frontiers | Frontiers in Aging Neuroscience

TYPE Review
PUBLISHED 19 July 2023
DOI 10.3389/fnagi.2023.1216163

OPEN ACCESS

EDITED BY
Robert Petersen,
Central Michigan University, United States

REVIEWED BY
Diego Castillo-Barnes,
University of Malaga, Spain
Carmen Jiménez-Mesa,
University of Granada, Spain

*CORRESPONDENCE
Beatriz Garcia Santa Cruz
garciasantacruz.beatriz@chl.lu

RECEIVED 03 May 2023
ACCEPTED 28 June 2023
PUBLISHED 19 July 2023

CITATION
Garcia Santa Cruz B, Husch A and Hertel F (2023) Machine learning models for diagnosis and prognosis of Parkinson's disease using brain imaging: general overview, main challenges, and future directions. *Front. Aging Neurosci.* 15:1216163. doi: 10.3389/fnagi.2023.1216163

Machine learning models for diagnosis and prognosis of Parkinson's disease using brain imaging: general overview, main challenges, and future directions

Beatriz Garcia Santa Cruz[1]*, Andreas Husch[2] and Frank Hertel[1]

[1]National Department of Neurosurgery, Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg, [2]Imaging AI Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Parkinson's disease (PD) is a progressive and complex neurodegenerative disorder associated with age that affects motor and cognitive functions. As there is currently no cure, early diagnosis and accurate prognosis are essential to increase the effectiveness of treatment and control its symptoms. Medical imaging, specifically magnetic resonance imaging (MRI), has emerged as a valuable tool for developing support systems to assist in diagnosis and prognosis. The current literature aims to improve understanding of the disease's structural and functional manifestations in the brain. By applying artificial intelligence to neuroimaging, such as deep learning (DL) and other machine learning (ML) techniques, previously unknown relationships and patterns can be revealed in this high-dimensional data. However, several issues must be addressed before these solutions can be safely integrated into clinical practice. This review provides a comprehensive overview of recent ML techniques analyzed for the automatic diagnosis and prognosis of PD in brain MRI. The main challenges in applying ML to medical diagnosis and its implications for PD are also addressed, including current limitations for safe translation into hospitals. These challenges are analyzed at three levels: disease-specific, task-specific, and technology-specific. Finally, potential future directions for each challenge and future perspectives are discussed.

KEYWORDS

Parkinson's disease, translational ML, neuroimaging, machine learning, deep learning, computer-aided diagnosis, digital health

Introduction

Computer-aided diagnosis (CAD) systems based on medical imaging has the potential to assist clinical practice in the diagnosis of Parkinson's disease (PD). However, the suitability of CAD systems for this application is still being evaluated, and several key aspects must be taken into consideration.

The primary objective of CAD systems is not to replace radiologists and clinicians, but to support them in improving the quality and efficiency of their diagnoses (Chen et al., 2013). Although CAD systems have been in use for several decades, with successful applications in detecting pulmonary nodules (Xu et al., 1997) and breast cancer (Mangasarian et al., 1995), they were previously reliant on manual feature extraction based on domain knowledge. However, with the recent emergence of Machine Learning (ML) techniques, such as Deep Learning (DL), the automatic extraction of features from imaging data has become possible

(Doi, 2007). Furthermore, the availability of large datasets and more powerful computational infrastructure has facilitated the development of advanced ML algorithms, which have the potential to significantly improve the accuracy of CAD systems (Neri et al., 2019).

Although CAD systems based on Artificial Intelligence (AI) have the potential to greatly enhance the effectiveness of clinical diagnosis and prognosis workflows, it is essential to carefully consider several key factors to ensure their safe and effective implementation in clinical practice. In fact, there is often a gap between the research literature on ML models and their final deployment in clinical applications. Closing this gap requires careful consideration and addressing several crucial aspects such as model robustness, data quality and bias, regulatory compliance, integration with existing clinical workflows, and ongoing validation in real-world settings.

A good example of a clinical deployment of an AI system that exemplified this gap is the AI-based tool by Google, Automated Retinal Disease Assessment (ARDA) system. Although this DL system was successfully developed and internally validated at the research level in 2016 (Gulshan et al., 2016), it faced several challenges in transitioning from the theoretical expectations to the reality of deploying the AI model tool in India and Thailand, as discussed in a recent paper highlighting the necessity of considering this gap (Widner et al., 2023).

While previous review papers have thoroughly covered the topic of using ML as a proof-of-concept for CAD systems (Sakai and Yamada, 2019; Mei et al., 2021), there has not been a previous review that specifically addresses the changes and potential solutions associated with the translation of these models into clinical practice for PD imaging using ML.

This review is organized as follows: first, a comprehensive background on PD, including related conditions and proposed clinical subtypes is presented. Second, the diagnosis and prognosis of PD is introduced, with a specific focus on the employment of magnetic resonance imaging (MRI). Lastly, a comprehensive analysis of the present status of computer-aided diagnosis, will be discussed, emphasizing the main limitations and future directions at three different levels. These considerations will take into account the unique features of PD, as well as the limitations of clinical brain imaging datasets, and the challenges associated with ML and DL approaches. By considering these factors, this review aims to provide insights into the potential of CAD in assisting clinical practice in the diagnosis of PD, while also highlighting the challenges that need to be addressed to ensure its safe and effective translation into clinical practice.

Parkinson's disease and related disorders

It has been more than 200 years since the first description of the symptoms of PD by James Parkinson in his essay "The Shaking Palsy" (Parkinson, 2002). This first description refers to some of the most prominent physical landmarks of the disease, such as tremors and flexed posture. Nowadays, we have a more holistic understanding of this complex neurodegenerative disease, but currently, there is no cure, and no established biomarker for differential diagnosis of the disease (Tolosa et al., 2021).

PD is the second most common neurodegenerative disorder after Alzheimer's disease (AD), with more than 10 million people affected worldwide (Marras et al., 2018). One of the main risk factor associated with PD is advanced age. Considering that the elderly population is expected to double by 2050, the number of PD patients is expected to increase accordingly (Nerius et al., 2017). It is characterized by visible motor symptoms such as slowness of movement, muscle rigidity, and tremors at rest (Sveinbjornsdottir, 2016). However, non-motor symptoms such as depression, anxiety, cognitive deficits, sleep disturbance, hyposmia, cardiovascular problems, and bladder dysfunction can also be debilitating and may present before the motor problems (Chaudhuri et al., 2006). Notably, there is growing evidence that PD is associated with gastrointestinal dysfunction and changes to the microbiome, which may have potential as a biomarker (Elfil et al., 2020). By the time the main physical symptoms of PD appear and the patient receives a diagnosis, 30%–50% of the dopamine neurons vulnerable to PD are already lost. Hence, a key goal is to detect and quantify PD biology before their symptoms appear, during the prodromal phase (Pellicano et al., 2007). Clinical markers of this phase are non-motor and motor symptoms. Non-motor symptoms include hyposmia, constipation, REM sleep behavior disorder (RBD), excessive daytime somnolence, depression and/or anxiety, global cognitive deficit, and orthostatic hypotension. Motor symptoms include voice and face akinesia (Hustad and Aasly, 2020).

PD affects various regions of the nervous system and different types of neurons. However, much attention has been given to neurons in brain regions associated with motor symptoms, particularly the *substantia nigra pars compacta* in the midbrain. This region is involved in a critical brain pathway that facilitates movements, known as the nigrostriatal pathway (Eriksen et al., 2009). One of the most widely accepted frameworks to describe the spread of sporadic PD is Braak's hypothesis, which suggests that PD progresses through six different stages, gradually evolving from the lower brain stem to the neocortex (Rietdijk et al., 2017). The gradual degeneration of dopaminergic neurons in the substantia nigra leads to the malfunction of this pathway and the characteristic motor problems. It has been proposed that not all patients follow this progression, and two subtypes have been suggested for the disease evolution: peripheral nervous system first (PNS-first) and central nervous system first (CNS-first) (Borghammer and Van Den Berge, 2019). The existence of these subtypes is supported by *in vivo* imaging studies of RBD-positive and RBD-negative patient groups (Borghammer and Van Den Berge, 2019), as well as for genetic makers (Blauwendraat et al., 2020).

Current treatments for deficits in dopamine often involve the use of drugs that either replace or mimic dopamine in the brain (Cools, 2006). However, over time, the effectiveness of these drugs tends to diminish. In addition to medication, physical therapy can be employed as a complementary approach to enhance cognitive function in individuals with dopamine deficits (da Silva et al., 2018). Physical therapy focuses on improving mobility, balance, and coordination, which can positively impact cognitive abilities. Furthermore, alternative therapeutic avenues are being explored. Probiotics have shown potential in reducing

constipation associated with Parkinson's disease (Tan et al., 2021). Additionally, anaerobic exercise has been investigated as a current approach for managing dopamine deficits (Schootemeijer et al., 2020). Moreover, emerging treatment options include drug repurposing, regenerative therapies, gene therapies, and cell-based treatments (Stoker and Barker, 2020). These innovative approaches offer promising prospects in the management of dopamine-related deficits.

Deep brain stimulation (DBS) is an effective treatment option for PD by targeting the subthalamic nucleus, globus pallidus (Lee et al., 2019), ventral intermedius nucleus (Fasano et al., 2012), and pedunculopontine nucleus (Thevathasan et al., 2018). Next-generation noninvasive DBS technologies, such as noninvasive or minimally invasive DBS (Lozano, 2017), transcranial direct current stimulation (tDCS) (Broeder et al., 2015), and transcranial magnetic stimulation (TMS) (Cantello et al., 2002), have also shown positive effects in reducing non-motor symptoms of PD when appropriate controls for side effects are in place. However, there is currently no cure for neurodegeneration, and current efforts focus on reducing symptoms to improve the quality of life.

Related conditions

Several neurological movement disorders are closely associated with PD, and differentiating it from other diseases can be challenging, especially during the initial stages of the disease (Poewe and Wenning, 2002). Related disorders that share similar clinical features with PD can be classified into two broad categories: degenerative disorders and non-degenerative disorders (Politis, 2014). Degenerative disorders, such as Multiple System Atrophy (MSA), Progressive Supranuclear Palsy (PSP), Corticobasal Degeneration (CBD), Dementia with Lewy Bodies (DLB), and AD, can present with clinical features that overlap with PD. On the other hand, non-degenerative disorders such as Essential Tremor (ET), dystonic tremor, exaggerated physiological tremor, tremor related to hyperthyroidism, vascular parkinsonism, normal pressure hydrocephalus (NPH) (Stolze et al., 2001) and drug-induced parkinsonism can also mimic some of the clinical features of PD.

Parkinson's disease clinical subtypes

The clinical and neuropathological heterogeneity of PD patients is well known, and consequently there have been many attempts to identify different subtypes. Initial approaches consisted of empirical classifications using *a priori* hypotheses (Zetusky et al., 1985; Jankovic et al., 1990). In recent years, research works have progressively employed data-driven cluster analysis that includes longitudinal assessment of motor and non-motor symptoms (De Pablo-Fernández et al., 2019; Zhang et al., 2019; Dadu et al., 2022). This classification method looks promising for informing patients about the future progression of the disease and for personalizing treatment. However, these criteria are not yet applied in clinics since more research is needed to unify and validate the criteria using well-curated longitudinal cohorts. Among the multiple attempts to separate the disease, several criteria have been applied, including early-onset vs. late-onset (Riboldi et al., 2022) slow vs. fast progression, with or without dementia or tremor-dominant vs. gait-dominant (Dadu et al., 2022).

Machine learning, deep learning and computer vision

In recent years, ML and DL have gained significant attention in healthcare and medical research. These computational tools enable the analysis of large and complex datasets to learn patterns and relationships, with DL algorithms utilizing multiple layers of artificial neural networks to extract abstract data representations such as images. Furthermore, Computer Vision (CV) seeks to enable computers to interpret and understand visual information from the surrounding environment. Supervised learning is a common type of ML employed in PD research, where labeled datasets are used to train the algorithm to make predictions on unseen data. Convolutional neural networks (CNNs) are the most frequently used type of neural network for image recognition to conduct tasks such as classification in medical imaging. In Figure 1, a graphical representation of the training and development of a ml-based system for clinical use is depicted.

The quality of data and labels are crucial factors that can significantly impact the performance of ML models. In current ML models, data is the most important component as the models learn from the data presented to them. Therefore, the quality of the data used in the training process is crucial. Other factors that can influence the quality of models include the choice of ML algorithms, feature engineering, hyperparameter tuning, and model selection. In addition to data quality, the quality of labels is also critical. Poor quality labels can result in biased models, incorrect predictions, and suboptimal performance. Moreover, data representation is equally important for a good model performance. A training set should be a representation of the event that we want to model, and a good validation strategy is essential for assessing the generability of the model.

