how the athlete was classified between tasks, athletes tended to be classified the same on all tasks that were targeting the same skill set (e.g., trunk stability, jumping, balance) and if there were discrepancies on how the left/right tasks were classified, the dominant side was usually classified as elite. This suggests that relevant differences between elite and novice-like movement patterns can be detected using both data types. A combined approach of using both sIMU and OPT data may provide even better classification rates than using sIMU or OPT alone due to the two types of data potentially capturing different movement features.

Previously, in order to assess how well the framework was classifying elite and novice athletes on a group basis, the percent of correctly classified elites and novices were calculated (Ross et al., 2018). SDT was chosen for this current study because it provides classification rates for each group (e.g., hit and correct rejection) as well as the additional information of response bias. For all tasks, both OMAT-OPT and OMAT-sIMU had higher rates of correctly classifying elite athletes (depicted by the increased hit and decreased miss rates) compared to novice athletes (depicted by the decreased CR and increased FA rates). For all tasks, D' was greater than 1.08 and 1.07 for OMAT-OPT and OMAT-sIMU, respectively, suggesting that elite and novice athletes are separable when using both OMAT-OPT and OMAT-sIMU. However, on average, the data are more robustly classified when using OMAT-sIMU data compared to OMAT-OPT data. Lastly, for all tasks, for both OMAT-OPT and OMAT-sIMU, the framework was more likely to classify the athlete as elite than the ideal observer. A potential reason for this could be that some of the novice athletes were attending an elite youth sports academy, which boasts a high percentage of students continuing to compete at the collegiate and professional levels. Therefore, some of the novice athletes were on track to become elite athletes at the time of testing. On average, OMAT-OPT acted more closely to the ideal observer than OMAT-sIMU, based on our definition of a hit and correct rejection; this is represented by the smaller C value.

Although on average the models using the OMAT-sIMU data as the inputs, had higher classification rates than OMAT-OPT, a limitation of the OMAT-sIMU data is that it is more difficult to interpret differences between elites and novices compared to OMAT-OPT, making it harder to train individuals to improve their movement patterns. With only linear accelerations and

angular velocities, and no video data, it is hard to discern exactly how the athlete is moving within space to obtain a score more representative of an elite or novice athlete. IMUs may offer an inexpensive measurement device to objectively screen movement abilities, where those individuals identified with weaknesses can then be tested more in depth with optical motion capture to inform targeted corrective exercise approaches.

A limitation of this study is the use of camera-based motion capture to calculate the linear accelerations and angular velocities for each segment and not raw data collected from IMUs. This technique has been used in previous research when the desired database does not contain IMU data (Young, 2010). Although linear accelerations and angular velocities would change when using IMUs, due to the inability to place IMUs at the CoG of segments, previous research has found strong agreement between IMU outputs and optical motion capture outputs (Mcginnis et al., 2014; Bolink et al., 2018). Even though the data were not raw data from an IMU, we purposefully took the Euclidean norm of the data to increase ecological validity and to remove the effect of different local coordinate system orientations of the linear acceleration and angular velocities within the global coordinate system. In addition, we differentiated positional data, which introduces noise to the data that would not be present when collecting data via IMU and were still able to get high classification rates. We are confident that these classification results are representative and may be lower than that of what would be achieved using sensors themselves, which we are in the process of testing. A second limitation of the study is the assumption athletes at the collegiate and professional level completed 10,000 h of deliberate practice. However, athletes competing at the professional and inter-collegiate levels would be in the higher echelon of athletes in their sport even if not completing 10,000 h. Nonetheless, this paper provides proof-of-concept that the OMAT is able to accurately classify athletes as novice or elite with consistent or improved accuracy when using data available from IMUs, relative to whole-body marker data. Future research should investigate the ability to classify athletes using OMAT using segment linear acceleration and angular velocity data collected using IMUs, fine-tuning algorithms to increase classification rates, and exploring other classifiers such as sport played, injury risk, and sex.

CONCLUSION

In conclusion, the introduction of feature selection increased the classification rates compared to using the first 35 PC scores and BLR, LDA, and SVM produced the highest classification

rates although there were minimal differences (<0.5%) between the three. Segment linear acceleration and angular velocity data readily available from an IMU could differentiate athletes' movement performance based on skill level when using a novel machine learning approach (Ross et al., 2018) with a level of accuracy consistent with the use of whole-body motion capture data. These data suggest that IMUs, in conjunction with OMAT, may provide an inexpensive and timely way to objectively characterize and classify movement performance in the field, providing a feasible method for coaches and clinicians to objectively measure performance.

DATA AVAILABILITY STATEMENT

Sample code and data are available at: https://doi.org/10.5281/zenodo.3575075.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Health Sciences and Science Research Ethics Board – University of Ottawa. Written informed consent to participate in this study was provided by the participant or if under the age of consent, by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

GR, RG, NT, and SF conceived the study and interpreted the results. GR and BD collected the data and performed the pre-processing of the data. GR implemented the OMAT, analyzed the results, and prepared the manuscript. All authors revised the manuscript.

REFERENCES

Abdi, H. (2007). "Signal detection theory (SDT)," in *Encyclopedia of Measurement and Statistics*, eds N. Salkini and K. Rasmussen (Thouse Oaks: SAGE Publications Ltd), 886–889.

Baker, J., Cote, J., and Abernethy, B. (2003). Sport-specific practice and the development of expert decision-making in team ball sports. *J. Appl. Sport Psychol.* 15, 12–25. doi: 10.1080/10413200305400

Bergmann, J. H. M., Langdon, P. M., Mayagoitia, R. E., and Howard, N. (2014). Exploring the use of sensors to measure behavioral interactions: an experimental evaluation of using hand trajectories. *PLoS One* 9:e0088080.

Bolink, S. A. A. N., Naisas, H., Senden, R., Essers, H., Heyligers, I. C., Meijer, K., et al. (2018). Validity of an inertial measurement unit to assess pelvic orientation angles during gait, sit – stand transfers and step-up transfers: comparison with an optoelectronic motion captur. *Med. Eng. Phys.* 38, 225–231. doi: 10.1016/j.medengphy.2015.11.009

Clouthier, A. L., Ross, G. B., Graham, R. B., and Graham, R. B. (2020). Sensor data required for automatic recognition of athletic tasks using deep neural networks. *Front. Bioeng. Biotechnol.* 7:473. doi: 10.3389/fbioe.2019.00473

Cook, G., Burton, L., and Hoogenboom, B. J. (2014). Functional movement screening: the use of fundamental movements as an asssment of function- part 2. *Int. J. Sports Phys. Ther.* 9, 549–563. doi: 10.1111/j.1600-0838.2010.01267.x

Donà, G., Preatoni, E., Cobelli, C., Rodano, R., and Harrison, A. J. (2009). Application of functional principal component analysis in race walking: an emerging methodology. *Sport. Biomech.* 8, 284–301. doi: 10.1080/14763140903414425

Federolf, P., Reid, R., Gilgien, M., Haugen, P., and Smith, G. (2014). The application of principal component analysis to quantify technique in sports. *Scand. J. Med. Sci. Sports* 24, 491–499. doi: 10.1111/j.1600-0838.2012.01455.x

Gulgin, H., and Hoogenboom, B. (2014). The functional movement screening (Fms)TM: an inter-rater reliability study between raters of varied experience. *Int. J. Sports Phys. Ther.* 9, 14–20.

Hadjidj, A., Souil, M., Bouabdallah, A., Challal, Y., and Owen, H. (2013). Wireless sensor networks for rehabilitation applications: challenges and opportunities. *J. Netw. Comput. Appl.* 36, 1–15. doi: 10.1016/j.jnca.2012.10.002

Helsen, W. F., Starkes, J. L., and Hodges, N. J. (1998). Team sports and the theory of deliberate practice. *J. Sport Exerc. Psychol.* 20, 12–34. doi: 10.1123/jsep.20.1.12

Hua, J., Xiong, Z., Lowey, J., Suh, E., and Dougherty, E. R. (2005). Gene expression Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 21, 1509–1515. doi: 10.1093/bioinformatics/bti171

Johnston, W., O'Reilly, M., Dolan, K., Reid, N., Coughlan, G., and Caulfield, B. (2016). "Objective classification of dynamic balance using a single wearable sensor," in *Proceedings of the 4th International Congress on Sports Sciences Research and Technology Support 2016*, Porto, 15–24.

Johnston, W., Reilly, M. O., Duignan, C., Liston, M., Mcloughlin, R., Coughlan, G. F., et al. (2019). Association of dynamic balance with sports-related concussion a prospective cohort study. *Am. J. Sports Med.* 47, 197–205. doi: 10.1177/0363546518812820

Kritz, M., Cronin, J., and Hume, P. (2009). The bodyweight squat: a movement screen for the squat pattern. *Strength Cond. J.* 31, 76–85. doi: 10.1519/SSC.0b013e318195eb2f

Madgwick, S. O. H., Harrison, A. J. L., and Vaidyanathan, R. (2011). "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR)*, Zurich, 1–7.

McCall, A., Carling, C., Nedelec, M., Davison, M., Le Gall, F., Berthoin, S., et al. (2014). Risk factors, testing and preventative strategies for non-contact injuries in professional football: current perceptions and practices of 44 teams from various premier leagues. *Br. J. Sports Med.* 48, 1352–1357. doi: 10.1136/bjsports-2014-093439

McCunn, R., aus der Fünten, K., Fullagar, H. H. K., McKeown, I., and Meyer, T. (2016). Reliability and association with injury of movement screens: a critical review. *Sport. Med.* 46, 763–781. doi: 10.1007/s40279-015-0453-1

Mcginnis, R. S., Cain, S. M., Davidson, S. P., Vitali, R. V., Mclean, S. G., and Perkins, N. C. (2014). "Validation of complementary filter based IMU data fusion for tracking torso angle and rifle orientation," in *Proceedings of the ASME 2014 International Mechanical Engineering Congress and Exposition IMECE*, Montreal.