Parkinson's disease diagnosis and prognosis

Accurate diagnosis of PD is essential, and achieving enough specificity to distinguish between similar conditions during the clinical phase is crucial. Developing monitoring tools to track disease progression and evaluate individual patient response, including the presence and magnitude of treatment side effects, is also necessary. Furthermore, quantifying the different systems, such as motor, memory, and limbic system, could help stratify patients. In terms of prognosis, ongoing efforts are focused on establishing clear criteria for patient stratification into different subtypes, which would aid in the development of targeted

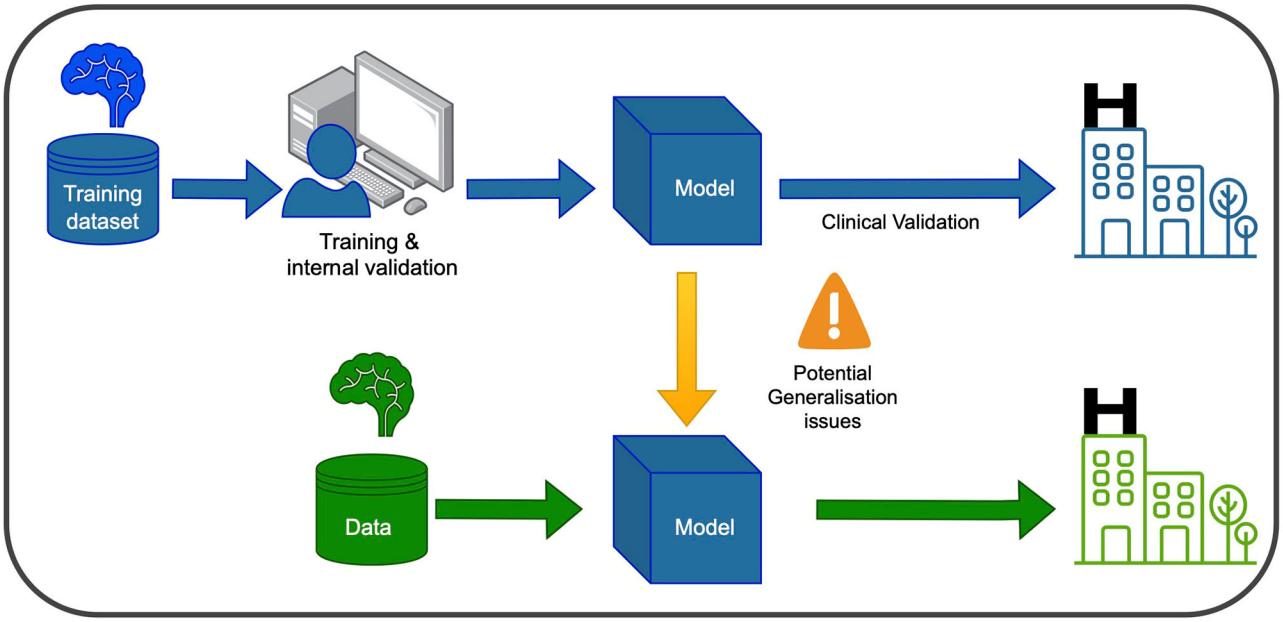

FIGURE 1
Training and using an ML model in the clinic involves two main phases. In the blue phase, the model is trained and validated using data from the same hospital. This ensures it learns from the hospital's specific context and performs well within that setting. After this, the model undergoes clinical validation to ensure its reliability and safety before deployment. In the green phase, the model can be used in new hospitals, but caution is needed to address potential generalization issues. Variations in healthcare systems and patient populations may affect its performance. Thorough testing and evaluation are necessary to ensure accurate and safe application in different healthcare settings.

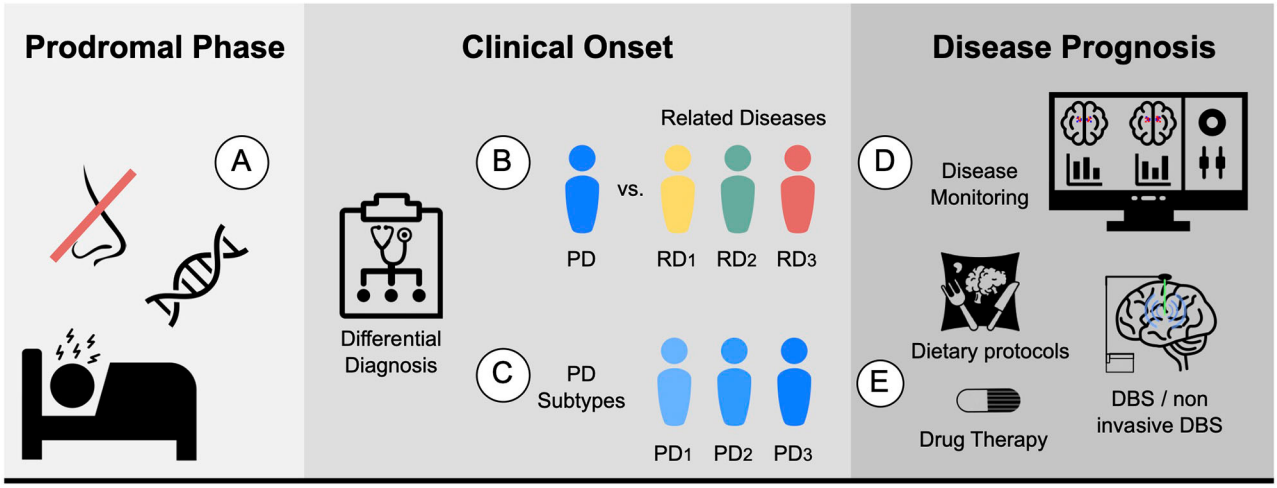

FIGURE 2
Proposed biomarkers for PD using MRI: **(A)** Prodromal biomarker: identifying brain changes during the prodromal phase. **(B)** Differential diagnosis biomarker: assisting in distinguishing PD from related diseases. For instance, ET or MSA. **(C)** Subtype biomarker: classifying PD patients into their corresponding subtypes. **(D)** Progression biomarker: aiding in predicting the progression of the disease and treatment response with disease monitoring. **(E)** Therapy response biomarker: facilitating personalized medicine by finding the best drug, dietary protocols, physical or cognitive therapies, and predicting the potential response to other therapies such as DBS and non-invasive DBS.

treatment approaches. Figure 2 proposes five different biomarkers that are relevant in the context of PD.

The current diagnostic criteria for PD is biased on a comprehensive evaluation of a patient's clinical presentation and medical history. Given the lack of a definitive diagnostic test for PD, clinicians rely on a variety of subjective and objective measures to make an accurate diagnosis. Clinical evaluation, involving detailed inquiry into the patient's symptoms, medical history, and family history, represents a fundamental component of the diagnostic process. Alongside this, a thorough physical examination aimed at assessing motor function, including muscle strength, reflexes, and coordination, as well as cognitive function and mood, is also

typically conducted. To support a clinical diagnosis, objective tests may be employed. Imaging modalities such as MRI or computed tomography (CT) scans are typically employed to rule out other conditions that may present similarly to PD. Furthermore, nuclear imaging techniques such as Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) can serve to buttress the diagnosis of PD.

Nowadays, there is a significant effort to find biomarkers for PD. In the preclinical phase, it highlights biomedical markers, such as those that measure the activity of mitochondria dysfunction and oxidative stress (He et al., 2018). Others focus on measuring abnormal protein aggregation and accumulation, such as alpha-synuclein (Foulds et al., 2011) or tau protein (Constantinescu and Mondello, 2013). Some try to measure established clinical features such as olfactory dysfunction, RBD, or constipation. During the prodromal phase, genetic biomarkers have been explored, such as mutations in Parkin (Pickrell and Youle, 2015), Leucine-rich repeat kinase 2 (LRRK2) (Tolosa et al., 2020), or Alpha-synuclein (SNCA) (Mata et al., 2010). Finally, neuroimaging techniques are also promising.

In the context of brain imaging, a biomarker is an objective characteristic derived from an *in vivo* image that measures a normal biological process, pathological process, or response to a therapeutic intervention (Mohammadi, 2013). It must fulfill the following criteria: be quantitative, repeatable, reproducible, precise, reliable, sensitive, and specific, and be measured on a ratio or interval scale (Smith et al., 2003).

Medical imaging in Parkinson's disease

The main advantage of brain imaging is that it allows for the visualization of the functional and structural brain changes that result from underlying pathophysiological abnormalities (Saeed et al., 2017). There are several imaging techniques that can be used to aid in the diagnosis and prognosis of PD.

On the one hand, there is a set of non-invasive techniques for investigating PD, such as **structural magnetic resonance imaging** (MRI) with T1, T2, and susceptibility-weighted sequences, which allow for volumetric and voxel-based morphometric analyses, as well as MRI-derived visual signatures (Saeed et al., 2017; Chougar et al., 2021). For instance, Schwarz et al. (2014) proposed that the appearance of the dorsolateral substantia nigra as a "swallow tail" shape on high-resolution, iron-sensitive, MRI at 3T, where healthy nigrosome-1 appears as a characteristic feature that could be employed as a marker of degeneration in that area. Further, a promising structural MRI sequence for PD diagnosis is neuromelanin-sensitive MRI (NM MRI), which can detect neuromelanin, a pigment synthesized by the substantia nigra dopamine neurons that is lost when neurons die in PD patients. NM's avid binding of iron enables its detection via magnetic resonance imaging (Sulzer et al., 2018). The use of NM MRI to define regions of interest (ROIs) in the substantia nigra pars compacta (SNpc) has shown promising results compared to using T2*-weighted contrasts. This approach has yielded consistent results, and studies have found that the mean R2* in the SNpc, as defined by neuromelanin-sensitive MRI, was significantly increased in PD patients (Langley et al., 2019).

Diffusion tensor MRI (DT-MRI) is another technique used to study the structural connectivity of the brain in PD. DT-MRI investigates the integrity of white matter tracts connecting different brain regions, and studies have shown that it can detect changes in white matter connectivity in PD patients. Specifically, Yoshikawa et al. (2004) demonstrated that DT-MRI can detect the loss of fractional anisotropy (FA) in the nigrostriatal projection, indicating that more than half of the dopaminergic neurons in this projection may be lost before the onset of PD.

Furthermore, **functional magnetic resonance imaging** (fMRI) can detect changes in blood flow in response to neural activity, which enables researchers to study brain function. In PD, fMRI has been used to investigate changes in brain activity related to both motor and non-motor symptoms. For instance, Tahmasian et al. (2015) employed resting-state (rs-fMRI) to assess the effect of dopamine replacement therapies, such as levodopa and dopamine agonists, on PD patients. Additionally, researchers have used fMRI techniques to investigate the effect of DBS therapy in the modulation of specific brain regions. An example of this is a study by Boutet et al. (2021), in which fMRI brain response patterns were used to predict the optimal parameters for DBS by identifying patterns associated with clinically effective stimulation that preferentially engages the motor circuit.

Additionally, **Transcranial sonography** (TCS) is an ultrasound-based neuroimaging technique that utilizes low frequency sound waves to generate images of the brain. In the context of PD diagnosis, TCS has been employed to investigate the structure and function of the SN, among other brain regions. Mahlknecht et al. (2013) demonstrated that TCS exhibits favorable diagnostic accuracy in detecting PD subjects based on the presence of hyperechogenicity in the SN Furthermore, TCS has been investigated as a potential tool to establish disease progression biomarkers that could provide real-time feedback on the rate of dopaminergic neuronal death in animal models (Zhang et al., 2020).

On the other hand, invasive **molecular imaging** techniques such as PET and SPECT can detect reduced density of dopaminergic nerve terminals in the basal ganglia. PET is an *in vivo* functional neuroimaging technique that utilizes a variety of radionuclides to assess the integrity of the dopaminergic system, cerebral metabolism, pathological protein accumulation, and inflammation in the brain (Saeed et al., 2017). Radiotracers, such as 18F-dopa (Morrish et al., 1996) and 11C-raclopride (Politis et al., 2008), can image the integrity of presynaptic and postsynaptic nigrostriatal and hypothalamus projections, respectively. Using SPECT, dopamine transporter SPECT (DAT SPECT) imaging is an objective tool for assessing dopaminergic function of presynaptic terminals, differentiating parkinsonian disorders related to striatal dopaminergic deficiency from those not related. DAT SPECT imaging can confirm or exclude a diagnosis of dopamine-deficient parkinsonism and detect dopaminergic dysfunction in presymptomatic subjects at risk for PD. Normal DAT SPECT findings exclude presynaptic striatal dopaminergic insufficiency, while abnormal findings indicate a variety of diseases with this insufficiency as a common pathophysiological process (Akdemir et al., 2021). For instance, DaT SPECT imaging with (123I)ioflupane is a useful tool to distinguish between PD-tremor and non-PD tremor, such as ET (Bajaj et al., 2013). Besides,

other non-dopaminergic imaging techniques such as glucose metabolism and PDE10A expression have been proposed to study PD (Pagano et al., 2016). Additionally, extrastriatal ^{123}I-FP-CIT SPECT impairment has been proposed to detect early cases of PD (Nicastro et al., 2020).

While imaging techniques are currently used for research purposes and can assist in challenging cases, they are not commonly used for diagnosing PD. However, it is worth noting that most PD diagnoses do not involve imaging. In the future, brain imaging could be integrated into the diagnostic process as advancements in techniques like ML and CV hold promise for improving the analysis of imaging data. These developments may enable more accurate and reliable diagnostic applications of imaging in PD.

Computer-aided diagnosis using brain imaging: main limitations and future directions

The main limitations of CAD systems in the context of PD can be grouped into three categories. The first set of limitations represented in Figure 3 pertains to the particularities of PD, its diagnosis, and prognosis. The second set of limitations is associated with the characteristics of datasets consisting of brain imaging. These limitations include factors such as the heterogeneity of the imaging modalities used, variability in image acquisition protocols, challenges in image preprocessing and feature extraction, and issues related to sample size and data quality. The third set of limitations is associated with the use of ML/DL-based algorithms for CAD systems. These limitations include challenges such as overfitting, lack of interpretability, bias and generalization issues, and difficulties in integrating multiple data sources. A summary of the main limitations can be found in Table 1, which will serve as a reference point throughout the discussion of potential solutions to address these limitations.

CAD systems have the potential to improve the accuracy and efficiency of diagnosing various diseases. By analyzing medical imaging data, genetic data, and clinical data, these systems can identify patterns and biomarkers associated with the disease that may be difficult to detect otherwise, which can accelerate the diagnostic and treatment workflows in clinical pathways. Moreover, CAD systems can be employed to evaluate disease progression, measure therapeutic responses to drugs in clinical trials, and speed up the development of new treatments.