Mcpherson, A. L., Dowling, B., Tubbs, T. G., and Paci, J. M. (2016). Sagittal plane kinematic differences between dominant and non-dominant legs in unilateral and bilateral jump landings. *Phys. Ther. Sport* 22, 54–60. doi: 10.1016/j.ptsp.2016.04.001

Onate, J. A., Dewey, T., Kollock, R. O., Thomas, K. S., Van Lunen, B. L., DeMaio, M., et al. (2012). Real-time intersession and interrater reliability of the functional movement screen. *J. Strength Cond. Res.* 26, 408–415. doi: 10.1519/JSC.0b013e318220e6fa

Padua, D. A., Marshall, S. W., Boling, M. C., Thigpen, C. A., Garrett, W. E., and Beutler, A. I (2009). The landing error scoring system (LESS) is a valid and reliable clinical assessment tool of jump-landing biomechanics: the JUMP-ACL study. *Am. J. Sports Med.* 37, 1996–2002. doi: 10.1177/0363546509343200

Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *J. Neuroeng. Rehabil.* 9, 1–17.

Pedregosa, F., Weiss, R., and Brucher, M. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Ross, G. B., Dowling, B., Troje, N. F., Fischer, S. L., and Graham, R. B. (2018). Objectively differentiating movement patterns between elite and novice athletes. *Med. Sci. Sport. Exerc.* 50, 1457–1464. doi: 10.1249/MSS.0000000000001571

Saeys, Y., Abeel, T., and Van de Peer, Y. V. (2008). "Robust feature selection using ensemble feature selection techniques," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (Berlin: Springer), 313–325. doi: 10.1007/978-3-540-87481-2_21

Sgro, F., Mango, P., Pignato, S., Schembri, R., Licari, D., and Lipoma, M. (2017). Assessing standing long jump developmental levels using an inertial measurement unit. *Percept. Mot. Skills* 124, 21–38. doi: 10.1177/0031512516682649

Smith, C. A., Chimera, N. J., Wright, N. J., and Warren, M. (2013). Interrater and intrarater reliability of the functional movement screen. *J. Strength Cond. Res.* 27, 982–987. doi: 10.1519/jsc.0b013e3182606df2

Smith, R. M., and Spinks, W. L. (1995). Discriminant analysis of biomechanical differences between novice, good and elite rowers. *J. Sport Sci.* 13, 377–385. doi: 10.1080/02640419508732253

Troje, N. F. (2002). Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *J. Vis.* 2, 371–387. doi: 10.1167/2.5.2

Wittmann, F., Lambercy, O., and Gassert, R. (2019). Magnetometer-based drift correction during rest in IMU arm motion tracking. *Sensors* 19, 1312–1330. doi: 10.3390/s19061312

Young, A. D. (2010). "From posture to motion: the challenge for real time wireless inertial motion capture," in *Proceedings of the Fifth International Conference on Body Area Networks*, Corfu, 131–137.

Young, C., and Reinkensmeyer, D. J. (2014). Judging complex movement performances for excellence: a principal components analysis-based technique applied to competitive diving. *Hum. Mov. Sci.* 36, 107–122. doi: 10.1016/j.humov.2014.05.009

Zago, M., Sforza, C., Dolci, C., Tarabini, M., and Galli, M. (2019). Use of machine learning and wearable sensors to predict energetics and kinematics of cutting maneuvers. *Sensors* 19, 3094–3105.

Zhou, H., and Hu, H. (2008). Human motion tracking for rehabilitation — A survey. *Biomed. Signal Process. Control* 3, 1–18. doi: 10.1016/j.bspc.2007.09.001

A Decision Support System to Facilitate Identification of Musculoskeletal Impairments and Propose Recommendations Using Gait Analysis in Children With Cerebral Palsy

Kohleth Chia[1], Igor Fischer[1], Pam Thomason[2], H. Kerr Graham[2,3] and Morgan Sangeux[1,2,3]*

[1] Murdoch Children's Research Institute, Melbourne, VIC, Australia, [2] The Royal Children's Hospital Melbourne, Parkville, VIC, Australia, [3] Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia

The identification of musculoskeletal impairments from gait analysis in children with cerebral palsy is a complex task, as is formulating (surgical) recommendations. In this paper, we present how we built a decision support system based on gait kinematics, anthropometrics, and physical examination data. The decision support system was trained to learn the association between these data and the list of impairments and recommendations formulated historically by experienced clinicians. Our aim was 2-fold, train a computational model that would be representative of data-based clinical reasoning in our center, and support new or junior clinicians by providing pre-processed impairments and recommendations with the associated supportive evidence. We present some of the challenges we faced, such as the issues of dimensionality reduction for kinematic data, missing data imputations, class imbalance and choosing an appropriate model evaluation metric. Most models, i.e., one model for each impairments and recommendations, achieved a weighted Brier score lower than 0.20, and sensitivity and specificity greater than 0.70 and 0.80, respectively. The results of the models are accessible through a web-based application which displays the probability predictions as well as the (up to) 5 best predictors.

Keywords: decision support system, gait analysis, cerebral palsy, orthopaedics, random forest, peadiatrics

INTRODUCTION

Cerebral palsy (CP) refers to a group of disorders due to a brain lesion that occurred shortly before or after birth (Graham et al., 2016). CP is the most common cause of physical disability in children, with a prevalence of around 2.5 per 1000 births in developed countries (McIntyre, 2018). Due to the brain lesion, secondary musculoskeletal impairments often develop and worsen during childhood and adolescence. Clinical and physical examinations as well as three-dimensional gait analysis (3DGA)

are utilized to determine how the musculoskeletal impairments affect the capacity of an individual to walk (Wren et al., 2011a,b).

Identifying neuro-musculoskeletal problems that impact the walking function of children with CP is a difficult process that involves multiple components. The diagnostic matrix (Davids et al., 2004) includes clinical history and diagnosis, classification and functional scales, physical examination, such as passive range of joint motion and muscle strength, and 3DGA which provides the kinematics and kinetics of the lower limb joints during walking as curves. In the latter analysis, the effects of the musculoskeletal impairments on gait may be detected from abnormal features present in the kinematic and kinetic curves. Features that might be interpreted include the magnitude and waveform of the different curves, the difference between the patient's curve and those from healthy individuals or the differences between the left and right limb curves. The final surgical recommendations also incorporate diagnostic imaging and examination under anesthesia (Davids et al., 2004).

A clinician needs to invest significant time and energy into assimilating all the available clinical information to determine the impairments that affect a child's capacity to walk. For example, 3DGA feature interpretation is not straightforward because: (i) one feature in one curve may relate to several impairments, (ii) there may be several features corresponding to several impairments, (iii) one impairment may lead to abnormal features in several other curves, and (iv) kinematics features may be primarily related to an impairment, compensatory or obligatory because of an impairment. Interpretation therefore requires analyzing the features observed simultaneously in the 24 kinematic and kinetic graphs in conjunction with results from the physical examinations and information from the other assessments.

We developed a decision support system to facilitate the process of identifying musculoskeletal impairments from the typical gait analysis assessments. Both the complexity of the identification process and the necessity to support this process using computerized tools has been recognized in the past, as early as the 1990s (Weintraub et al., 1990; Johnson et al., 1996; Simon, 2004), however, with limited uptake in practice by the clinical gait analysis community. Work to design decision support systems for gait analysis is continuing (Wagner et al., 2019). Machine learning concepts and algorithms are now applied to a range of tasks pertaining to gait analysis thanks to advancement in computation power and widespread use of databases to store clinical and gait analysis data. These tasks range from automated classification (Lai et al., 2009; Rozumalski and Schwartz, 2009; Sangeux et al., 2015) to predicting outcomes (Schwartz et al., 2013) to data-driven optimal clinical decision making (Ries et al., 2014).

The core of the decision support system we developed is a group of predictive models which use physical examination and kinematic data to identify impairments and surgical recommendations. The models were trained on a historical dataset to predict, based on the current clinical findings for the current child, what the impairment list and surgical recommendations would have been identified by the clinicians in the past. In other words, the predictive models are concerned

about replicating the behaviors of clinicians in the past and may be viewed as an objective and probabilistic storage of clinical reasoning. Our objective was to develop models able to support a clinician by providing an answer to the question: "What my-past-self, or experienced predecessors in this center, would have decided based on similar data?" As such, these models may be particularly useful to clinicians with less clinical experience.

In our gait laboratory, the clinical decision-making process is separated into two components. Firstly, we identify the impairments, objectively from the clinical data. Secondly, management options which may include orthopaedic surgery are selected. We therefore developed this decision support system for both components. However, instead of being sequential, we developed these two decision support components independently of each other. The surgical recommendation system depends on the clinical data only not the (machine) identified impairments. The reason for making these independent is so that the quality of the surgical recommendation system will not be limited by the quality of the impairment identification system. Thus, even if the system fails to return a correct impairment list, there is still a possibility for the second component of the system to make the correct surgical recommendation.

This article explains how we developed and evaluated the impairment identification and surgical recommendation decision support systems. The decision support system was designed in tight partnership with clinicians in our center, including how the outputs are presented.