Other benefits of CAD systems include the objectification of diagnosis, as the current diagnosis relies on subjective evaluation of motor and non-motor symptoms, making CAD systems promising tools for the objective evaluation of symptoms. In the context of MRI for PD, CAD systems can provide quantitative measures of the changes associated with the disease at physical, functional, and metabolic levels. Furthermore, the employment of CAD systems could aid in the unification of clinical diagnosis criteria. Additionally, CV solutions, including those that employ DL as an optimisation technique, have been shown to excel at detecting subtle changes and complex patterns in comparison with human vision. Therefore, CAD systems have the potential to serve as a valuable second or supporting opinion, as they do not experience a reduction in productivity over time, as can happen with human experts.

There are many research-level papers proposing proof-of-concept approaches for CAD systems in PD, emphasizing the importance of robust models. For instance, Castillo-Barnes et al. (2018) utilized the PPMI dataset and proposed an Ensemble Classification model to classify PD patients. Similarly, Augimeri et al. (2016) demonstrated the potential of support vector machines in combination with careful feature extraction to analyze DaTSCAN scans for PD applications. In line with these studies, Martínez-Murcia et al. (2014) also proposed a PD classification method using DaTSCAN scans.

Similarly, machine learning (ML) has been employed to distinguish between PD and related disorders. For instance, Talai et al. (2021) propose a multimedia approach using T1-weighted, T2-weighted, and diffusion tensor imaging (DTI) to aid in the differential diagnosis of progressive supranuclear palsy Richardson's syndrome (PSP-RS). In the same vein, Martins et al. (2021) reported on the use of PET uptake and MRI for distinguishing Parkinsonian syndromes. Similarly, Castillo-Barnes et al. (2020) conducted a study that employed SPECT scans from the PPMI database and compared different ML methods.

More recently, CNN has been successfully proposed for the classification of brain imaging in PD. For instance, Chakraborty et al. (2020) proposed a classification using T1 weighted MRI scans using CNNs. Similarly, Martinez-Murcia et al. (2019) demonstrated the use of autoencoders to classify complex neurological diseases such as Alzheimer's. Finally, Shinde et al. (2019) also demonstrated the potential of CNNs in the modality of neuromelanin-sensitive MRI with great performance (Biondetti et al., 2020).

The mentioned research-level papers and alike ones, provide a valuable insights into the potential of CAD systems for PD. However, it is crucial to acknowledge that these studies primarily focus on demonstrating the effectiveness of specific methodologies or models in isolated aspects of PD diagnosis or classification. While their findings are promising and essential to the progress in the area, they represent only a fraction of what is required for the development of comprehensive and practical clinical systems.

To build end-to-end clinically useful CAD systems for PD, various aspects need to be considered beyond the individual proof-of-concept models. These aspects may include data acquisition and quality assurance, integration with existing clinical workflows, interpretability of the models, regulatory compliance, ethical considerations, scalability, and validation in diverse patient populations. The following sections of the paper will delve into these critical considerations and discuss potential solutions to ensure the successful implementation and utilization of CAD systems in real-world clinical settings.

Limitations associated with Parkinson's disease

Disease heterogeneity: intra-class variance and inter-class similarity

Medical conditions may have several etiologies. Moreover, one etiology may lead to more than one disease (Coleman and Tsongalis, 2009). Consequently, medical conditions are commonly

10.3389/fnagi.2023.1216163

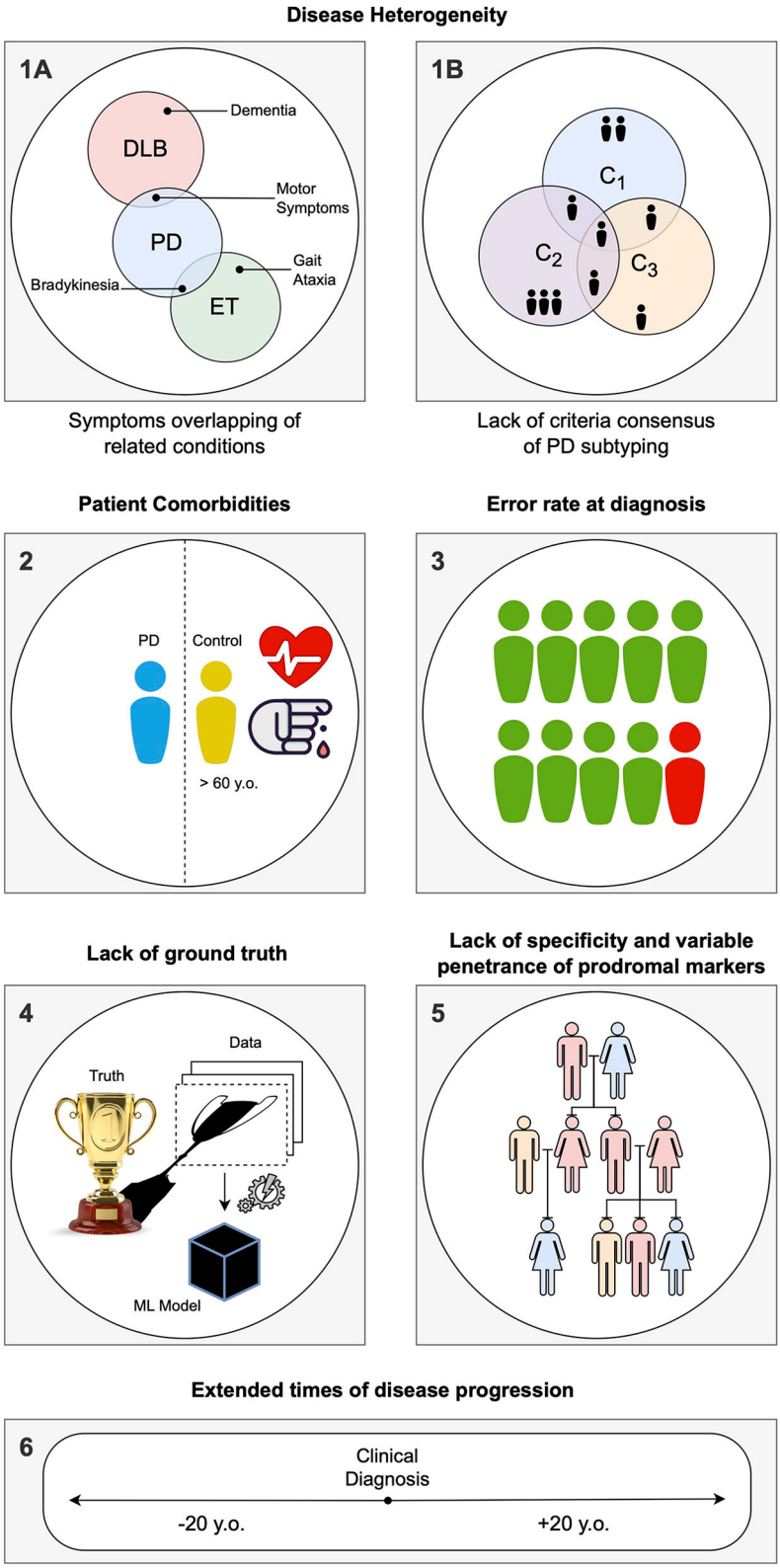

FIGURE 3
Summary of the specific limitations in computer-aided diagnosis (CAD) for Parkinson's disease (PD) associated with idiosyncrasies of the disease, as addressed in Section Limitations associated with Parkinson's disease: **(1)** During the labeling of datasets for supervised learning, several problems can be encountered. **(1A)** Building a solution for differential diagnosis can be challenging due to the overlapping symptoms of PD and related disorders. This challenge is especially significant during the initial phases of clinical diagnosis, where such solutions would be most useful. **(1B)** PD is known to have several subtypes with implications for clinical treatment, but there is a lack of clear global consensus, adding another layer of complexity. **(2)** PD being an age-related disorder, the control subjects used in age-pairing may have additional health conditions or factors that can affect their

(Continued)

FIGURE 3 (Continued)
representatives as healthy individuals. **(3)** Due to the complexity of PD, there is a notable rate of misdiagnosis, even in specialized centers, particularly during the early phases of clinical diagnosis. This hampers the accuracy of labels used in supervised learning solutions. **(4)** When acquiring data and building a model, a simplification of the disease within the context of human biology is necessary, as it is the case with any other data-driven solution. Consequently, any developed solution will have errors, particularly if the model is used in different conditions than those it was designed for. **(5)** Detecting PD in the prodromal phase is particularly challenging. A common approach is to employ known markers that increase the probability of developing the disease, such as genetic mutations. However, the specificity of these markers to PD is variable. **(6)** Conducting long-term longitudinal studies that are consistent in terms of acquisition protocol while maintaining low levels of drop-out rates is extremely difficult for PD, given its nature as a complex, long-term neurodegenerative disease.

TABLE 1 Overview of limitations and future directions at the three levels: disease-specific, task-specific, and technology-specific.

Limitations	Directions
Parkinson's disease	
Disease heterogeneity	Considering subgroups of PD and careful assessment of controls
Patients' comorbidities	Large and Long studies and control of unwanted correlations
Error rate at diagnosis	Acknowledging errors and employing noise-labeled techniques
Extended times of disease progression	Institutional incentives, importance of consistency in protocols
High variability of prodromal markers	Multimodal prodromal markers, epigenetics changes
Lack of ground truth	Objective measures, holistic multidisciplinary approach
Clinical brain imaging datasets	
Complexity of brain imaging	Multimodal approach, combination with clinical measures
Lack of standardization in acquisition	Standardization of acquisition, sharing study assumptions
Lack of standardization in preprocessing	Sharing raw data and reproducible code ability
Lack of standardization in annotation	Assisted annotation with guidelines and unsupervised learning
Machine learning/deep learning	
Generalization issues	Avoid overfitting, control for spurious correlations
Algorithmic Bias	Acknowledge algorithm bias and prioritize fairness strategies
Need for better interpretability	Prioritize transparency and ethics, GDPR compliance
Model explainability	Use explainable ML algorithms, employ interpretability methods
Model uncertainty	Documentation of uncertainty sources, calibration methods
Costly systems to develop and maintain	Pre-train models, cloud computing, decentralized ML
Security and privacy challenges	Proactive security and privacy strategies

defined clinically or pathologically (instead of etiologically). PD presents high variability at both prodromal and clinical phases (He et al., 2018). We can refer to this variability as an intra-class variance. However, another level of complexity exists due to the overlap of PD symptoms with those from other diseases, which calls for thorough differential diagnosis (Kalia and Lang, 2015). For instance, patients with arterial hypertension may exhibit distinct neuroimaging abnormalities detectable by brain MRI (van Veluw et al., 2014), which may complicate the diagnosis of PD using medical imaging techniques in these individuals. Thus, we can find a high inter-class similarity. Finally, diseases are described based on a definable deviation from a normal phenotype made evident through symptoms, and pathological markers, to then become grouped into categories. However, studies and taxonomies struggle to find a consensus for PD subtypes (Albrecht et al., 2022). Hence, studies may employ different subtypes to refer to the same biological mechanism and therapy response.

Patients' comorbidities

In addition to the aforementioned complexity, the onset age of PD in patients is typically around 60 years, making it difficult to differentiate symptoms caused by aging and other comorbidities from those of PD (Deeb et al., 2019). For instance, common comorbidities in PD patients, such as hypertension and diabetes, have an unknown effect on the pathogenesis and progression of PD (Santiago et al., 2017). This presents a twofold challenge: first, it complicates the identification of a reliable set of control and diseased subjects, making it difficult to distinguish between groups. Second, due to the lack of knowledge regarding the effects of comorbidities on PD onset and development, controlling for these characteristics is challenging. As a result, researchers may face a "lose–lose" situation, as ML models may make assumptions that cannot be refuted or confirmed by the researcher. This situation is also referred to as butterfly bias, in which a variable or feature may

be considered both a confounder and a source of M-bias (Ding and Miratrix, 2015).

To mitigate the effects of comorbidities and the heterogeneity of PD, researchers often employ large sample sizes to account for the variability in the population and the disease. For example, datasets like the Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2011) and the Oxford Parkinson's Disease Centre discovery cohort (OPDC) (Lawton et al., 2015) acknowledge the presence of subtypes and follow patients over extended periods, presenting clinical data in addition to imaging data. Moreover, studies frequently use statistical techniques such as propensity score matching (Huang et al., 2013), stratification (Virreira Winter et al., 2021), and multivariable regression (Pechevis et al., 2005) to control for confounding variables. Another approach is to utilize ML algorithms that can handle multiple confounders and nonlinear relationships between variables, such as random forest (Oprescu et al., 2019) or support vector machine models (Westreich et al., 2010).

Error rate at diagnosis

The aforementioned challenges are further compounded by the difficulty of accurately diagnosing PD. According to Hess and Okun (2016), the misdiagnosis rate of PD can range from 10 to 20% or greater, depending on clinician experience. Other studies have reported misdiagnosis rates of 20%–30% in the early stages, with the main causes being the failure to recognize atypical parkinsonian disorders such as dementia with Lewy bodies or multiple system atrophy (Poewe and Wenning, 2002). Consequently, researchers must address the challenges of training models with noisy labeled data (Karthik et al., 2021), where label noise can potentially degrade model performance.

To address noisy labeled data several approaches have been proposed, including semi-supervised learning, where a small set of labeled data is combined with a large set of unlabeled data to improve the model's accuracy (Adeli et al., 2018). Another approach is active learning, where the model is iterative trained on a small set of labeled data, and the most informative samples are selected for annotation by a human expert, reducing labeling costs while maintaining or even improving the model's accuracy (Settles, 2009; Garcia Santa Cruz et al., 2022a). Recent developments in DL have led to the emergence of new techniques that can handle label noise more robustly, such as the label Smoothing technique (Müller et al., 2019) that reduces the impact of noisy labels on the loss function by smoothing the label distribution. Ensemble techniques also help mitigate the impact of label noise on model performance by combining the predictions of multiple models, each trained on a slightly different subset of the data (Adeli et al., 2018).

Extended times of disease progression

PD is characterized by a slow progression, with a period of up to 20 years before the clinical phase (Kalia and Lang, 2015), and can survive up to 20 years in the clinical phase (Hassan et al., 2015), with a mean survival onset of 12 years (Rajput, 1992). This slow progression impacts longitudinal follow-up of study participants, which becomes difficult and prone to high dropout rates and protocol changes. It also brings another important dimension into play, as data subjects may showcase both different ages and distinct PD stages. Moreover, assumed control subjects may reveal PD symptoms in the long term, increasing the risk of ascertainment bias.

The extended duration of longitudinal studies can lead to higher rates of dropout and protocol changes. To mitigate these issues, researchers can employ remote monitoring technologies that allow patients to be monitored from their homes, reducing the need for in-person visits. Wearable sensors can also provide continuous, objective measurements of symptoms and mobility (Kubota et al., 2016; Arroyo-Gallego et al., 2018). Additionally, providing incentives to patients and institutions can help improve retention rates (Smith et al., 2019). For brain imaging studies, it is important to maintain consistent imaging protocols and analysis methods to reduce the risk of acquisition bias (Castro et al., 2020).

Lack of specificity and variable penetrance of prodromal markers

Finding markers for the prodromal phase of PD is complex in many aspects. One of the key factors hindering the discovery of such markers is the low frequency of the disease, which is estimated to be under 2% (Muangpaisan et al., 2011). This low frequency makes it challenging to find participants in the prodromal phase of the disease, as large sample sizes are required for such studies. To overcome this challenge, researchers often employ non-specific markers to identify individuals who may be in the prodromal phase of PD. These non-specific markers include rapid eye movement sleep behavior disorder (RBD), hyposmia (reduced ability to smell), depression, gastrointestinal symptoms, and mild motor symptoms. However, the use of non-specific markers has limitations, as they are not specific to PD and may be present in individuals who do not develop the disease (Durcan et al., 2019). Although specific markers such as genetic markers have been identified, their use is limited by their variable penetrance, which is often incomplete and dependent on the population. Some of the most commonly associated genes with PD are LRRK2, Glucocerebrosidase (GBA), and SNCA (Niotis et al., 2022). This means that even if an individual has a genetic marker associated with an increased risk of developing PD, there is still a significant chance that they may never develop the disease.

Finding markers for the prodromal phase of PD is complex, but one potential solution to overcome the challenge of low disease frequency and the need for large sample sizes is to collaborate with multiple research centers and establish consortium. Another approach to identifying specific markers for the prodromal phase of PD is to consider multiple sources of data, such as the hyposmia test (Siderowf et al., 2012). Finally, to address the limitations of genetic markers with incomplete penetrance, researchers can focus on identifying epigenetic modifications associated with the prodromal phase of PD, which may provide more accurate and specific markers for early detection of the disease (Chen and Ritz, 2018).

Lack of ground truth

In addition to the challenges of finding markers for the prodromal phase, there are also challenges related to generating accurate ground truth data for supervised learning. PD is not fully understood yet, which can lead to errors in the models. Deliberate idealisations are inherent in any model, but inaccurate assumptions based on insufficient knowledge can lead to biased and inaccurate representations. An example of this is the lack of understanding about comorbidity effects. Disparities in these regards can affect coherence between studies, as causal assumptions may vary across research teams and over time. Conducting further research on the disease could be a potential solution to enhance the understanding of the disease. This research can include a better understanding of the various aspects that contribute to the disease, such as adopting a complex systems approach (Cohen et al., 2022). Another solution is to develop more objective and quantitative measures of motor symptoms using wearable sensors and digital technologies.

Current diagnosis relies on assessments by physicians, often employing the current gold standard, the Unified Parkinson's Disease Rating Scale (UPDRS) (Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 2003). Furthermore efforts are underway to develop more objective and continuous measures of motor symptoms using wearable sensors and digital technologies (Parisi et al., 2015; Lu et al., 2021). These emerging technologies can provide more accurate and reliable data for the diagnosis and monitoring of PD (Kubota et al., 2016). By replacing subjective evaluations with objective measurements, the accuracy of diagnoses may be improved, leading to earlier identification and treatment of PD. Further research on the missing link between genetic and environmental causes of the disease can also contribute to a better understanding of PD (Hill-Burns et al., 2017). Additionally, standardizing diagnostic criteria and protocols across research teams and clinical settings can increase coherence between studies and improve the accuracy of the diagnosis. One such criterion is the UK Brain Bank criteria (Postuma et al., 2018). Enhanced collaboration and communication between researchers and clinicians may serve as a valuable means to reinforce the aforementioned efforts.

Limitations associated with clinical brain imaging datasets

Diversity and complexity of *in vivo* imaging brain markers

The pathology underlying PD motor symptoms such as tremors and bradykinesia is mainly associated with the loss of dopaminergic neurons in the *substantia nigra* and other gray matter alterations visible through brain imaging. However, non-motor symptoms of PD such as hyposmia, sleep disturbances, and depression do not present a clear *in vivo* imaging brain marker, even though some NMS-related brain alterations have been described. In particular, Prell (2018) state that imaging NMS characteristics may require different modalities, e.g., rs-fMRI for fatigue, fMRI and FDG-PET for mild cognitive impairment. In addition, studies have shown that quantitative iron imaging techniques such as R2*, SWI, and QSM are reliable markers of iron content in PD. These measurements have also been found to correlate with the severity of motor symptoms. Among these techniques, QSM has been identified as more robust and reproducible than R2* and is more adequate for use in multicenter studies (Pyatigorskaya et al., 2020). Finally, some authors have even discouraged the routine use of neuroimaging techniques in clinical practice for PD (Pagano et al., 2016). As stated by Pagano et al. (2016), "despite significant evidence for the utility of neuroimaging in assessing parkinsonian patients, none of the neuroimaging techniques is specifically recommended for routine use in clinical practice."

Therefore, the impact of this variety is threefold. First, the symptoms may not associate with structural or functional brain patterns. Second, when existing, such patterns require particular brain imaging modalities. Finally, such patterns may not be specific to PD. On top of these three circumstances, the temporal evolution of the disease adds another layer of complexity. Each stage calls for different symptoms, which in turn require dedicated imaging modalities with different diagnosis specificity. In this light, accurate PD subtyping becomes challenging, as obtaining a complete view of the brain manifestations of PD symptoms requires image acquisition of several modalities or the employment of multimodal approaches (Saeed et al., 2017; Chougar et al., 2020; Albrecht et al., 2022).

One potential solution to address this issue is to use a combination of multiple imaging techniques. Multimodal approaches can provide a more complete and accurate picture of the disease by capturing different aspects of brain function and structure, as well as the density of neurotransmitter receptors such as dopamine receptors. Additionally, clinical assessments can be supplemented by specific neuropsychological questionnaires or physiological tests, with subsequent confirmation by imaging or a biochemical marker, as different modalities are suitable at different stages of disease progression (Michell et al., 2004). Moreover, the use of multi-modal data, combining clinical, motor, cognitive, and neuroimaging data, can aid in subtyping PD and potentially identifying correlations between the pathology manifested in the brain and the motor and non-motor symptoms of the patient (Albrecht et al., 2022). However, it is important to note that using multiple imaging modalities can also pose some challenges, such as the need for specialized expertise, the complexity of data integration (Behrad and Abadeh, 2022), and the increased cost and time required for imaging and analysis.

Lack of standardization in acquisition, preprocessing, and annotation pipelines

After image acquisition, another set of problems may compromise research. First, variations in the acquisition parameters may alter the observed changes in longitudinal studies. Chua et al. (2015) showed how variability in MRI acquisition parameters between scans can confound observations. Then, the diversity of preprocessing pipelines across studies presents another dimension for potential unwanted interactions and errors. For instance, the exclusion criteria for head motion may vary across studies without common criteria. Strother (2006) highlighted how

the preprocessing steps interact with every decision taken during the design and execution of fMRI experiments. The authors argue that "applying a new processing pipeline to a raw dataset may result in significantly modified spatial activation patterns as a result of changing/optimizing preprocessing techniques and/or the data analysis approach." Similarly, Power et al. (2017) identified several contributors to global fMRI signals such as hardware artifacts and head motion that were not removed from scans through denoising techniques, affecting the observed covariances. Bhagwat et al. (2021) underscored the variability introduced by preprocessing in neuroimaging pipelines. Hence, the lack of standardization in acquisition, preprocessing, and annotation pipelines can lead to unwanted interactions and errors, which has significant implications for the reliability and reproducibility of neuroimaging research (Brauneck et al., 2023).

To address this issue, it is crucial to develop and validate standardized protocols and criteria for data acquisition, preprocessing, and analysis. This can be achieved through a variety of approaches, such as establishing international consortia, promoting open data sharing, and providing training and resources for researchers. For example, the International Society for Magnetic Resonance in Medicine (ISMRM) has developed several standards for MRI data acquisition and analysis, including quantitative MR (Weingärtner et al., 2022). In addition, promoting open data sharing and encouraging researchers to openly share their raw data and analysis pipelines can help to identify potential sources of variability and errors in data processing and analysis. This can facilitate the development of more robust and reliable methods for data preprocessing and analysis. Several initiatives have already been developed to promote open data sharing in neuroimaging, such as the OpenfMRI (Poldrack et al., 2013) and NeuroVault (Gorgolewski et al., 2015) repositories. Furthermore, educating researchers about the importance of standardization in neuroimaging research (Laird et al., 2011) and providing them with the necessary tools and resources to implement standardized protocols and criteria in their research is crucial, including standardization of the metadata as a way to reflect the causal and anti-causal assumptions made during the data collection and annotation (Garcia Santa Cruz et al., 2022b). Further, standardization of the annotation pipeline is important to improve the consistency and quality of annotations. To tackle this issue, it is important to have standardized guidelines and procedures. This can reduce misinterpretation, which may result in inconsistency, making the subsequent training of the machine learning solution difficult (Miceli et al., 2020). Additionally, it's crucial to have a good way to integrate annotations from multiple annotators, carefully considering how to deal with labeling merging in unmatched results when and the seniority of the experts. Furthermore, as labeling is an expensive task, unsupervised or semi-supervised techniques could be employed to generate cheaper but potentially more consistent labels (dos Santos Ferreira et al., 2019).

To fully exploit the potential for personalized healthcare, collecting metadata may be necessary. However, current General Data Protection Regulation (GDPR) regulations impose limitations to ensure both data privacy and security. To address this challenge, several approaches have been proposed, including federated machine learning, multi-party computation, and differential privacy. These methods provide a win-win solution by enabling the collection of necessary data while preserving the privacy and security of sensitive information (Brauneck et al., 2023).

This can be achieved through training programs, workshops, and online resources that provide guidance on best practices for data acquisition, preprocessing, and analysis in neuroimaging (Borghi and Van Gulick, 2018). The development of established protocols in standardization and analysis, such as those proposed for other neurodegenerative diseases like the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Wyman et al., 2013), can also serve as important models for promoting consistency and reliability in neuroimaging research.

Limitations associated with machine learning/deep learning

Generalization issues that hinder transferability

Neural networks (NNs) have been shown to be highly effective in approximating complex functions and achieving accurate predictions by leveraging large and high-quality datasets. However, despite demonstrating good performance on the training data, there is no guarantee that the model will continue to perform well on new and unseen data. This phenomenon, known as overfitting, occurs when the model is too closely tailored to the training data, and thus, is not generalizable to new data. Out-of-distribution and out-of-domain examples can cause neural networks to learn incorrect correlations and make inaccurate predictions. Common causes of overfitting include domain shift (Kondrateva et al., 2021), task mismatch (Castro et al., 2020), and catastrophic forgetting (Gupta et al., 2021). Poor generalization can lead to unreliable and incorrect predictions on real-world tasks where the data distribution may differ significantly from the training data (Yagis et al., 2019; Ge et al., 2023). In the context of CAD for PD, this may result in incorrect predictions that could lead to misdiagnosis or failure to detect the disease, ultimately resulting in incorrect treatment or delayed diagnosis.

To reduce overfitting, techniques such as regularization (Kukačka et al., 2017) and early stopping (Prechelt, 1998) can be employed. Data augmentation techniques can also expand the dataset size and improve internal generalization (Chlap et al., 2021). However, data augmentation alone cannot address demographic representativeness issues. Thorough internal and external validation is essential to ensure reliable and accurate model performance, especially for new and unseen data (Garcia Santa Cruz et al., 2021). Cross-validation techniques such as stratified cross-validation (Zeng and Martinez, 2000) and leave-one-out cross-validation (Hastie et al., 2009) can be used for internal validation, while external validation can be achieved through external datasets. These techniques can enhance model transferability and promote generalizability.

Additionally, when dealing with a small sample size, as is often the case in biomedical datasets, splitting the dataset for cross-validation may lead to a loss of the algorithm's generalization capacity. This limitation arises from the fact that when the sample size is small, dividing it into training and validation sets further reduces the amount of data available for training, potentially hindering the algorithm's ability to generalize well. Despite the

conventional wisdom that attributes this small generalization error to properties of the model family or regularization techniques used during training (Zhang et al., 2021), it has been demonstrated that even with explicit regularization, state-of-the-art convolutional networks can fit random labeling of the training data, suggesting that these models have enough capacity to memorize the training data. A potential solution is to employ distribution-free performance bounds (Jakubovitz et al., 2019), which have been successfully implemented in neuroimaging (Górriz et al., 2019; Jimenez-Mesa et al., 2023).

To address data drift, various techniques can be employed. Calibration techniques (Wald et al., 2021) and appropriate metrics for evaluating model generalization (Jiang et al., 2019) can be used. Additionally, selecting the appropriate model architecture and hyperparameters can significantly enhance the model's generalization ability. Techniques such as grid search or Bayesian optimization (Kandasamy et al., 2018) can be employed to optimize hyperparameters. Furthermore, transfer learning has been demonstrated as an effective approach for improving model generalization, particularly when working with limited data (Yosinski et al., 2014).