MATERIALS AND METHODS

Dataset for Impairment Modeling
Data
We collated 3DGA records from the Hugh Williamson Gait Analysis Laboratory (HWGAL) with the following inclusion criteria:

(1) A diagnosis of CP as determined by appropriate clinicians, and registration on the state-wide CP Register.
(2) Data collected from 2008 onward (Prior to 2008, a set of identified impairments was not a mandatory reporting requirement).
(3) 3DGA data must contain at least one barefoot, unassisted walk. No other condition (e.g., orthosis) was included. Typical data collection includes six walking trials with at least three with kinetics data, however, sometimes only the most representative walking trial was uploaded to the report. Representative trials were chosen visually before 2015 and computationally after 2015 (Sangeux and Polak, 2015).
(4) Physical Examination data must be available.
(5) The 3DGA report must list a clear set of identified impairments.

This procedure led to 689 3DGA records being used, stemming from 423 children (mean (SD) age: 10 years (2.4 years), range: 2–21 years). However, we modeled each side (left and right) separately, hence doubling the data to 1378 records. We
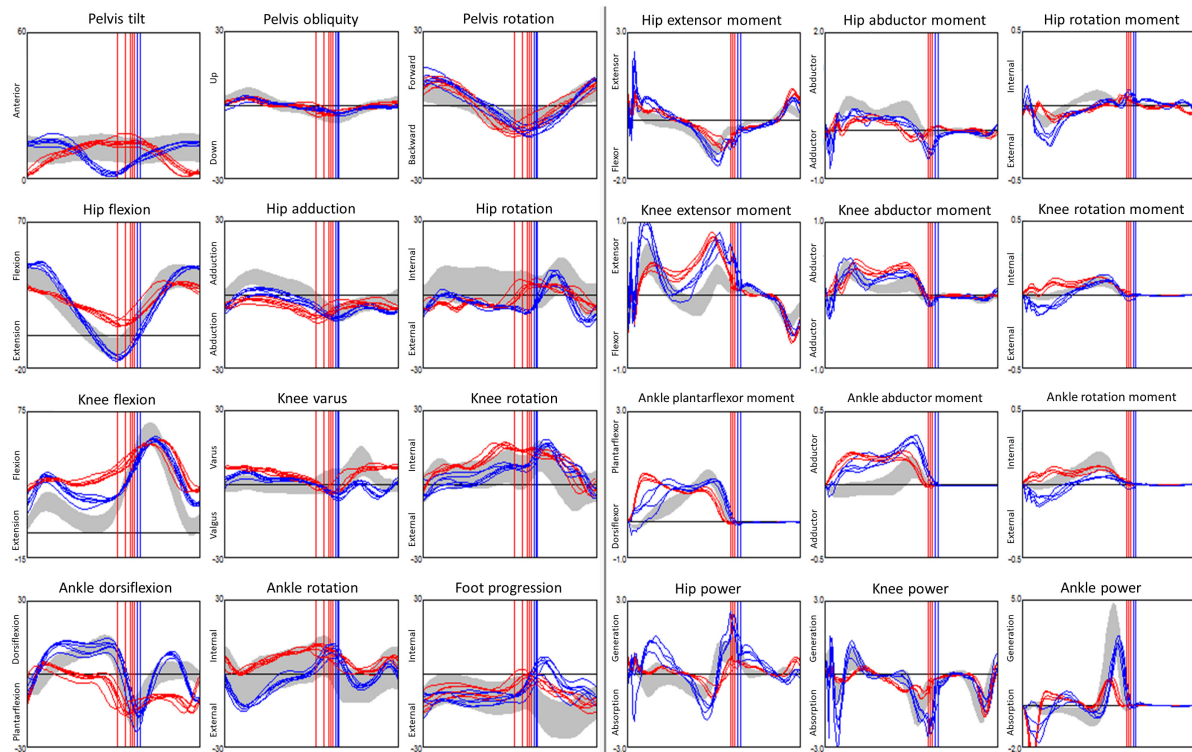
FIGURE 1 | Example of a dataset from gait analysis: 24 kinematic (3 columns from the left, angles in °) and kinetic (3 columns from the right, moments are internal and in N.m/kg bodyweight, powers are in W/kg bodyweight) graphs for a child with CP. Kinematics and kinetics data are plotted along time *t*, in % of the gait cycle. Data from 5 walks were superimposed, the left limb is in red, the right limb in blue, data from typically developed children displayed as a gray band. Vertical lines denote the timing of the ipsilateral foot off.

initially attempted modeling on the individual level, but we found the result to be worse than modeling on each side. We suspect the main reason is that we do not have enough data points to support accurate estimation of many predictors, so when we model both sides together, which effectively doubles the number of predictors, the final predictive accuracy drops. This has been observed previously (Trunk, 1979).

Predictors

The predictors that we used can be grouped into two sets: kinematics and physical examination data.

The kinematics were collected using the following protocol. The children were equipped with the Plug-in-Gait marker set (Vicon, Oxford Metrics Group) by registered physiotherapists. During the 3DGA session, children walked barefoot at their self-selected speed and chose their cadence freely. The three-dimensional marker trajectories were obtained using Vicon motion capture systems including 10 cameras (Oxford Metrics Group, United Kingdom) recording at 100 Hz. The foot strike events, and ground reaction force were captured from 2 (before 2009) to 6 (after 2009) AMTI force plates (AMTI Inc, United States) embedded in the floor. Force plate signals were sampled at 2000 Hz. Lower limb kinematics and kinetics were calculated with Plug-in-Gait in Nexus software (VICON, Oxford Metrics Group) after filtering marker trajectories

(Woltring, 1986). We used kinematic data from the lower limbs only namely: Pelvis (sagittal, coronal, transverse), Hip (sagittal, coronal, transverse), Knee (sagittal, coronal, transverse), Ankle (sagittal, transverse), and Foot Progression angle (transverse). All kinematics and kinetics data were normalized to the gait cycle, each curve was described from 101 points, one every % of the gait cycle (**Figure 1**).

Physical examination data was collected using standard protocols published elsewhere (Keenan et al., 2004; Thomason et al., 2014). Not all physical examination measurements were collected for all children due to difficulty with compliance or physical ability. In our model, we excluded any predictors that were not collected for at least 90% of the children.

Table 1 lists the physical examination predictors that we used in the model.

Impairments

We modeled frequently occurring impairments, with impairments that were listed with at least 100 occurrences. Impairments were extracted, tabulated, and added to the database from the text available in the clinical reports. In 2007, clinicians at our center agreed to follow a template to report clinical interpretation of gait analysis. The reporting policy was in line with the concept of impairment focused interpretation which was first described by Baker (2013). Specifically, the

Category	Measurements
Anthropometric	Age, Height, Weight
Strength	Knee extensors, Quadriceps lag, Abdominals, Knee flexors, Hip extensors, Hip abductors, Dorsiflexors, Plantarflexors, Invertors, Hip flexors
ROM/Spasticity (Tardieu fast)	True popliteal angle, Popliteal angle, Dynamic popliteal angle (fast), Dorsiflexion (knee flexed), Dorsiflexion (knee extended), Dynamic dorsiflexion (fast), Hip abduction (knee extended), Hip extension, Duncan-Ely (slow), Duncan-Ely (fast), Hip Internal rotation, Hip external rotation, Selective Motor Control at the ankle
Bone	Femoral anteversion (trochanteric prominence test), Tibial torsion (Bimalleolar axis), Thigh heel angle, Foot posture - Midfoot, Forefoot, Hindfoot sagittal, Hindfoot coronal

Spasticity measurements were according to Tardieu as the angle of arrest under fast passive range of movement (ROM).

TABLE 2 | List of impairments with at least 100 occurrences and number of occurrences.

Impairments	Number of occurrences
Hamstring spasticity	497
Gastrocnemius spasticity	434
Increased femoral neck anteversion	383
Soleus spasticity	358
Gastrocnemius contracture	342
Increased external tibial torsion	338
Rectus femoris spasticity	243
Soleus contracture	237
Knee fixed flexion deformity	177
Gluteal weakness	129
Soleus weakness	128
Hip fixed flexion deformity	125
Hamstring contracture	117
Gastrocnemius weakness	107

TABLE 3 | Physical examination measurements used for predicting surgical recommendation in the model.

Category	Measurements
Anthropometric	Age, Height, Weight
Strength	Nil
ROM/Spasticity	Hip abduction (knee extended), Dorsiflexion (knee extended), Duncan-Ely (fast), True popliteal angle, Popliteal angle, Hip internal rotation, Hip external rotation
Bone	Tibial torsion (Bimalleolar axis)

TABLE 4 | Surgical procedures conducted at least 100 times.

Surgeries	Number of times conducted
Femoral derotation osteotomy	159
Semitendinosus transfer	143
Gastrocnemius lengthening (Strayer)	142
Adductor longus lengthening	128

impairments listed in the reports were those deemed to impact gait function by the clinician who conducted the 3DGA and completed the interpretation.

Table 2 lists such impairments.

Dataset for Surgical Recommendations
Data

The inclusion criteria for data used for building the dataset for surgery modeling was the same as for impairment modeling, (1 to 4 listed above) but without the requirement for an impairment list and the addition of:

(6) The child must have undergone surgery within 2 years of the recommendations from the 3DGA report.

This led to 384 3DGA analysis records, stemming from 309 children being included. Again, we modeled each side separately which then doubled the data record number to 618.

Predictors

We went through a similar procedure of removing measurements that were not collected for at least 90% of the children. **Table 3** lists the physical examination predictors that we used in this model.