Another big issue that can hinder the generalization of models is when they fail to learn the desirable patterns that characterize the phenomena we are trying to model, and instead learn spurious correlations. This can result in the model learning potential confounders, colliders, and other unwanted biases.

To address these issues, it is important to carefully evaluate the data used to train the model, identify potential confounders and colliders biases, and use appropriate statistical methods to account for them (Wang et al., 2018). Additionally, confounding removal strategies such as domain adaptation techniques can be employed during the harmonization phase (Dinsdale et al., 2021) and during the training process (Qin et al., 2020). Finally, it is crucial to regularly monitor the performance of the model and validate its results against independent and temporally updated data sets to identify and correct potential unwanted biases (Tamburri, 2020).

Algorithmic bias

This can be considered an extension of a generalization issue. Algorithmic bias is another significant challenge in ML, particularly in medical diagnosis and other decision-making applications. Societal biases and data acquisition biases can result in systematic and repeatable errors that lead to unfair outcomes and lower accuracy for certain groups (Ricci Lara et al., 2022). It is essential to address these biases in the design, training, and evaluation of NNs to ensure fairness and avoid perpetuating existing inequalities. These biases can result in systematic and repeatable errors, leading to unfair outcomes that favor certain groups over others, ultimately lowering the accuracy of the recommendation for some patient groups, particularly when there are racial biases. These biases can originate from existing inequality (Ricci Lara et al., 2022) or can also stem from selection bias introduced during the acquisition process (Garcia Santa Cruz et al., 2022b).

For example, Obermeyer et al. (2019) identified some systemic conditional disparities in risk scores based on the medical history of Black patients. In such cases, bias-correcting techniques can be employed (Wiens et al., 2020). Bias can also be introduced during the data acquisition process, resulting in technical debt and downstream effects known as data cascades (Sambasivan et al., 2021). Moreover, it is essential to address the issue of unwanted biases in the data used for current AI systems, as these systems not only have the risk of making incorrect predictions, but also of perpetuating and amplifying biases present in the data (Zhao et al., 2017).

The ML community has made interdisciplinary efforts to address the aforementioned issues, leading to the development of a range of solutions that fall under the umbrella of fairness (Mehrabi et al., 2021). By implementing such strategies in algorithm design, training, and evaluation, performance across groups can be improved, thereby mitigating the risk of unfairness in the final solution. These solutions typically target characteristics that have traditionally been the source of unfair discrepancies, such as gender and ethnicity. However, it is also crucial to ensure that algorithms perform well in cases where diseases have subgroups, such PD subtypes (Thenganatt and Jankovic, 2014) and varying degrees of disease penetrance (Espay et al., 2017). In such cases, similar metrics can be used, with the subgroups or disease penetrance considered as protected attributes.

Need for better interpretability

Another significant issue with NNs is their inability to accurately represent uncertainty in their predictions (Abdar et al., 2021). Since NNs are deterministic, they cannot capture the notion of what they know and what they do not know, or the confidence level of their predictions. Furthermore, current NNs are limited to accessing the knowledge contained in the dataset. This lack of uncertainty estimation can lead to overconfidence in their predictions, which can be problematic in critical applications such as medical diagnosis or self-driving cars.

Before implementing CAD systems for PD as decision-making tools in clinical practice, it is essential to establish an interpretability strategy (Chan et al., 2020). CAD systems with low interpretability can have severe consequences, such as decreased trust and acceptance among clinicians and patients, misdiagnoses, and ineffective treatment strategies. A transparent and understandable model can help clinicians validate the system's predictions and ensure that the model is not making decisions based on spurious correlations or biases. Additionally, interpretability can help researchers gain new insights into PD and refine the diagnostic criteria.

The lack of uncertainty estimation can lead to overconfidence in their predictions, which can have severe consequences such as misdiagnoses and ineffective treatment strategies. Therefore, it is essential to establish an interpretability strategy before implementing CAD systems in clinical practice. Furthermore, the limitations of current explainability methods used in ML decision-making systems suggest that unless there are significant advances in explainable ML, we must treat these systems as black boxes, justified by their reliable and experimentally confirmed performance. Finally, it is recommended that healthcare workers exercise caution when using explanations from ML systems and regulators be judicious in listing explanations among the requirements needed for clinical deployment of ML (Ghassemi et al., 2021).

Recent regulations, such as the GDPR in the European Union (EU), emphasize the right to be informed and the right to contest an automated decision. In such cases, interpretability of AI becomes crucial for auditing the decision-making of automated agents such as ML models. In particular, Article 22 of the GDPR deals with the rights related to automated individual decision-making since data subjects cannot be subject to a decision based solely on automated processing (Council of European Union, 2016). Additionally, Articles 12 and 13 specify the right to be informed about the use of their data in an easily understandable and accessible manner. The most common use cases for participant data fall into two main scenarios: (1) data subjects provide their data to train AI models, and (2) data subjects receive a result from an AI model after providing some data. The first scenario requires informing the participants about the purpose and usage of their data. However, the second scenario requires additional clarification, as the participants should understand how a decision was made and, in particular, which input data was relevant for obtaining a specific result.

To meet the above requirements, ML solutions must be designed with transparency in mind. Some ML approaches produce models that are inherently easier to inspect. Decision tree predictive models are popular due to their intelligibility and simplicity. However, this approach does not suit all tasks. Essentially, models optimize a function that draws the boundary to separate the given classes (e.g., healthy vs. diseased) by grouping nearby instances. However, the definition of proximity differs across ML learners and interpretability measures become complex. For instance, random forest methods constitute an evolution of decision trees but at the cost of intrinsic interpretability since their internal model consists of a collection of decision trees, obfuscating the "reasoning" of the trained model (Nair et al., 2013). Another approach includes tracking the decision-making process on CNNs. For instance, Magesh et al. (2020) employ Local Interpretable Model-Agnostic Explainer (LIME) to increase the explainability of CNN-based models for PD diagnosis. Two key elements to improve interpretability are solutions to improve model *explainability* and model *uncertainty*.

Model explainability

Model explainability refers to the ability to understand how a ML model makes its predictions. It is important because in critical applications, such as healthcare or finance, it is necessary to understand why the model makes certain decisions, especially when human lives or significant resources are at stake. For example, if a model is predicting whether a patient has PD or make a recomendation about the treatment, it is important to know which factors the model is considering in its decision-making process.

Explainability and interpretability terms are is frequently used interchangeably and for this work, we do not distinguish between them. Of course, interpretability tools vary across ML methods, but there are some important methods worth mentioning that can facilitate the interpretability of the results. Molnar (2020) provides an overview of the available techniques for ML interpretability. The author distinguishes between intrinsic and *post hoc* methods. The first group concerns models whose simple structure permits human interpretation, e.g., short decision trees. The second group of methods are used after model training. Additionally, the author divides interpretability methods into model-specific and model-agnostic. The author provides yet another criterion to separate the methods into two groups, i.e., local (for methods that explain a particular result) and global (for methods that explain the whole model behavior) interpretability.

Aside from the above, solution design can impact model interpretability as well. Often models are designed in an end-to-end way that attempts to map input data with the final result with a single model. For instance, a medical imaging CAD system can be designed as a chain of several models, with the first dedicated to finding pathologies and the subsequent models mapping pathologies to diseases or conditions (e.g., through several one class classifiers) (Vega, 2021). This approach eases solution maintainance and increases interpretability, allowing inspection of the intermediate results.

To address this challenge, researchers have proposed various methods for interpreting and explaining the decisions of ML models, including model-agnostic techniques such as LIME (Visani et al., 2022) and SHapley Additive exPlanations (SHAP) (Kaur et al., 2020), as well as model-specific approaches such as attention mechanisms (Vaswani et al., 2017) and gradient-based attribution methods (Ancona et al., 2019).

Model uncertainty

In the context of medical diagnosis, the concept of model uncertainty plays a crucial role in determining the degree of confidence or uncertainty that a model has in its predictions. This consideration is particularly pertinent given the high stakes involved in clinical decision-making. The degree of certainty or uncertainty in a model's output is a crucial factor in determining appropriate actions to be taken based on the model's predictions. As such, accounting for model uncertainty can enhance the transparency and reliability of medical diagnosis, leading to more effective treatment strategies and improved patient outcomes.

Uncertainty in ML can stem from multiple sources. Some of them include data variance, lack of representativity in the data sample, label noise, and the intrinsic imperfections of any ML model developed from such data. The literature also refers to these types of uncertainty as systematic, aleatoric and epistemic, (Hüllermeier and Waegeman, 2021; Gal et al., 2022). Most of these issues cannot be fixed *a posteriori* and must be avoided through careful data acquisition design. However, documenting uncertainty sources and quantifying its magnitude in data, labels and model is of uttermost importance, in the same way we should document other aspects such as the representativity of the sample. This information is key to assess the generalization power of the solutions to new settings. For instance, reporting probability estimates together with the model prediction can indicate the model prediction confidence. However, these estimates may not accurately reflect model uncertainity calling for calibration methods (Lemay et al., 2022).

Costly systems to develop and maintain

ML solutions are also expensive in terms of data and computation. Developing and training ML models requires a

substantial amount of data, computing power, and specialized expertise. Acquiring large and diverse datasets can be challenging, and data collection, cleaning, and preprocessing can be time-consuming and labor-intensive (Ngiam and Khor, 2019). Moreover, the development and training of ML models often require specialized hardware, such as Graphics Processing Units (GPUs), which can increase energy consumption and carbon footprint (Patterson et al., 2021). It is important to consider the environmental impact of ML and take steps to reduce it, such as using energy-efficient hardware or exploring alternative training methods that require fewer computing resources (Wang et al., 2020).

In addition, ML models require ongoing monitoring, updating, and maintenance to ensure their continued accuracy. As data changes over time, the models may need to be retrained or updated to account for new patterns or trends. In the case of PD, this can be particularly challenging due to the variability in disease progression across patients, making it difficult to develop models that accurately capture the underlying patterns of the disease. Furthermore, implementing ML systems in clinical practice requires careful consideration of regulatory and ethical concerns to ensure patient safety and privacy. ML models used in clinical practice must undergo rigorous testing and validation to ensure their safety, efficacy, and reliability. The validation process involves evaluating the model's performance on independent datasets and comparing it to other established diagnostic methods (Liu et al., 2019). Additionally, models must be regularly audited to identify and mitigate biases and errors that may affect their performance (Reddy et al., 2020).

To address the challenges of cost and development associated with ML, there has been a concerted effort to develop open-source platforms and tools that make ML more accessible to researchers and clinicians. For instance, several open-source libraries, including TensorFlow (Abadi et al., 2016a), PyTorch (Paszke et al., 2019) and MONAI (Cardoso et al., 2022) provide pre-built ML models and algorithms that can be readily adapted and customized for specific applications. In addition, cloud computing platforms, such as Amazon Web Services and Google Cloud, offer scalable and cost-effective solutions for training and deploying ML models. Moreover, there is a growing trend toward collaborative and decentralized approaches to ML development (Castiglioni et al., 2021). One such approach is federated learning, which allows multiple parties to train a shared ML model without sharing their data, thus preserving data privacy and security (Tedeschini et al., 2022). Another approach is to use blockchain technology to create decentralized ML models that are transparent, auditable, and resistant to tampering (Neelakandan et al., 2022). These developments are expected to enhance the accessibility and affordability of ML solutions, thereby facilitating their wider adoption and implementation in clinical practice.

Security and privacy challenges

Healthcare institutions are frequent targets of malicious hackers, resulting in data breaches and ransomware attacks (Branch et al., 2019; Devi, 2023). In March 2023, the Hospital Clinic de Barcelona, which serves half a million people, suffered a ransomware attack by the RansomHouse group, resulting in the theft of 4.4 TB of data (Toulas, 2023). Healthcare ML models often deal with very sensitive patient data, making them attractive targets for malicious attacks.

Adversarial training is a technique used to improve the robustness of ML models against adversarial attacks (Madry et al., 2017). It involves training the model on adversarial examples generated by an adversary system to make the model more resilient to similar attacks. However, these techniques can also be used maliciously. Adversarial attacks can cause the model to make incorrect predictions, which could potentially expose personal information from healthcare ML models. In membership inference attacks, an adversary attempts to determine whether a particular individual's data was used to train a machine learning model (Hu et al., 2022). In model inversion attacks, the aim is to reconstruct an individual's data from the outputs of a machine learning model. This can be achieved by generating adversarial examples that maximize the likelihood of the individual's data, given the model outputs (Fredrikson et al., 2015). These attacks highlight the need for robust security measures to be in place to protect healthcare ML models from malicious attacks.

The most effective safety measure for healthcare ML models is to restrict access to the trained models to authorized personnel. Additionally, privacy-preserving machine learning techniques such as differential privacy and homomorphic encryption can help prevent these attacks (Abadi et al., 2016b; Aono et al., 2017). It is advisable to take a proactive approach to healthcare privacy and security during the solution design instead of a reactive approach (Song et al., 2019; Bhuyan et al., 2020).

Concluding remarks and perspectives

During recent years, both the ML and the medical community have begun to consider data quality as the most crucial factor impacting the performance of the solutions and their robustness, (Sambasivan et al., 2021). However, acquiring high-quality data, building a suitable model for the task, and determining the appropriate use for such models, remain challenging objectives toward clinically relevant models. In particular, Sambasivan et al. (2021) insist on building incentive structures across all stakeholders, stating that "many practitioners described data work as time-consuming, invisible to track, and often done under pressures to move fast due to margins–investment, constraints, and deadlines often came in the way of focusing on improving data quality." Data bootstrapping is yet another source of issues in high-stakes AI domains, as many researchers begin the AI/ML work employing existing data or data collected for non-AI purposes that leads to poor generalization. It is essential to ensure that ML models are rigorously validated and tested before they can be used in clinical practice. The employment of datasets from multiple independent studies can boost the statistical power and lead to more accurate, reliable and reproducible research. In ML, a common practice to this end is to mix several datasets. However, if the mixed datasets do not share certain degree of methodological similarity, biases may be introduced due to differences in acquisition, preprocessing or annotation.

The circumstances previously described hinder the availability of large datasets containing multiple imaging modalities as large datasets often consist of multi-center cohorts employing different acquisition devices, protocols and pipelines. Overall, developing and maintaining ML systems for clinical practice can be a costly and time-consuming process that requires significant expertise and resources. However, the potential benefits, such as improved diagnosis and treatment outcomes for patients with PD, make it a worthwhile investment. The use of CAD tools to interpreted brain images is the context of PD is very promising. However, as previously mentioned, these solutions will be used as assisting tools in a very specific context and under specialized supervision and must pass a series of verification before they can be used, as is the case with other medical products or treatments. To achieve this, the models must be accompanied by interpretability methods to ensure that clinicians can understand how the model makes its predictions.

While this review focuses primarily on brain imaging, it has become increasingly clear that a single measure is unlikely to be sufficient for diagnosing PD in the foreseeable future. Instead, a combination of measures will likely be necessary. The most critical aspect of a biomarker is not its ability to diagnose PD in its early stages, but rather its ability to reflect the disease's pathogenesis and progression. By using a multimodal approach that combines various imaging biomarkers, clinicians can make early, accurate, and objective diagnostic decisions, identify neuroanatomical and pathophysiological mechanisms, and evaluate disease progression and therapeutic responses to drugs in clinical trials.

A common approach in developing multimodal CAD systems involves combining multiple imaging modalities as well as leveraging ensemble learning to integrate data from various sources for obtaining the final result. A concrete example of a multimodal approach in PD is the employment of multiple modalities to characterize a specific pathological process in certain regions of the brain. For instance, multimodal approaches employing hybrid images created through the integration of different MRI parameters offer a valuable tool. By combining T1-, T2*-, and diffusion-weighted MRI, Barbagallo et al. (2016) proposed to enable the detection and analysis of macro- and micro-structural abnormalities in the nigrostriatal pathway. The key benefit of integrating hybrid images enhances the accuracy and reliability of CAD systems by capturing diverse aspects of neurodegeneration.

Another example of a multimodal approach consists in combining MRI techniques, particularly those visualizing pathological changes in the substantia nigra using diffusion, iron-sensitive susceptibility, and neuromelanin-sensitive sequences, which offer a more accessible imaging tool. However, these techniques may be insufficient for phenotyping or prognostication due to the heterogeneous nature of PD resulting from extranigral pathologies. In Siderowf et al. (2023) highlight the emerging role of retinal optical coherence tomography as a non-invasive technique to visualize structural changes in the retina, which can serve as potential biomarkers for early diagnosis and prognostication in PD. Ensemble learning, a popular technique employed in multimodal CAD systems, plays a crucial role in fusing information from diverse data sources. Through ensemble learning, multiple models are trained independently on different subsets of data or using distinct feature representations. Ensemble learning had been successfully applied in PD classification using multimodal voice and speech data (Ali et al., 2021).

Recent promising markers that use the biochemistry of alpha-synuclein seed amplification assays have shown potential (Siderowf et al., 2023). For instance when recommending DBS as a therapy option for PD, it is important to consider genetic information, specifically whether the patient is a carrier of mutations in the glucocerebrosidase (GBA) gene. PD patients with GBA mutations are at particularly high risk for cognitive impairment with DBS due to dysfunction of the glucocerebrosidase (GCase) enzyme, resulting in more rapid accumulation and spread of Lewy bodies. Recent research has shown that PD patients experience cognitive impairment after DBS, and this risk is even greater for those with GBA mutations. Therefore, models that assist with therapy recommendations for PD patients should carefully evaluate whether patients are carriers of GBA mutations before recommending DBS as a treatment option (Pal et al., 2022).

Furthermore, there is an extended literature of ML models that have the potential to become CAD systems in the future from diagnosis and monitoring of PD, by providing more accurate and objective measurements of motor symptoms and disease progression. However, until this model are properly validated there are far to be ready for its used in clinical settings to ensure their safety and effectiveness in clinical practice.

Ultimately, our review emphasizes the critical importance of taking a multidisciplinary approach and putting in extensive effort during the data preparation and clinical validation phases of developing ML models. It is crucial to recognize that proper design and clinical validation may be undervalued in comparison to the training of ML models, but they are indispensable for data-driven CAD solutions that are safe for a clinical use. We hope that this review will inspire both future users and developers of these systems in the context of MRI for PD.

Author contributions

BG conceived, structured, and wrote the manuscript. AH and FH supervised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

BG was supported by the FNR within the PARK-QC DTU (PRIDE17/12244779/PARK-QC) and Pelican award from the Fondation du Pelican de Mie et Pierre Hippert-Faber, under the aegis of the Fondation de Luxembourg.

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.

Publisher's note

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., et al. (2016b). "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Savannah, GA), 308–318. doi: 10.1145/2976749.2978318

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016a). "Tensorflow: a system for large-scale machine learning," in *OSDI*, Volume 16 (Savannah, GA), 265–283.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* 76, 243–297. doi: 10.1016/j.inffus.2021.05.008

Adeli, E., Thung, K.-H., An, L., Wu, G., Shi, F., Wang, T., et al. (2018). Semi-supervised discriminative classification robust to sample-outliers and feature-noises. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 515–522. doi: 10.1109/TPAMI.2018.2794470

Akdemir, Ü. Ö., Bora, H. A. T., and Atay, L. Ö. (2021). Dopamine transporter spect imaging in Parkinson's disease and parkinsoniandisorders. *Turk. J. Med. Sci.* 51, 400–410. doi: 10.3906/sag-2008-253

Albrecht, F., Poulakis, K., Freidle, M., Johansson, H., Ekman, U., Volpe, G., et al. (2022). Unraveling Parkinson's disease heterogeneity using subtypes based on multimodal data. *Parkinsonism Relat. Disord.* 102, 19–29. doi: 10.1016/j.parkreldis.2022.07.014

Ali, L., He, Z., Cao, W., Rauf, H. T., Imrana, Y., Bin Heyat, M. B., et al. (2021). MMDD-ensemble: a multimodal data-driven ensemble approach for Parkinson's disease detection. *Front. Neurosci.* 15, 754058. doi: 10.3389/fnins.2021.754058

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, eds W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. R. Müller (Cham: Springer), 169–191. doi: 10.1007/978-3-030-28954-6_9

Aono, Y., Hayashi, T., Wang, L., Moriai, S., et al. (2017). Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans. Inf. Forensics Secur.* 13, 1333–1345. doi: 10.1109/TIFS.2017.2787987

Arroyo-Gallego, T., Ledesma-Carbayo, M. J., Butterworth, I., Matarazzo, M., Montero-Escribano, P., Puertas-Martín, V., et al. (2018). Detecting motor impairment in early Parkinson's disease via natural typing interaction with keyboards: validation of the neuroqwerty approach in an uncontrolled at-home setting. *J. Med. Internet Res.* 20, e89. doi: 10.2196/jmir.9462

Augimeri, A., Cherubini, A., Cascini, G. L., Galea, D., Caligiuri, M. E., Barbagallo, G., et al. (2016). Coflupane in diagnosi–computer-aided datscan analysis. *EJNMMI Phys.* 3, 1–13. doi: 10.1186/s40658-016-0140-9

Bajaj, N., Hauser, R. A., and Grachev, I. D. (2013). Clinical utility of dopamine transporter single photon emission CT (DAT-SPECT) with (123I) ioflupane in diagnosis of parkinsonian syndromes. *J. Neurol. Neurosurg. Psychiatry* 84, 1288–1295. doi: 10.1136/jnnp-2012-304436

Barbagallo, G., Sierra-Peña, M., Nemmi, F., Traon, A. P.-L., Meissner, W. G., Rascol, O., et al. (2016). Multimodal MRI assessment of nigro-striatal pathway in multiple system atrophy and Parkinson disease. *Mov. Disord.* 31, 325–334. doi: 10.1002/mds.26471

Behrad, F., and Abadeh, M. S. (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Syst. Appl.* 200, 117006. doi: 10.1016/j.eswa.2022.117006

Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., et al. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience* 10, giaa155. doi: 10.1093/gigascience/giaa155

Bhuyan, S. S., Kabir, U. Y., Escareno, J. M., Ector, K., Palakodeti, S., Wyant, D., et al. (2020). Transforming healthcare cybersecurity from reactive to proactive: current status and future recommendations. *J. Med. Syst.* 44, 1–9. doi: 10.1007/s10916-019-1507-y

Biondetti, E., Gaurav, R., Yahia-Cherif, L., Mangone, G., Pyatigorskaya, N., Valabrègue, R., et al. (2020). Spatiotemporal changes in substantia nigra neuromelanin content in Parkinson's disease. *Brain* 143, 2757–2770. doi: 10.1093/brain/awaa216

Blauwendraat, C., Nalls, M. A., and Singleton, A. B. (2020). The genetic architecture of Parkinson's disease. *Lancet Neurol.* 19, 170–178. doi: 10.1016/S1474-4422(19)30287-X

Borghammer, P., and Van Den Berge, N. (2019). Brain-first versus gut-first Parkinson's disease: a hypothesis. *J. Parkinsons Dis.* 9, S281–S295. doi: 10.3233/JPD-191721

Borghi, J. A., and Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: practices and perceptions of MRI researchers. *PLoS ONE* 13, e0200562. doi: 10.1371/journal.pone.0200562

Boutet, A., Madhavan, R., Elias, G. J., Joel, S. E., Gramer, R., Ranjan, M., et al. (2021). Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. *Nat. Commun.* 12, 3043. doi: 10.1038/s41467-021-23311-9

Branch, L., Eller, W., Bias, T., McCawley, M., Myers, D., Gerber, B., et al. (2019). Trends in malware attacks against united states healthcare organizations, 2016-2017. *Glob. Biosecur.* 1, 15–24. doi: 10.31646/gbio.7

Brauneck, A., Schmalhorst, L., Kazemi Majdabadi, M. M., Bakhtiari, M., Völker, U., Baumbach, J., et al. (2023). Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: scoping review. *J. Med. Internet Res.* 25, e41588. doi: 10.2196/41588

Broeder, S., Nackaerts, E., Heremans, E., Vervoort, G., Meesen, R., Verheyden, G., et al. (2015). Transcranial direct current stimulation in Parkinson's disease: neurophysiological mechanisms and behavioral effects. *Neurosci. Biobehav. Rev.* 57, 105–117. doi: 10.1016/j.neubiorev.2015.08.010

Cantello, R., Tarletti, R., and Civardi, C. (2002). Transcranial magnetic stimulation and Parkinson's disease. *Brain Res. Rev.* 38, 309–327. doi: 10.1016/S0165-0173(01)00158-8

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., et al. (2022). MONAI: an open-source framework for deep learning in healthcare. *arXiv*. [preprint]. doi: 10.48550/arXiv.2211.0270

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, F., et al. (2021). AI applications to medical images: from machine learning to deep learning. *Phys. Med.* 83, 9–24. doi: 10.1016/j.ejmp.2021.02.006

Castillo-Barnes, D., Martinez-Murcia, F. J., Ortiz, A., Salas-Gonzalez, D., RamÍrez, J., and Górriz, J. M. (2020). Morphological characterization of functional brain imaging by isosurface analysis in Parkinson's disease. *Int. J. Neural Syst.* 30, 2050044. doi: 10.1142/S0129065720500446

Castillo-Barnes, D., Ramírez, J., Segovia, F., Martínez-Murcia, F. J., Salas-Gonzalez, D., and Górriz, J. M. (2018). Robust ensemble classification methodology for I123-ioflupane spect images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's disease. *Front. Neuroinform.* 12, 53. doi: 10.3389/fninf.2018.00053

Castro, D. C., Walker, I., and Glocker, B. (2020). Causality matters in medical imaging. *Nat. Commun.* 11, 3673. doi: 10.1038/s41467-020-17478-w

Chakraborty, S., Aich, S., and Kim, H.-C. (2020). Detection of Parkinson's disease from 3t t1 weighted MRI scans using 3D convolutional neural network. *Diagnostics* 10, 402. doi: 10.3390/diagnostics10060402

Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Med. Phys.* 47, e218–e227. doi: 10.1002/mp.13764

Chaudhuri, K. R., Healy, D. G., and Schapira, A. H. (2006). Non-motor symptoms of Parkinson's disease: diagnosis and management. *Lancet Neurol.* 5, 235–245. doi: 10.1016/S1474-4422(06)70373-8

Chen, C.-M., Chou, Y.-H., Tagawa, N., and Do, Y. (2013). Computer-aided detection and diagnosis in medical imaging. *Comput. Math. Methods Med.* 2013, 790608. doi: 10.1155/2013/790608

Chen, H., and Ritz, B. (2018). The search for environmental causes of Parkinson's disease: moving forward. *J. Parkinsons Dis.* 8, S9–S17. doi: 10.3233/JPD-181493

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., et al. (2021). A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* 65, 545–563. doi: 10.1111/1754-9485.13261

Chougar, L., Faouzi, J., Pyatigorskaya, N., Yahia-Cherif, L., Gaurav, R., Biondetti, E., et al. (2021). Automated categorization of parkinsonian syndromes using magnetic resonance imaging in a clinical setting. *Mov. Disord.* 36, 460–470. doi: 10.1002/mds.28348

Chougar, L., Pyatigorskaya, N., Degos, B., Grabli, D., and Lehéricy, S. (2020). The role of magnetic resonance imaging for the diagnosis of atypical parkinsonism. *Front. Neurol.* 11, 665. doi: 10.3389/fneur.2020.00665

Chua, A. S., Egorova, S., Anderson, M. C., Polgar-Turcsanyi, M., Chitnis, T., Weiner, H. L., et al. (2015). Handling changes in MRI acquisition parameters in modeling whole brain lesion volume and atrophy data in multiple sclerosis subjects: comparison of linear mixed-effect models. *Neuroimage Clin.* 8, 606–610. doi: 10.1016/j.nicl.2015.06.009

Cohen, A. A., Ferrucci, L., Fülöp, T., Gravel, D., Hao, N., Kriete, A., et al. (2022). A complex systems approach to aging biology. *Nat. Aging* 2, 580–591. doi: 10.1038/s43587-022-00252-6

Coleman, W. B., and Tsongalis, G. J. (2009). *Molecular Pathology: The Molecular Basis of Human Disease*. Cambridge, MA: Academic Press.