Surgical Procedures

Table 4 lists the surgical procedures included in the model, each of which were conducted at least 100 times.

Modeling
Kinematic Feature Extraction

Forty-nine kinematic binary features, all of the form "has X or has not X," were extracted from the raw kinematic curves. We derived the feature definitions by first using a published DELPHI consensus study (Nieuwenhuys et al., 2016) as a starting point, and then we conducted our own discussion session with clinicians to fine tune the features into the final form as shown in **Table 5**.

Computationally, these features are detected by applying some function to both the curve under consideration and a set of standard kinematic curves measured from typically developing children (Pinzone et al., 2014; Sangeux et al., 2016), and then making some comparison using means and standard deviation (SD). For example, the feature "Increased Hip extension at Mid Stance" has a definition "Mean angle at $t \in [20,45] < 1$ SD of typical mean." This translates to a four-step procedure:

(1) Calculate the mean angle for the curve under consideration at $t \in [20,45]$.

TABLE 5 | Features extracted from raw kinematic curves.

Structure	Plane	Features	Definition
Pelvis	Sagittal	Increased ROM (double bump)	● ROM > 2 SD of typical ROM, and
			● period is 2, and ● difference between L&R is <0.25 ROM, and
			● Correlation with our reference double bump curves >0.8.
		Decreased Pelvic Tilt	● Mean angle < 1 SD of typical mean.
		Decreased Pelvic Tilt + Increased ROM	● Decreased Pelvic Tilt, and
			● ROM > 2 SD of typical ROM.
		Increased Pelvic Tilt	● Mean angle > 1 SD from typical mean.
		Increased Pelvic Tilt + Increased ROM	● Increased Pelvic Tilt, and
			● ROM > 2 SD of typical ROM.
		Unilateral Bump	● ROM > 2 SD of typical ROM, and
			● Not double bump.
	Coronal	Increased Pelvic ROM	● ROM > 2 SD of typical ROM
		Pelvic Elevation/Depression	● Mean difference between L&R > 1 SD of typical difference.
	Transverse	Increased Pelvic Rotation ROM	● ROM > 2 SD of typical ROM
		Pelvic Pro / Retraction	● Mean difference between L&R > 1 SD of typical difference.
		Reversed ROM	● Correlation with reference reversed ROM curves >0.8.
Hip	Sagittal	Decreased Hip Flexion at Initial Contact	● Angle at $t = 0$ < 2 SD of typical angle.
		Hip Extension Deficit	● Mean angle in stance > 1 SD of typical angle, and
			● ROM < 2 SD of typical ROM.
		Hip Hyper-Flexion	● Mean angle in stance within 1 SD of typical range, and
			● Peak angle in swing > 2 SD of typical peak.
		Increased Hip extension at Mid Stance	● Mean angle at $t \in [20,45]$ < 1 SD of typical mean.
		Increased Hip Flexion	● Mean angle in stance > 1 SD of typical mean, and
			● All angles > 0.
		Increased Hip Flexion + Decreased ROM	● Increased Hip Flexion, and ● ROM < 2 SD of typical ROM.
	Coronal	Excessive Hip Abduction	● Mean angle in stance < 1 SD of typical mean, and
			● Mean angle in swing < 1 SD of typical mean.
		Excessive Hip Abduction in Swing	● Mean angle in swing < 1 SD of typical mean.
		Excessive Hip Adduction	● Mean angle in stance > 1 SD of typical mean, and
			● Mean angle in swing > 1 SD of typical mean.
		Hip Adduction in Stance	● Mean angle in stance > 1 SD of typical mean.
	Transverse	Hip External Rotation	● Mean angle < 1 SD of typical mean.
		Hip Internal Rotation	● Mean angle > 1 SD of typical mean.
		Increased Hip Internal Rotation at Late Stance	● mean angle in $t \in [40,60]$ > 1 SD of typical mean, and
			● peak occurs before $t = 70$, and
			● no pit in $t \in [20,80]$
Knee	Sagittal	Reduced Flexion at Loading	● Mean angle in $t \in [0,20]$ < 1 SD of typical mean.
		Decreased Peak Knee Flexion	● Peak in swing < 2 SD of typical peak.
		Delayed + Decreased Peak Knee Flexion	● Peak occurs after $t = 75$, and
			● Decreased Peak Knee Flexion.
		Delayed + Increased Peak Knee Flexion	● Peak occurs after $t = 75$, and ● Peak in swing > 2 SD of typical peak
		Delayed Peak Knee Flexion	● Peak occurs after $t = 75$.
		Knee Flexion in Mid Stance	● Mean angle in $t \in [20,45]$ > 1 SD of typical mean.
		Knee Hyperextension	● Mean angle in $t \in [20,45]$ < 1 SD of typical mean.
		Increased Flexion at Initial Contact	● Angle at $t = 0$ > 2 SD of typical angle.
		Increased flexion at Initial Contact+ Early Knee Extension	● Increased flexion at initial contact, and
			● Pit occurs before $t = 25$, and
			● Difference between angle @ IC and at pit > 10, and
			● Min angle in $t \in [10,25]$ < mean+1 SD of typical angle
		Increased Peak Knee Flexion	● Peak angle in swing > 2 SD of typical peak.
Ankle	Sagittal	Reduced Dorsiflexion	● Mean angle in $t \in [0,50]$ < 1 SD of typical mean.
		Descending 2nd Rocker	● Angle at $t = 45$ minus the angle at $t = 20$ < −5, and ● Angle at $t = 45$ < 0

(Continued)

TABLE 5 | Continued

Structure	Plane	Features	Definition
		Dorsiflexion in Swing	• Mean angle in swing > 1 SD of typical mean.
		Foot Drop	• Mean angle in $t∈[80,100]$ < 1 SD of typical mean.
		Horizontal 2nd Rocker	• ROM in $t∈[20,45]$ < 5, and
			• Absolute slope of the same period <0.1, and
			• Angle at $t = 45$ < 0.
		Increased Dorsiflexion	• Mean angle at $t∈[20,45]$ > 1 SD of typical mean.
		Increased Max. Dorsiflexion	• Max angle in stance > 2 SD of typical stance.
		Increased Plantarflexion	• Mean angle in $t∈[20,45]$ < 1 SD of typical mean.
		Insufficient Pre-positioning	• Angle at $t = 100$ (final) > 2 SD of typical angle.
		No 1st Rocker	• Angle at $t = 1$ > angle at $t = 0$.
		Short 2nd Rocker	• Peak exists in $t∈[0,20]$, and
			• Slope in $t∈[20,45]$ < 0, and
			• Correlation with reference short 2nd rocker curves >0.8.
	Transverse	Ankle Internal Rotation	• Mean angle in stance > 1 SD of typical angle.
Foot Progression	Transverse	External Foot Progression (Wave) in Swing only	• ROM in swing > 2 SD of typical ROM, and
			• Correlation with referenced External foot progression curve >0.8.
		In-toe	• Mean angle in stance > 1 SD of typical mean.
		Out-toe	• Mean angle in stance < 1 SD of typical mean.

ROM, Range of motion; SD, Standard Deviation; R, Right; L, Left.

(2) Calculate the same for each individual curve in the reference (typically developed) dataset.

(3) Calculate the mean and SD of the mean angles calculated in step 2.

(4) If the statistics calculated in step 1 is less than the mean minus 1 SD (as calculated in step 3), then the feature is deemed present in the curve under consideration.

The R code to detect the default kinematic features and create custom-designed feature detectors is available here: https://github.com/Morgan435/gaitFeature/.

Most features in **Table 5** are defined over the entire gait cycle or a sub-phase of gait, e.g., stance or swing phase, or at $t = 0$. Markedly different walking speeds between the mean of the typically developing reference dataset and that of the subject under consideration may lead to some time-shift when the features are considered over a specific time period, e.g., $t∈[20,45]$ as above. In this case, it is possible to add a pre-processing step that apply dynamic time warping to the curves before features extraction. We did not deem necessary to include such a pre-processing step in this instance.

Missing Data Imputation

Missing data were imputed just prior to the model training step. Continuous variables were imputed by their median, and categorical ones were imputed by the most frequently appearing category. We acknowledge that such a simple imputation scheme imposed certain assumptions to the data missing mechanism, such as missing at random, which is unlikely true in clinical practice. But we did check that variables with high missing data did not rank highly in the variable importance matrix.

Models

There are many different machine learning algorithms that could be used. We have trained many different models (e.g., support vector machine, linear discriminant model, partial least square, naïve bayes), however, here we report the models with the best overall balance between ease-of-use, computation speed, and predictive performance.

For each impairment / surgical procedure, a standard random forest (Breiman, 2001), a stratified random forest (Kuhn, 2008), and a regularized logistic regression (with elastic net penalty, a.k.a. "glmnet" (Friedman et al., 2010) were fitted. The stratified random forest is like the standard one, except that each tree in the forest is trained on a balanced resample. That is, a resample where the number of instances in each class is equal. The reason for using the stratified version is to tackle class imbalance. However, such stratified resamples are essentially down sampling techniques, that is, throwing away instances of the majority class. Therefore, the cost of a balanced training is higher variance. Hence, the other models were still retained as feasible candidates.

The hyper-parameters were tuned according to the weighted Brier score as defined in the next section.