Constantinescu, R., and Mondello, S. (2013). Cerebrospinal fluid biomarker candidates for parkinsonian disorders. *Front. Neurol.* 3, 187. doi: 10.3389/fneur.2012.00187

Cools, R. (2006). Dopaminergic modulation of cognitive function-implications for l-dopa treatment in Parkinson's disease. *Neurosci. Biobehav. Rev.* 30, 1–23. doi: 10.1016/j.neubiorev.2005.03.024

Council of European Union (2016). *General Data Protection Regulation*. Avaialble online at: http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502anduri=CELEX:32014R0269 (accessed July 1, 2023).

da Silva, F. C., Iop, R. R., de Oliveira, L. C., Boll, A. M., de Alvarenga, J. G. S., Gutierres Filho, P. J. B., et al. (2018). Effects of physical exercise programs on cognitive function in Parkinson's disease patients: a systematic review of randomized controlled trials of the last 10 years. *PLoS ONE* 13, e0193113. doi: 10.1371/journal.pone.0193113

Dadu, A., Satone, V., Kaur, R., Hashemi, S. H., Leonard, H., Iwaki, H., et al. (2022). Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ Parkinsons Dis.* 8, 172. doi: 10.1038/s41531-022-00439-z

De Pablo-Fernández, E., Lees, A. J., Holton, J. L., and Warner, T. T. (2019). Prognosis and neuropathologic correlation of clinical subtypes of Parkinson disease. *JAMA Neurol.* 76, 470–479. doi: 10.1001/jamaneurol.2018.4377

Deeb, W., Nozile-Firth, K., and Okun, M. S. (2019). Parkinson's disease: diagnosis and appreciation of comorbidities. *Handb. Clin. Neurol.* 167, 257–277. doi: 10.1016/B978-0-12-804766-8.00014-5

Devi, S. (2023). Cyber-attacks on health-care systems. *Lancet Oncol.* 24, e148. doi: 10.1016/S1470-2045(23)00119-5

Ding, P., and Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of m-bias and butterfly-bias. *J. Causal Inference* 3, 41–57. doi: 10.1515/jci-2013-0021

Dinsdale, N. K., Jenkinson, M., and Namburete, A. I. (2021). Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *Neuroimage* 228, 117689. doi: 10.1016/j.neuroimage.2020.117689

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput. Med. Imaging Graph.* 31, 198–211. doi: 10.1016/j.compmedimag.2007.02.002

dos Santos Ferreira, A., Freitas, D. M., da Silva, G. G., Pistori, H., and Folhes, M. T. (2019). Unsupervised deep learning and semi-automatic data labeling in weed discrimination. *Comput. Electron. Agric.* 165, 104963. doi: 10.1016/j.compag.2019.104963

Durcan, R., Wiblin, L., Lawson, R. A., Khoo, T. K., Yarnall, A., Duncan, G. W., et al. (2019). Prevalence and duration of non-motor symptoms in prodromal Parkinson's disease. *Eur. J. Neurol.* 26, 979–985. doi: 10.1111/ene.13919

Elfil, M., Kamel, S., Kandil, M., Koo, B. B., and Schaefer, S. M. (2020). Implications of the gut microbiome in Parkinson's disease. *Mov. Disord.* 35, 921–933. doi: 10.1002/mds.28004

Eriksen, N., Stark, A. K., and Pakkenberg, B. (2009). "Age and Parkinson's disease-related neuronal death in the substantia nigra pars compacta," in *Birth, Life and Death of Dopaminergic Neurons in the Substantia Nigra*, eds G. Giovanni, V. Di Matteo, and E. Esposito (Vienna: Springer), 203–213. doi: 10.1007/978-3-211-92660-4_16

Espay, A. J., Brundin, P., and Lang, A. E. (2017). Precision medicine for disease modification in parkinson disease. *Nat. Rev. Neurol.* 13, 119–126. doi: 10.1038/nrneurol.2016.196

Fasano, A., Daniele, A., and Albanese, A. (2012). Treatment of motor and non-motor features of Parkinson's disease with deep brain stimulation. *Lancet Neurol.* 11, 429–442. doi: 10.1016/S1474-4422(12)70049-2

Foulds, P. G., Mitchell, J. D., Parker, A., Turner, R., Green, G., Diggle, P., et al. (2011). Phosphorylated α-synuclein can be detected in blood plasma and is potentially a useful biomarker for Parkinson's disease. *FASEB J.* 25, 4127–4137. doi: 10.1096/fj.10-179192

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (New York, NY), 1322–1333. doi: 10.1145/2810103.2813677

Gal, Y., Koumoutsakos, P., Lanusse, F., Louppe, G., and Papadimitriou, C. (2022). Bayesian uncertainty quantification for machine-learned models in physics. *Nat. Rev. Phys.* 4, 573–577. doi: 10.1038/s42254-022-00498-4

Garcia Santa Cruz, B., Bossa, M. N., Sölter, J., and Husch, A. D. (2021). Public covid-19 x-ray datasets and their impact on model bias-a systematic review of a significant problem. *Med. Image Anal.* 74, 102225. doi: 10.1016/j.media.2021.102225

Garcia Santa Cruz, B., Slter, J., Gomez-Giro, G., Saraiva, C., Sabate-Soler, S., Modamio, J., et al. (2022a). Generalising from conventional pipelines using deep learning in high-throughput screening workflows. *Sci. Rep.* 12, 11465. doi: 10.1038/s41598-022-15623-7

Garcia Santa Cruz, B., Vega, C., and Hertel, F. (2022b). "The need of standardised metadata to encode causal relationships: towards safer data-driven machine learning biological solutions," in *Computational Intelligence Methods for Bioinformatics and Biostatistics: 17th International Meeting, CIBB 2021, Virtual Event, November 15-17, 2021*. Revised Selected Papers (Cham: Springer), 200–216. doi: 10.1007/978-3-031-20837-9_16

Ge, W., Lueck, C., Suominen, H., and Apthorp, D. (2023). Has machine learning over-promised in healthcare? A critical analysis and a proposal for improved evaluation, with evidence from Parkinson's disease. *Artif. Intell. Med.* 139, 102524. doi: 10.1016/j.artmed.2023.102524

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* 3, e745–e750. doi: 10.1016/S2589-7500(21)00208-9

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., et al. (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* 9, 8. doi: 10.3389/fninf.2015.00008

Górriz, J. M., Ramirez, J., Suckling, J., Consortium, M. A., et al. (2019). On the computation of distribution-free performance bounds: application to small sample sizes in neuroimaging. *Pattern Recognit.* 93, 1–13. doi: 10.1016/j.patcog.2019.03.032

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Gupta, S., Singh, P., Chang, K., Qu, L., Aggarwal, M., Arun, N., et al. (2021). Addressing catastrophic forgetting for medical domain expansion. *arXiv*. [preprint]. doi: 10.48550/arXiv.2103.13511

Hassan, A., Wu, S. S., Schmidt, P., Simuni, T., Giladi, N., Miyasaki, J. M., et al. (2015). The profile of long-term Parkinson's disease survivors with 20 years of disease duration and beyond. *J. Parkinsons Dis.* 5, 313–319. doi: 10.3233/JPD-140515

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Volume 2*. Cham: Springer. doi: 10.1007/978-0-387-84858-7

He, R., Yan, X., Guo, J., Xu, Q., Tang, B., Sun, Q., et al. (2018). Recent advances in biomarkers for Parkinson's disease. *Front. Aging Neurosci.* 10, 305. doi: 10.3389/fnagi.2018.00305

Hess, C. W., and Okun, M. S. (2016). Diagnosing parkinson disease. *Contin. Lifelong Learn. Neurol.* 22, 1047–1063. doi: 10.1212/CON.0000000000000345

Hill-Burns, E. M., Debelius, J. W., Morton, J. T., Wissemann, W. T., Lewis, M. R., Wallen, Z. D., et al. (2017). Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* 32, 739–749. doi: 10.1002/mds.26942

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., Zhang, X., et al. (2022). Membership inference attacks on machine learning: a survey. *ACM Comput. Surv.* 54(11s), 1–37. doi: 10.1145/3523273

Huang, Y.-P., Chen, L.-S., Yen, M.-F., Fann, C.-Y., Chiu, Y.-H., Chen, H.-H., et al. (2013). Parkinson's disease is related to an increased risk of ischemic stroke–a population-based propensity score-matched follow-up study. *PLoS ONE* 8, e68314. doi: 10.1371/journal.pone.0068314

Hüllermeier, E., and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* 110, 457–506. doi: 10.1007/s10994-021-05946-3

Hustad, E., and Aasly, J. O. (2020). Clinical and imaging markers of prodromal Parkinson's disease. *Front. Neurol.* 11, 395. doi: 10.3389/fneur.2020.00395

Jakubovitz, D., Giryes, R., and Rodrigues, M. R. (2019). "Generalization error in deep learning," in *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, eds H. Boche, G. Caire, R. Calderbank, G. Kutyniok, R. Mathar, and P. Petersen (Cham: Springer), 153–193. doi: 10.1007/978-3-319-73074-5_5

Jankovic, J., McDermott, M., Carter, J., Gauthier, S., Goetz, C., Golbe, L., et al. (1990). Variable expression of Parkinson's disease: a base-line analysis of the dat atop cohort. *Neurology* 40, 1529–1529. doi: 10.1212/WNL.40.10.1529

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv*. [preprint]. doi: 10.48550/arXiv.1912.02178

Jimenez-Mesa, C., Ramirez, J., Suckling, J. Vöglein, J., Levin, J., Gorriz, J. M., Initiative, A. D. N., et al. (2023). A non-parametric statistical inference framework for deep learning in current neuroimaging. *Inf. Fusion* 91, 598–611. doi: 10.1016/j.inffus.2022.11.007

Kalia, L. V., and Lang, A. E. (2015). Parkinson's disease. *Lancet* 386, 896–912. doi: 10.1016/S0140-6736(14)61393-3

Kandasamy, K., Neiswanger, W., Schneider, J., Poczos, B., and Xing, E. P. (2018). Neural architecture search with bayesian optimisation and optimal transport. *Adv. Neural Inf. Process. Syst*. 31, 2016–2026. doi: 10.5555/3326943.3327130

Karthik, S., Revaud, J., and Chidlovskii, B. (2021). Learning from long-tailed data with noisy labels. *arXiv*. [preprint]. doi: 10.48550/arXiv.2108.11096

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J., et al. (2020). "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY), 1–14. doi: 10.1145/3313831.3376219

Kondrateva, E., Pominova, M., Popova, E., Sharaev, M., Bernstein, A., Burnaev, E., et al. (2021). "Domain shift in computer vision models for MRI data analysis: an overview," in *Thirteenth International Conference on Machine Vision*, Volume 11605 (Bellingham, WA: SPIE), 126–133. doi: 10.1117/12.2587872

Kubota, K. J., Chen, J. A., and Little, M. A. (2016). Machine learning for large-scale wearable sensor data in Parkinson's disease: concepts, promises, pitfalls, and futures. *Mov. Disord*. 31, 1314–1326. doi: 10.1002/mds.26693

Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: a taxonomy. *arXiv*. [preprint]. doi: 10.48550/arXiv.1710.10686

Laird, A. R., Eickhoff, S. B., Fox, P. M., Uecker, A. M., Ray, K. L., Saenz, J. J., et al. (2011). The brainmap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res. Notes* 4, 1–9. doi: 10.1186/1756-0500-4-349

Langley, J., He, N., Huddleston, D. E., Chen, S., Yan, F., Crosson, B., et al. (2019). Reproducible detection of nigral iron deposition in 2 Parkinson's disease cohorts. *Mov. Disord*. 34, 416–419. doi: 10.1002/mds.27608

Lawton, M., Baig, F., Rolinski, M., Ruffman, C., Nithi, K., May, M. T., et al. (2015). Parkinson's disease subtypes in the oxford parkinson disease centre (OPDC) discovery cohort. *J. Parkinsons Dis*. 5, 269–279. doi: 10.3233/JPD-140523

Lee, D. J., Lozano, C. S., Dallapiazza, R. F., and Lozano, A. M. (2019). Current and future directions of deep brain stimulation for neurological and psychiatric disorders: JNSPG 75th anniversary invited review article. *J. Neurosurg*. 131, 333–342. doi: 10.3171/2019.4.JNS181761

Lemay, A., Hoebel, K., Bridge, C. P., Befano, B., De Sanjosé, S., Egemen, D., et al. (2022). Improving the repeatability of deep learning models with Monte Carlo dropout. *Npj Digit. Med*. 5, 174. doi: 10.1038/s41746-022-00709-3

Liu, X. CONSORT-AI, T., and Group, S.-A. S. (2019). Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med*. 25, 1467–1468. doi: 10.1038/s41591-019-0603-3

Lozano, A. M. (2017). Waving hello to noninvasive deep-brain stimulation. *N. Engl. J. Med*. 377, 1096–1098. doi: 10.1056/NEJMcibr1707165

Lu, M., Zhao, Q., Poston, K. L., Sullivan, E. V., Pfefferbaum, A., Shahid, M., et al. (2021). Quantifying Parkinson's disease motor severity under uncertainty using mds-updrs videos. *Med. Image Anal*. 73, 102179. doi: 10.1016/j.media.2021.102179

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv*. [preprint]. doi: 10.48550/arXiv.1706.06083