Model Evaluation Metric

One way to assess classification models is to see how well they predict the class of new observations. Metrics such as accuracy, precision, sensitivity, specificity are all designed to assess this aspect of the prediction (Altman and Bland, 1994). However, we advocate moving away from predicting the class to predicting the probability of class membership. This is because, first, unlike predicted class, the predicted probabilities also convey the uncertainty of the prediction. Most of us would probably view 51 and 99% chance of something happening very differently, but in the class prediction sense they would both be considered the

same positive prediction, so the uncertainty is lost. Secondly, unlike many artificial intelligence application (e.g., hand-written postcode recognition by the post office), the machine (model) in our clinical setting does not actually have to make any decision, as this responsibility lies with the clinician. Therefore, predicting class membership is in a sense one step more than what is required.

Hence, we decided to use the Brier score (Brier, 1950) as our model evaluation metric. The Brier score is defined as $\sum_i \left(\widehat{P}_i - O_i\right)^2 / N$, where $\widehat{P}_i$ denotes the predicted probability that child i had the impairment / surgery, O_i is 1 if the child i had the impairment / surgery and 0 otherwise, and N is the total number of children. Therefore, Brier score is the mean-squared-error equivalent for binary classification. Another advantage of using Brier score is that it is a proper scoring metric (Harrell, 2015), meaning its expectation is minimized by the true, unknown, probability.

The weighted Brier score is the weighted version where each term in the summation is weighted by the reciprocal of their true class frequency. The reason we used the weighted Brier score is to tackle the issue of class imbalance. By setting the weight to be the reciprocal of the class frequency, error in the minority class would be penalized more heavily, thus forcing us to choose a model that has a more balanced performance in both classes.

A null model which predicts a constant 50% regardless of predictors will achieve a Brier score (weighted or not) of 0.25. Therefore, any admissible model should achieve a score of less than 0.25.

Training and Validation

Training and validation were done using 25 out-of-bag bootstrap resample. That is, the entire dataset was used to recreate 25 bootstrap resamples, and the models were trained on each resample, and validated on the portion that did not appear in the corresponding training sample.

Probability Calibration

Because the predicted probability is our primary output, we carried out a further step of calibrating the prediction probability, using an isotonic regression. The procedure is as followed:

(1) From the previous steps (i.e., model training) collect the prediction made on the out-of-bag (i.e., validation) data, treat this as our new training dataset.
(2) Train an isotonic regression using the predicted probability as predictor, and the observed class as response variable. Such training is also done using the out-of-bag bootstrap resampling scheme.
(3) Using the predicted probability from the trained isotonic regression model as the final predicted probability that we report to the end user.

Probability calibration can sometimes improve the accuracy of the predicted probability, but sometimes impairs it. To ensure this step is beneficial, we once again computed the weighted Brier score and other secondary evaluation metrics and compare them with those from the uncalibrated models.

Explanation

We tried to make the model's prediction as transparent as possible, reasoning that the more transparent the model is, the more informative, and thus helpful, it will be to the end-user clinician. A clinician may not trust a black box that simply declares a certain child has impairment X without providing reason. But the same clinician will likely find a system which suggests a child might have impairment X with Y% certainty because of Z reason(s) more helpful and credible. The reasons for a decision support tool are at least 2-fold. Firstly, the clinician can compare and combine the model's confidence with their own. If a clinician already has strong belief that impairment X exists, then a model prediction of 60% confidence is most likely enough to reinforce that belief. On the contrary, if the clinician's prior belief was "highly unlikely," then even a 60% model confidence may not sway them in their belief. If, however, the model output is 90% confidence, then it might prompt the clinician to further investigate the issue. Secondly, the clinician has the information to examine the reason underlying the prediction, which empowers them to agree with or overrule the prediction. This makes a transparent and trustworthy computer system that has been the subject of many research papers (Ferri et al., 2002; Allahyari and Lavesson, 2011; Freitas, 2013; Castelvecchi, 2016; Lipton, 2016; Ribeiro et al., 2016).

We explain our prediction by firstly, outputting the predicted probability instead of predicted class. Secondly, we display the measurement values of the (up to) 5 most important predictors, where the importance of the predictors is judged by the trained models. In a typical random forest model, the importance metric is related to how much node impurity is reduced by a split on that variable, whereas in a regularized regression, it is related to the absolute magnitude of the coefficients. We say "up to" 5 because for some impairments (e.g., hip fixed flexion deformity), most of the result is explained by less than five predictors (e.g., hip extension ROM). In these cases, it is only meaningful to retain the predictors that have enough weight. Thirdly, we constructed a partial dependence plot, which shows how the models react to changes in each individual predictor, while holding all other predictors constant (or averaging over them). If our models were simply linear regression, these plots would simply be a straight line with the slope being to the estimated coefficient. But for algorithms such as random forest, the dependency can be non-linear.

RESULTS

Impairment Diagnosis Model

Table 6 reports the weighted Brier score of the various impairment diagnosis models on the validation sample. Recall that a Brier score can be loosely thought of as the mean squared error of the predicted probability (thus lower equals better), and a null model has a score of 0.25. For the calibrated random forest, we have also reported the mis-classification rate, sensitivity, and specificity. We have placed the models which fail to give a weighted Brier score or less than 0.25 at the bottom of the Table. Overall, the calibrated (standard) random forest is usually the

TABLE 6 | Weighted (out-of-bag/validation) Brier score for impairment diagnosis model, and the associated important variables.

Impairments	Glmnet		Stratified R.F.		Random Forest		Important Predictors
	Raw	Calibrated	Raw	Calibrated	Raw	Calibrated	By Random Forest
Hamstring Spasticity	0.186	0.187	0.174	0.169	0.172	0.166 Mis:0.21 Sens:0.71 Spec:0.83	1. Dynamic Popliteal Angle 2. Height 3. Weight 4. Popliteal Angle 5. True Popliteal Angle
Gastrocnemius Spasticity	0.216	0.219	0.198	0.200	0.196	0.195 Mis:0.23 Sens:0.68 Spec:0.80	1. Dynamic Dorsiflexion 2. Height 3. Weight 4. Increased Plantarflexion 5. Dorsiflexion (Knee Flexed)
Increased Femoral-Neck Anteversion	0.176	0.174	0.165	0.168	0.163	0.163 Mis:0.17 Sens:0.70 Spec:0.89	1. Hip Internal Rotation 2. Anteversion 3. Feature: Hip Internal Rotation 4. Weight 5. Hip External Rotation
Soleus Spasticity	0.245	0.246	0.223	0.226	0.220	0.218 Mis:0.22 Sens:0.60 Spec:0.82	1. Dynamic Dorsiflexion 2. Height 3. Weight 4. Dorsiflexion (Knee Extended) 5. Feature: Increased Plantarflexion
Gastrocnemius Contracture	0.140	0.152	0.132	0.147	0.132	0.144 Mis:0.13 Sens:0.74 Spec:0.91	1. Dorsiflexion (Knee Extended) 2. Dorsiflexion (Knee Flexed) 3. Dynamic Dorsiflexion
Increased External Tibial Torsion	0.194	0.197	0.183	0.188	0.183	0.181 Mis:0.17 Sens:0.71 Spec:0.86	1. Tibial Torsion 2. Thigh Heel Angle
Rectus Femoris Spasticity	0.192	0.194	0.176	0.179	0.166	0.158 Mis:0.09 Sens:0.83 Spec:0.92	1. Feature: Increased ROM (Pelvis) 2. Duncan-Ely (Fast) 3. Duncan-Ely (Slow) 4. Weight 5. Height
Soleus Contracture	0.165	0.179	0.158	0.174	0.161	0.174 Mis:0.11 Sens:0.71 Spec:0.92	1. Dorsiflexion (Knee Flexed) 2. Dorsiflexion (Knee Extended) 3. Dynamic Dorsiflexion
Hip Fixed Flexion Deformity	0.184	0.191	0.148	0.186	0.148	0.172 Mis:0.07 Sens:0.71 Spec:0.95	1. Hip Extension ROM

Models Which Are Not Better Than The Null Model (In Terms Of Weighted Brier Score)

Hamstring Contracture	0.296	0.310	0.262	0.291	0.249	0.245 Mis:0.08 Sens:0.62 Spec:0.94	1. Popliteal Angle 2. Dynamic Popliteal Angle 3. Weight 4. True Popliteal Angle 5. Height
Knee Fixed Flexion Deformity	0.277	0.292	0.276	0.298	0.276	0.281 Mis:0.12 Sens:0.64 Spec:0.90	1. Quadriceps Strength 2. True Popliteal Angle 3. Hip Abduction (Knee Extended) 4. Hip External Rotation 5. Popliteal Angle

(Continued)

TABLE 6 | Continued

Impairments	Glmnet		Stratified R.F.		Random Forest		Important Predictors
	Raw	Calibrated	Raw	Calibrated	Raw	Calibrated	By Random Forest
Gluteal Weakness	0.307	0.324	0.310	0.333	0.302	0.280 Mis:0.08 Sens:0.71 Spec:0.93	1. Dynamic Popliteal 2. Height 3. Weight 4. Popliteal Angle 5. True Popliteal Angle
Soleus Weakness	0.303	0.315	0.304	0.330	0.297	0.272 Mis:0.08 Sens:0.59 Spec:0.94	1. True Popliteal Angle 2. Hip Abduction (Knee Extended) 3. Hip Extension ROM 4. Weight 5. Anteversion
Gastrocnemius Weakness	0.331	0.342	0.335	0.361	0.325	0.298 Mis:0.07 Sens:0.59 Spec:0.94	1. True Popliteal Angle 2. Popliteal Angle 3. Dorsiflexion (Knee Extended) 4. Hip Extension ROM 5. Hip Abduction (Knee Extended)

For calibrated random forest, the mis-classification rate, sensitivity, and specificity are also reported. RF, random forest; ROM, Range of motion; Mis, mis-classification rate; Sen, sensitivity; and Spec, specificity.

best model in terms of having the lowest weighted Brier score. However, for all the muscle weakness impairments, all models failed to be better than the null model in the weighted Brier score.