Magesh, P. R., Myloth, R. D., and Tom, R. J. (2020). An explainable machine learning model for early detection of Parkinson's disease using lime on datscan imagery. *Comput. Biol. Med*. 126, 104041. doi: 10.1016/j.compbiomed.2020.104041

Mahlknecht, P., Seppi, K., Stockner, H., Nocker, M., Scherfler, C., Kiechl, S., et al. (2013). Substantia nigra hyperechogenicity as a marker for Parkinson's disease: a population-based study. *Neurodegener. Dis*. 12, 212–218. doi: 10.1159/000348595

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Oper. Res*. 43, 570–577. doi: 10.1287/opre.43.4.570

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The parkinson progression marker initiative (PPMI). *Prog. Neurobiol*. 95, 629–635. doi: 10.1016/j.pneurobio.2011.09.005

Marras, C., Beck, J., Bower, J., Roberts, E., Ritz, B., Ross, G., et al. (2018). Prevalence of Parkinson's disease across north america. *NPJ Parkinsons Dis*. 4, 1–7. doi: 10.1038/s41531-018-0058-0

Martínez-Murcia, F. J., Górriz, J. M., Ramírez, J., Illán, I., Ortiz, A., Initiative, P. P. M., et al. (2014). Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging. *Neurocomputing* 126, 58–70. doi: 10.1016/j.neucom.2013.01.054

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J.-M., Ramirez, J., and Castillo-Barnes, D. (2019). Studying the manifold structure of alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform*. 24, 17–26. doi: 10.1109/JBHI.2019.2914970

Martins, R., Oliveira, F., Moreira, F., Moreira, A. P., Abrunhosa, A., Januário, C., et al. (2021). Automatic classification of idiopathic Parkinson's disease and atypical parkinsonian syndromes combining [11C] raclopride pet uptake and MRI grey matter morphometry. *J. Neural. Eng*. 18, 046037. doi: 10.1088/1741-2552/abf772

Mata, I. F., Shi, M., Agarwal, P., Chung, K. A., Edwards, K. L., Factor, S. A., et al. (2010). SNCA variant associated with parkinson disease and plasma α-synuclein level. *Arch. Neurol*. 67, 1350–1356. doi: 10.1001/archneurol.2010.279

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv*. 54, 1–35. doi: 10.1145/3457607

Mei, J., Desrosiers, C., and Frasnelli, J. (2021). Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Front. Aging Neurosci*. 13, 633752. doi: 10.3389/fnagi.2021.633752

Miceli, M., Schuessler, M., and Yang, T. (2020). Between subjectivity and imposition: power dynamics in data annotation for computer vision. *Proc. ACM Hum.-Comput. Interact*. 4(CSCW2), 1–25. doi: 10.1145/3415186

Michell, A., Lewis, S., Foltynie, T., and Barker, R. (2004). Biomarkers and Parkinson's disease. *Brain* 127, 1693–1705. doi: 10.1093/brain/awh198

Mohammadi, D. (2013). The harvard biomarker study's big plan. *Lancet Neurol*. 12, 739–740. doi: 10.1016/S1474-4422(13)70155-8

Molnar, C. (2020). *Interpretable machine learning*. Available online at: https://christophm.github.io/interpretableml-book/ (accessed July 1, 2023).

Morrish, P., Sawle, G., and Brooks, D. (1996). An [18F] dopa-pet and clinical study of the rate of progression in Parkinson's disease. *Brain* 119, 585–591. doi: 10.1093/brain/119.2.585

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease (2003). The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Mov. Disord*. 18, 738–750. doi: 10.1002/mds.10473

Muangpaisan, W., Mathews, A., Hori, H., and Seidel, D. (2011). A systematic review of the worldwide prevalence and incidence of Parkinson's disease. *J. Med. Assoc. Thailand* 94, 749.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Adv. Neural Inf. Process. Syst*. 32, 4671–4681. Available online at: https://dl.acm.org/doi/10.5555/3454287.3454709

Nair, S. R., Tan, L. K., Mohd Ramli, N., Lim, S. Y., Rahmat, K., Mohd Nor, H., et al. (2013). A decision tree for differentiating multiple system atrophy from Parkinson's disease using 3-T MR imaging. *Eur. Radiol*. 23, 1459–1466. doi: 10.1007/s00330-012-2759-9

Neelakandan, S., Beulah, J. R., Prathiba, L., Murthy, G., Irudaya Raj, E. F., Arulkumar, N., et al. (2022). Blockchain with deep learning-enabled secure healthcare data transmission and diagnostic model. *Int. J. Model. Simul. Sci. Comput*. 13, 2241006. doi: 10.1142/S1793962322410069

Neri, E., de Souza, N., Brady, A., Bayarri, A. A., Becker, C. D., Coppola, F., et al. (2019). What the radiologist should know about artificial intelligence-an ESR white paper. *Insights Imaging* 10, 44. doi: 10.1186/s13244-019-0738-2

Nerius, M., Fink, A., and Doblhammer, G. (2017). Parkinson's disease in germany: prevalence and incidence based on health claims data. *Acta Neurol. Scand*. 136, 386–392. doi: 10.1111/ane.12694

Ngiam, K. Y., and Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 20, e262–e273. doi: 10.1016/S1470-2045(19)30149-4

Nicastro, N., Garibotto, V., and Burkhard, P. R. (2020). Extrastriatal 123 I-FP-CIT spect impairment in Parkinson's disease-the PPMI cohort. *BMC Neurol*. 20, 1–9. doi: 10.1186/s12883-020-01777-2

Niotis, K., West, A. B., and Saunders-Pullman, R. (2022). Who to enroll in parkinson disease prevention trials?: the case for genetically at-risk cohorts. *Neurology* 99(7 Supplement 1), 10–18. doi: 10.1212/WNL.0000000000200812

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). "Orthogonal random forest for causal inference," in *International Conference on Machine Learning* (New York, NY: PMLR), 4932–4941.

Pagano, G., Niccolini, F., and Politis, M. (2016). Imaging in Parkinson's disease. *Clin. Med*. 16, 371. doi: 10.7861/clinmedicine.16-4-371

Pal, G., Mangone, G., Hill, E. J., Ouyang, B., Liu, Y., Lythe, V., et al. (2022). Parkinson disease and subthalamic nucleus deep brain stimulation: cognitive effects in GBA mutation carriers. *Ann. Neurol*. 91, 424–435. doi: 10.1002/ana.26302

Parisi, F., Ferrari, G., Giuberti, M., Contin, L., Cimolin, V., Azzaro, C., et al. (2015). Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and gait tasks in Parkinson's disease. *IEEE J. Biomed. Health Inf.* 19, 1777–1793. doi: 10.1109/JBHI.2015.2472640

Parkinson, J. (2002). An essay on the shaking palsy. *J. Neuropsychiatry Clin. Neurosci*. 14, 223–236. doi: 10.1176/jnp.14.2.223

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, 7994–8006. Available online at: https://dl.acm.org/doi/10.5555/3454287.3455008

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., et al. (2021). Carbon emissions and large neural network training. *arXiv*. [preprint]. doi: 10.48550/arXiv.2104.10350

Pechevis, M., Clarke, C., Vieregge, P., Khoshnood, B., Deschaseaux-Voinet, C., Berdeaux, G., et al. (2005). Effects of dyskinesias in Parkinson's disease on quality of life and health-related costs: a prospective european study. *Eur. J. Neurol.* 12, 956–963. doi: 10.1111/j.1468-1331.2005.01096.x

Pellicano, C., Benincasa, D., Pisani, V., Buttarelli, F. R., Giovannelli, M., Pontieri, F. E., et al. (2007). Prodromal non-motor symptoms of Parkinson's disease. *Neuropsychiatr. Dis. Treat.* 3, 145. doi: 10.2147/nedt.2007.3.1.145

Pickrell, A. M., and Youle, R. J. (2015). The roles of pink1, parkin, and mitochondrial fidelity in Parkinson's disease. *Neuron* 85, 257–273. doi: 10.1016/j.neuron.2014.12.007

Poewe, W., and Wenning, G. (2002). The differential diagnosis of Parkinson's disease. *Eur. J. Neurol.* 9, 23–30. doi: 10.1046/j.1468-1331.9.s3.3.x

Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fmri data: the openfmri project. *Front. Neuroinform.* 7, 12. doi: 10.3389/fninf.2013.00012

Politis, M. (2014). Neuroimaging in parkinson disease: from research setting to clinical practice. *Nat. Rev. Neurol.* 10, 708–722. doi: 10.1038/nrneurol.2014.205

Politis, M., Piccini, P., Pavese, N., Koh, S.-B., and Brooks, D. J. (2008). Evidence of dopamine dysfunction in the hypothalamus of patients with Parkinson's disease: an in vivo 11c-raclopride pet study. *Exp. Neurol.* 214, 112–116. doi: 10.1016/j.expneurol.2008.07.021

Postuma, R. B., Poewe, W., Litvan, I., Lewis, S., Lang, A. E., Halliday, G., et al. (2018). Validation of the mds clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* 33, 1601–1608. doi: 10.1002/mds.27362

Power, J. D., Plitt, M., Laumann, T. O., and Martin, A. (2017). Sources and implications of whole-brain fmri signals in humans. *Neuroimage* 146, 609–625. doi: 10.1016/j.neuroimage.2016.09.038

Prechelt, L. (1998). "Early stopping-but when?" in *Neural Networks: Tricks of the Trade*, eds G. B. Orr, and K. R. Müller (Berlin: Springer), 55–69. doi: 10.1007/3-540-49430-8_3

Prell, T. (2018). Structural and functional brain patterns of non-motor syndromes in Parkinson's disease. *Front. Neurol.* 9, 138. doi: 10.3389/fneur.2018.00138

Pyatigorskaya, N., Sanz-Morère, C. B., Gaurav, R., Biondetti, E., Valabregue, R., Santin, M., et al. (2020). Iron imaging as a diagnostic tool for Parkinson's disease: a systematic review and meta-analysis. *Front. Neurol.* 11, 366. doi: 10.3389/fneur.2020.00366

Qin, R., Zhang, H., Jiang, L., Qiao, K., Hai, J., Chen, J., et al. (2020). Multicenter computer-aided diagnosis for lymph nodes using unsupervised domain-adaptation networks based on cross-domain confounding representations. *Comput. Math. Methods Med.* 2020, 3709873. doi: 10.1155/2020/3709873

Rajput, A. (1992). Frequency and cause of Parkinson's disease. *Can. J. Neurol. Sci.* 19, 103–107. doi: 10.1017/S0317167100041457

Reddy, S., Allan, S., Coghlan, S., and Cooper, P. (2020). A governance model for the application of ai in health care. *J. Am. Med. Inform. Assoc.* 27, 491–497. doi: 10.1093/jamia/ocz192

Riboldi, G. M., Frattini, E., Monfrini, E., Frucht, S. J., and Di Fonzo, A. (2022). A practical approach to early-onset parkinsonism. *J. Parkinsons Dis.* 12, 1–26. doi: 10.3233/JPD-212815

Ricci Lara, M. A., Echeveste, R., and Ferrante, E. (2022). Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* 13, 4581. doi: 10.1038/s41467-022-32186-3

Rietdijk, C. D., Perez-Pardo, P., Garssen, J., Van Wezel, R. J., and Kraneveld, A. D. (2017). Exploring Braak's hypothesis of Parkinson's disease. *Front. Neurol.* 8, 37. doi: 10.3389/fneur.2017.00037

Saeed, U., Compagnone, J., Aviv, R. I., Strafella, A. P., Black, S. E., Lang, A. E., et al. (2017). Imaging biomarkers in Parkinson's disease and parkinsonian syndromes: current and emerging concepts. *Transl. Neurodegener.* 6, 1–25. doi: 10.1186/s40035-017-0076-6

Sakai, K., and Yamada, K. (2019). Machine learning studies on major brain diseases: 5-year trends of 2014-2018. *Jpn. J. Radiol.* 37, 34–72. doi: 10.1007/s11604-018-0794-4

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., Aroyo, L. M., et al. (2021). "Everyone wants to do the model work, not the data work," *in Data Cascades in High-stakes AI* (New York, NY), 1–15. doi: 10.1145/3411764.3445518

Santiago, J. A., Bottero, V., and Potashkin, J. A. (2017). Biological and clinical implications of comorbidities in Parkinson's disease. *Front. Aging Neurosci.* 9, 394. doi: 10.3389/fnagi.2017.00394

Schootemeijer, S., van der Kolk, N. M., Bloem, B. R., and de Vries, N. M. (2020). Current perspectives on aerobic exercise in people with Parkinson's disease. *Neurotherapeutics* 17, 1418–1433. doi: 10.1007/s13311-020-00904-8

Schwarz, S. T., Afzal, M., Morgan, P. S., Bajaj, N., Gowland, P. A., Auer, D. P., et al. (2014). The 'swallow tail' appearance of the healthy nigrosome-a new accurate test of Parkinson's disease: a case-control and retrospective cross-sectional MRI study at 3T. *PLoS ONE* 9, e93814. doi: 10.1371/journal.pone.0093814

Settles, B. (2009). *Active Learning Literature Survey*. Madison, WI: University of Wisconsin–Madison.

Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P. K., et al. (2019). Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. *Neuroimage Clin.* 22, 101748. doi: 10.1016/j.nicl.2019.101748

Siderowf, A., Concha-Marambio, L., Lafontant, D.-E., Farris, C. M., Ma, Y., Urenia, P. A., et al. (2023). Assessment of heterogeneity among participants in the Parkinson's progression markers initiative cohort using α-synuclein seed amplification: a cross-sectional study. *Lancet Neurol.* 22, 407–417. doi: 10.1016/S1474-4422(23)00109-6

Siderowf, A., Jennings, D., Eberly, S., Oakes, D., Hawkins, K. A., Ascherio, A., et al. (2012). Impaired olfaction and other prodromal features