Figure 2 shows the partial dependence plot for the impairment model (Friedman, 2001). The vertical axis is predicted probability that the impairment is present on the log scale. Some plots only show a straight line, as the predictors are binary (most likely one of the kinematic feature variables).

Surgical Recommendation Model

Table 7 presents the weighted Brier score for the surgical recommendation prediction model. Similarly, for the calibrated random forest, the mis-classification rate, sensitivity, and specificity are also provided. Overall, the two flavors of random forest perform very similarly, with the standard version being marginally better.

Figure 3 shows the partial dependence plot for the surgery model.

System Output

In order to give readers a sense of the decision support system, **Figure 4** shows the output from the impairment model. The green reflects the predicted probability that the impairment is present, and the orange is the complement to that. As can be seen, the predicted probability rather than predicted class is the primary output, and the explanation behind the prediction is reported.

DISCUSSION

We have developed two sets of models to facilitate identification of musculoskeletal impairments and surgical recommendations. These models are meant to reproduce the essence of past clinical reasoning to new data. The models were trained to associate kinematic features and physical examination data with the list of impairments from the clinical interpretation report and the list of surgical treatments. Our approach is original because it proposes an intermediate step, between the use of computerized tools using explicit knowledge, e.g., (Wagner et al., 2019), and data-driven approach to determine which treatment should be recommended (or should a particular treatment be recommended) by maximizing the likelihood of good outcome measures, e.g., (Ries et al., 2014). The primary output was the calibrated confidence that a certain impairment is present/absent, together with the values of the five most important predictors. Partial dependence plots were also supplied to assist understanding of the general reasoning of the models.

Physical examination measurements were the most important predictors for impairments, listed as impacting gait in the interpretation report, and surgical recommendations. The partial dependence plots in **Figures 2**, **3** provide some indications about the soft threshold values linking certain physical examination measurements with impairments. For many impairments and surgeries, these were the only predictors deemed important by the models. Of note, only a reduced set of physical examination measurements could be used for surgical recommendation models because of missing data. Only three impairments: increased femoral neck anteversion (feature: increased hip internal rotation), soleus spasticity (feature: increased plantarflexion) and rectus femoris spasticity (feature: double bump) included a kinematic feature as an important predictor. For models to predict surgical recommendations, only the femoral derotation osteotomy surgery included a
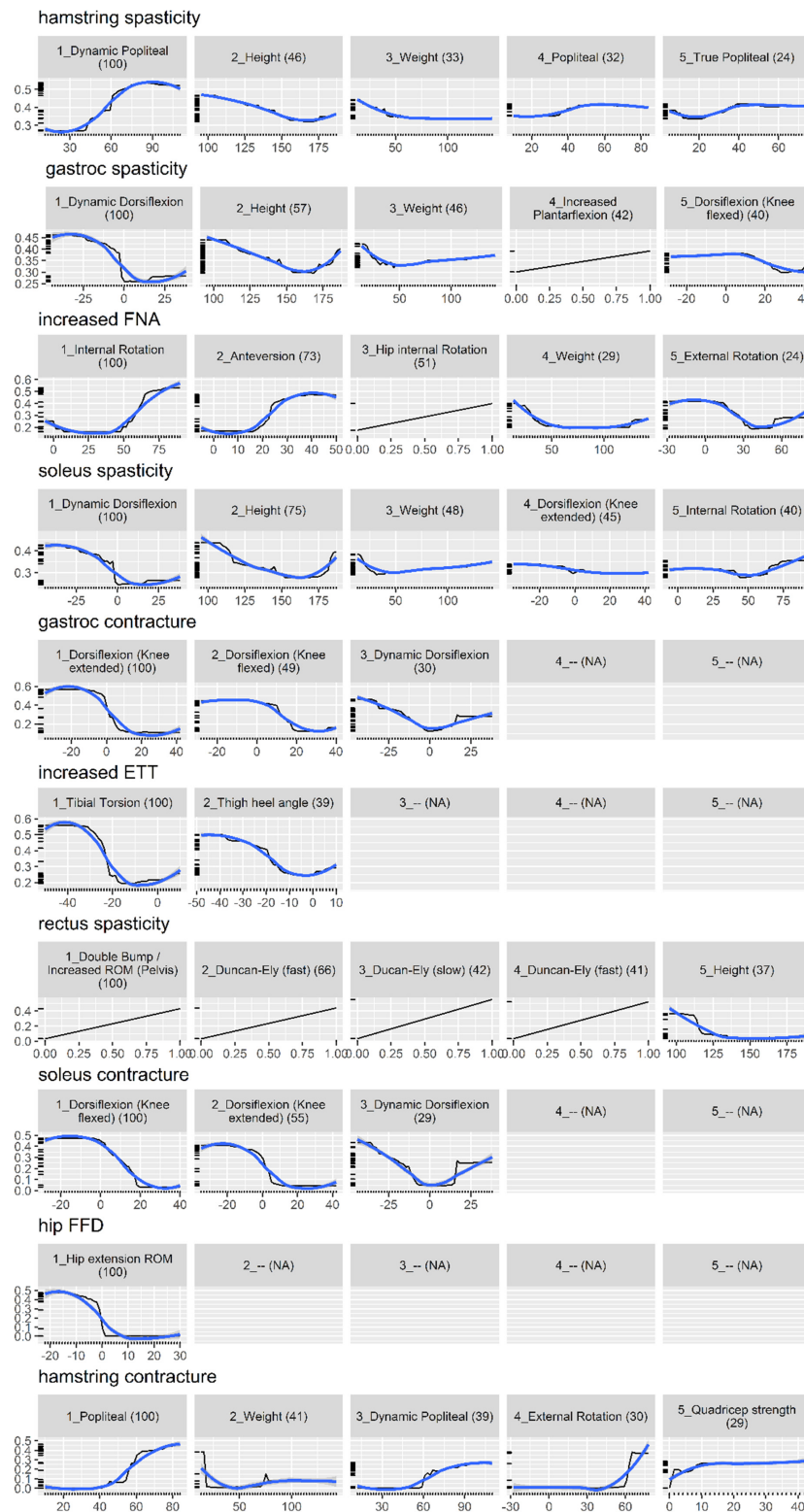
FIGURE 2 | Partial dependence plot for the impairment prediction model (black line, 30). The vertical axis is the predicted probability (between 0 and 1) that the impairment is present, on a log scale. The horizontal axis is the measurement for the predictor. The (blue) smooth line is added for visual aid. For some impairments, there are less than five important predictors, resulting in blank panels. Finally, the number in parenthesis indicates the importance of that predictor, relative to the most important predictor (so it always starts from 100 and decreases).
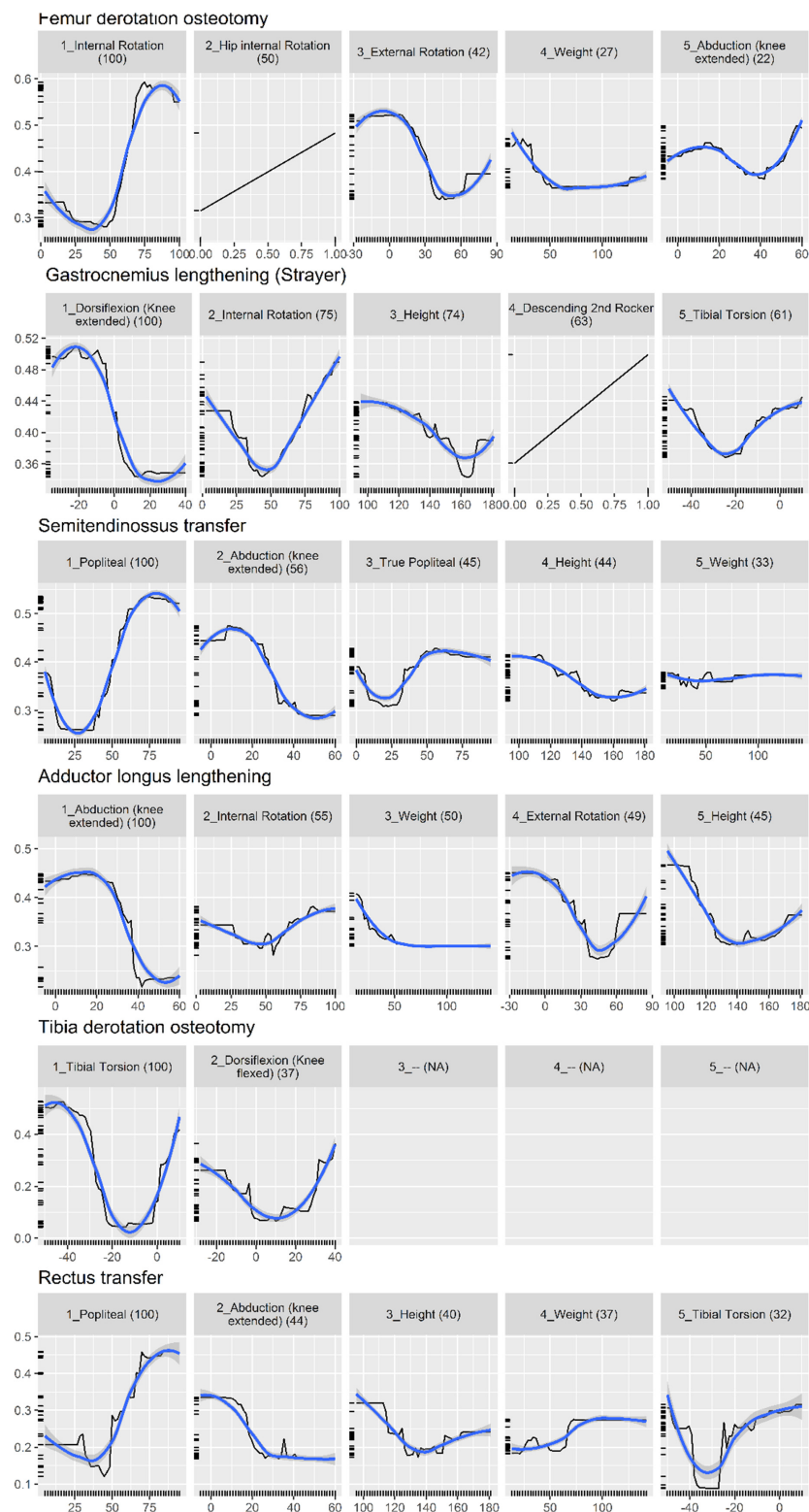
FIGURE 3 | Partial dependence plot for surgery prediction model (black line, 30). The vertical axis shows the predicted probability (between 0 and 1) that the surgery is needed, on a log scale. The horizontal axis is the measurement for the predictor. The (blue) smooth line is added for visual aid. For some surgeries, there are less than five important predictors, resulting in blank panels. Finally, the number in parenthesis indicates the importance of that predictor, relative to the most important predictor (so it always starts from 100 and decreases).

TABLE 7 | Weighted Brier score surgery recommendation models.

Surgery	Glmnet		Stratified R.F.		Random Forest		Important predictors
	raw	calibrated	Raw	calibrated	raw	calibrated	by random forest
Femur derotation osteotomy	0.173	0.172	0.163	0.163	**0.163**	0.164 mis:0.22 sens:0.78 spec:0.78	1. Internal Rotation 2. Feature: Hip internal Rotation 3. External Rotation 4. Weight 5. Abduction (knee extended)
Gastrocnemius lengthening (Strayer)	0.226	0.227	0.202	0.207	**0.201**	0.204 mis:0.24 sens:0.67 spec:0.80	1. Dorsiflexion (Knee extended) 2. Height 3. Internal Rotation 4. Abduction (knee extended) 5. Dorsiflexion (Knee flexed)
Semitendinosus transfer	0.196	0.197	0.177	0.176	0.175	**0.172** **mis:0.20** **sens:0.73** **spec:0.83**	1. Popliteal 2. Abduction (knee extended) 3. True Popliteal 4. Height 5. Weight
Adductor longus lengthening	0.240	0.241	**0.223**	0.226	0.226	0.223 mis:0.24 sens:0.64 spec:0.79	1. Abduction (knee extended) 2. Internal Rotation 3. External Rotation 4. Weight 5. Height
Tibial derotation osteotomy	0.233	0.235	0.209	0.224	**0.208**	0.215 mis:0.14 sens:0.67 spec:0.89	1. Tibial Torsion 2. Dorsiflexion (Knee flexed)
Rectus transfer	0.255	0.257	0.226	0.250	0.220	**0.219** **mis:0.13** **sens:0.75** **spec:0.88**	1. Popliteal 2. Tibial Torsion 3. Weight 4. Abduction (knee extended) 5. Height

For the calibrated random forest, mis-classification rate, sensitivity, and specificity are also provided. The right most column lists the top five important predictors as judged by the random forest. RF, random forest; ROM, Range of motion; Mis, mis-classification rate; Sen, sensitivity; and Spec, specificity.

kinematic feature (increased hip internal rotation) as an important predictor.

These results were unexpected and somewhat contrary to our opinion about the influence of lower limb kinematics on clinical decision making. One explanation may be that kinematic data, as curves, are difficult to include in a predictive statistical model because of their high dimensionality (101 points times 15 curves). We proposed a dimension reduction process that summarizes the kinematics curves into a set of kinematic features. We may have lost some important information during this process. However, we initially also treated the curves directly and obtained worse weighted Brier scores for the various models (Chia et al., 2017). Another explanation may be that kinematic data are essential to confirm the impact of a physical examination measurement on the gait pattern, but that decisions ultimately hinge upon physical examination measurements.

There are arguably more advanced methods to achieve transparent predictions than those we implemented. For example, one approach may be to use a comprehensible model, such as a (single) decision tree to approximate the behavior of the models, either globally (Domingos, 1998; Martens et al., 2007) or locally (Ribeiro et al., 2016). The advantage of this approach is that the full decision pathway is explained, instead of our current approach which simply returns the values of some predictors which are deemed important. However, the problem of all approximation is that it loses predictive power. In addition, the fact that we calibrate our model further complicates the process of explaining the prediction. Another area of challenge is the class imbalance problem. The class imbalance is both in terms of occurrence, as well as cost of error. For imbalance in occurrence, we tried to tackle them with stratified random forest, but the overall performance was not better than the standard version. For imbalance in cost, we could have trained cost-sensitive models, which penalize error in both classes differently, and according to a pre-specified cost. Eliciting such cost structure is non-trivial (e.g., what is the cost of missing an impairment or recommending an unnecessary surgery?) but would be a worthwhile pursuit.

Our predictive models seek to imitate the past behavior of the clinicians at our center. The limitation of this approach
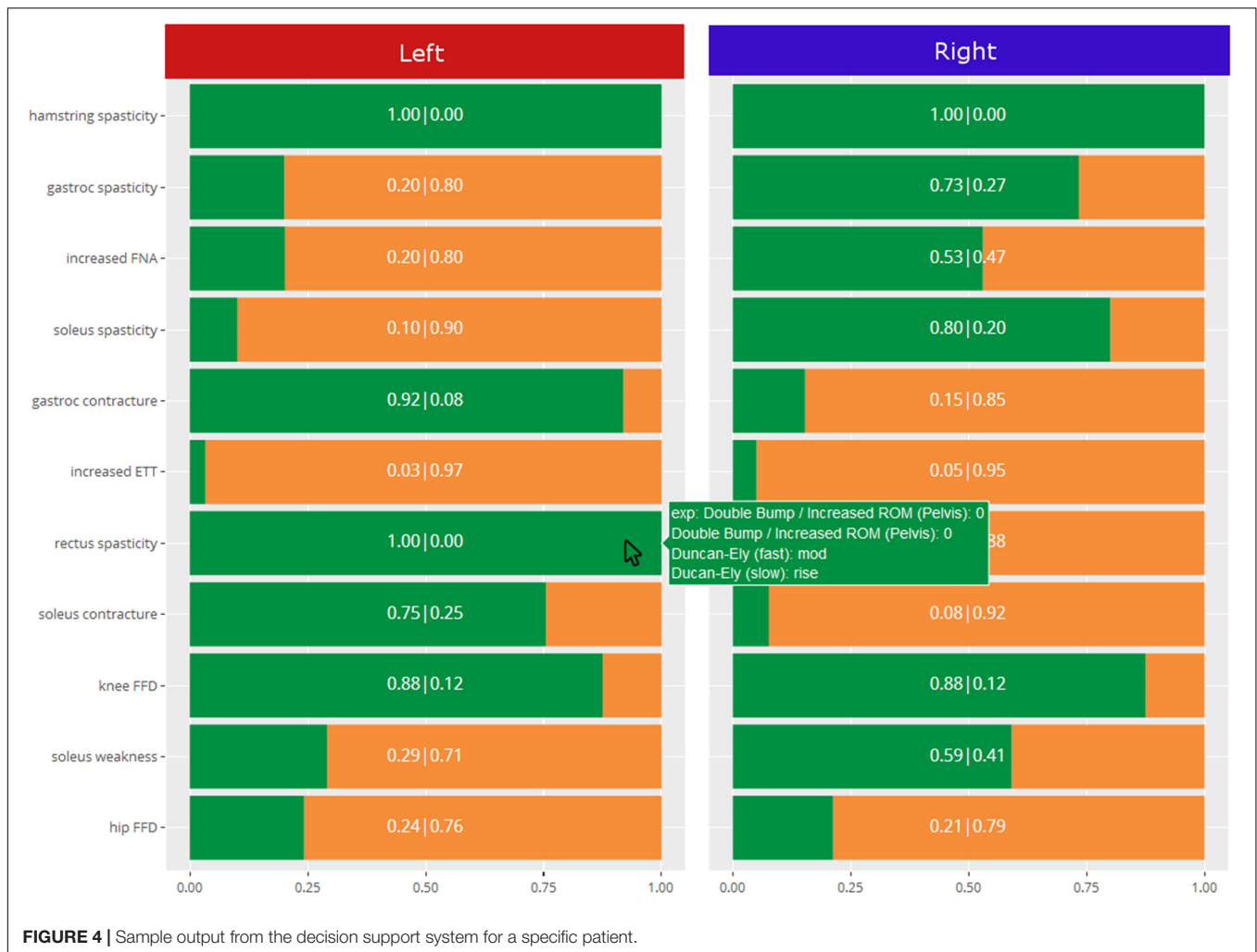
FIGURE 4 | Sample output from the decision support system for a specific patient.

is that past errors will also be learnt by the model, which is why it is important to constantly refresh and update the model with new data, which will dilute the influence of past decisions.

Another limitation is that these predictive models are, by intent and because the models were trained on data and clinical reasoning from a single center, not generalizable to other centers. However, there is also a benefit to this approach. In our gait laboratory, data are discussed at a team reporting meeting where recommendations are made. The models capture the collective, therefore representative, reasoning of the clinical team. If each center trains their own set of models, the comparison of these models would highlight the similarities and differences between the clinical practices objectively. For example, it would be possible to perform virtual visit(s), whereby the same clinical and gait analysis data would be fed to the models from different centers, likely leading to different conclusions (Noonan et al., 2003). This would initiate fruitful discussion about the rationale behind the differences in reasoning and would allow the comparison of outcomes drawn from independent samples. This could be a key element in making

progress toward the search for evidence-based optimal treatment recommendations.

CONCLUSION

We presented a decision-support system able to propose a list of impairments and surgical recommendations based on past decisions and gait analysis datasets. Machine learning models were trained and validated to predict the probability that clinicians, experienced in the interpretation of gait analysis data in children with CP, recommend an impairment or a surgical procedure. The random forest algorithm provided the best evaluation metrics (weighted Brier score) in most cases. Overall, the models achieved a weighted Brier score lower than 0.20, and sensitivity and specificity greater than 0.70 and 0.80, respectively. Once trained, these models collectively store the relationship between clinical decisions and gait data at our centre. The implementation of similar models in other center would facilitate objective comparison of clinical decision making, or "philosophy," between centers.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. The dataset is composed of patient sensitive information.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Royal Children's Hospital Melbourne Ethics Committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KC, IF, and MS contributed to the conception of the study and development of the machine learning models and infrastructure. KC performed the statistical analyses. KC and MS wrote the first draft of the manuscript, with sections contributed by PT and HG. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Allahyari, H., and Lavesson, N. (2011). User-oriented Assessment of Classification Model Understandability. *Front. Artif. Intell. Appl.* 227, 11–19. doi: 10.3233/978-1-60750-754-3-11

Altman, D. G., and Bland, J. M. (1994). Statistics Notes: Diagnostic tests 1: Sensitivity and specificity. *BMJ* 308:1552. doi: 10.1136/bmj.308.6943.1552

Baker, R. (2013). *Measuring Walking: A Handbook of Clinical Gait Analysis.* New Jersey: Wiley.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brier, G. W. (1950). Verification of forecasts expersses in terms of probaility. *Mon. Weather Rev.* 78, 1–3. doi: 10.1126/science.27.693.594

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature* 538, 20–23. doi: 10.1038/538020a

Chia, K., Fischer, I., Thomason, P., Graham, K., and Sangeux, M. (2017). O50: Is it feasible to use an automated system to recommend orthopaedic surgeries?. *Gait Posture* 57:89. doi: 10.1016/j.gaitpost.2017.06.304

Davids, J. R., Ounpuu, S., DeLuca, P. A., and Davis, R. B. III (2004). Optimization of walking ability of children with cerebral palsy. *Instr. Course Lect.* 53, 511–522.

Domingos, P. (1998). Knowledge discovery via multiple models. *Intell. Data Anal.* 2, 187–202. doi: 10.3233/IDA-1998-2303

Ferri, C., Hern ndez-Orallo, J., and Ram rez-Quintana, M. (2002). From Ensemble Methods to Comprehensible Models. *Discov. Sci.* 2534, 223–234.

Freitas, A. A. (2013). Comprehensible Classification Models – a position paper. *ACM SIGKDD Explor. Newsl.* 15, 1–10. doi: 10.1145/2594473.2594475

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Graham, H. K., Rosenbaum, P., Paneth, N., Dan, B., Lin, J.-P., Damiano, D. L., et al. (2016). Cerebral palsy. *Nat. Rev. Dis. Prim.* 2:15082. doi: 10.1038/nrdp.2015.82

Harrell, F. E. (2015). *Regression Modeling Strategies.* Netherland: Springer.

Johnson, K. A., Denning, R., Smith, P. J., Smith, J. W., and Simon, S. R. (1996). Dr. Gait III – A multimedia system for gait analysis. *Gait Posture* 4, 195–196. doi: 10.1016/0966-6362(96)80634-5

Keenan, W. N., Rodda, J., Wolfe, R., Roberts, S., Borton, D. C., and Graham, H. K. (2004). The static examination of children and young adults with cerebral palsy in the gait analysis laboratory: technique and observer agreement. *J. Pediatr. Orthop. Part B* 13, 1–8. doi: 10.1097/01202412-200401000-00001

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Lai, D. T. H. H., Begg, R. K., and Palaniswami, M. (2009). Computational intelligence in gait research: a perspective on current applications and future challenges. *IEEE Trans. Inf. Technol. Biomed.* 13, 687–702. doi: 10.1109/TITB.2009.2022913

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ICML Work. Hum. Interpret. Mach. Learn.* 16:3233231.

Martens, D., Baesens, B., Van Gestel, T., and Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *Eur. J. Oper. Res.* 183, 1466–1476. doi: 10.1016/j.ejor.2006.04.051

McIntyre, S. (2018). The continually changing epidemiology of cerebral palsy. *Acta Paediatr. Int. J. Paediatr.* 107, 374–375. doi: 10.1111/apa.14232

Nieuwenhuys, A., Ounpuu, S., Van Campenhout, A., Theologis, T., De Cat, J., Stout, J., et al. (2016). Identification of joint patterns during gait in children with cerebral palsy: A Delphi consensus study. *Dev. Med. Child Neurol.* 58, 306–313. doi: 10.1111/dmcn.12892

Noonan, K. J., Halliday, S., Browne, R., O'Brien, S., Kayes, K., and Feinberg, J. (2003). Interobserver variability of gait analysis in patients with cerebral palsy. *J. Pediatr. Orthop.* 23, 279–287. doi: 10.1097/00004694-200305000-00001

Pinzone, O., Schwartz, M. H., Thomason, P., and Baker, R. (2014). The comparison of normative reference data from different gait analysis services. *Gait Posture* 40, 286–290. doi: 10.1016/j.gaitpost.2014.03.185

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). *Why Should I Trust You?': Explaining the Predictions of Any Classifier.* New Jersey: IEEE.

Ries, A. J., Novacheck, T. F., and Schwartz, M. H. (2014). A data driven model for optimal orthosis selection in children with cerebral palsy. *Gait Posture* 40, 539–544. doi: 10.1016/j.gaitpost.2014.06.011

Rozumalski, A., and Schwartz, M. H. (2009). Crouch gait patterns defined using k-means cluster analysis are related to underlying clinical pathology. *Gait Posture* 30, 155–160. doi: 10.1016/j.gaitpost.2009.05.010

Sangeux, M., Passmore, E., Graham, H. K. K., and Tirosh, O. (2016). The gait standard deviation, a single measure of kinematic variability. *Gait Posture* 46, 194–200. doi: 10.1016/j.gaitpost.2016.03.015

Sangeux, M., and Polak, J. (2015). A simple method to choose the most representative stride and detect outliers. *Gait Posture* 41, 726–730. doi: 10.1016/j.gaitpost.2014.12.004

Sangeux, M., Rodda, J., and Graham, H. K. (2015). Sagittal gait patterns in cerebral palsy: The plantarflexor-knee extension couple index. *Gait Posture* 41, 586–91. doi: 10.1016/j.gaitpost.2014.12.019

Schwartz, M. H., Rozumalski, A., Truong, W., and Novacheck, T. F. (2013). Predicting the outcome of intramuscular psoas lengthening in children with

cerebral palsy using preoperative gait data and the random forest algorithm. *Gait Posture* 37, 473–479. doi: 10.1016/j.gaitpost.2012.08.016

Simon, S. R. (2004). Quantification of human motion: gait analysis-benefits and limitations to its application to clinical problems. *J. Biomech.* 37, 1869–1880. doi: 10.1016/j.jbiomech.2004.02.047

Thomason, P., Rodda, J., Willoughby, K. L., and Graham, H. K. (2014). "Orthopaedic Assessment," in *Cerebral Palsy*, eds B. Dan, M. Mayston, N. Paneth, and L. Rosenbloom (New Jersey: Mac Keith Press), 287–312.

Trunk, G. V. (1979). A Problem of Dimensionality: A Simple Example. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 306–307. doi: 10.1109/TPAMI.1979.4766926

Wagner, M., Slijepcevic, D., Horsak, B., Rind, A., Zeppelzauer, M., and Aigner, W. (2019). KAVAGait: Knowledge-assisted visual analytics for clinical gait analysis. *IEEE Trans. Vis. Comput. Graph.* 25, 1528–1542. doi: 10.1109/TVCG.2017.2785271

Weintraub, M. A., Bylander, T., and Simon, S. R. (1990). Quawds: a Composite Diagnostic System for Gait Analysis. *Comput. Methods Programs Biomed.* 32, 91–106. doi: 10.1016/0169-2607(90)90089-R

Woltring, H. J. (1986). A Fortran package for generalized, cross-validatory spline smoothing and differentiation. *Adv. Eng. Softw.* 8, 104–113. doi: 10.1016/0141-1195(86)90098-7

Wren, T. A., Gorton, G. E. III, Ounpuu, S., and Tucker, C. A. (2011a). Efficacy of clinical gait analysis: A systematic review. *Gait Posture* 34, 149–153. doi: 10.1016/j.gaitpost.2011.03.027

Wren, T. A., Otsuka, N. Y., Bowen, R. E., Scaduto, A. A., Chan, L. S., Sheng, M., et al. (2011b). Influence of gait analysis on decision-making for lower extremity orthopaedic surgery: Baseline data from a randomized controlled trial. *Gait Posture* 34, 364–369. doi: 10.1016/j.gaitpost.2011.06.002

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership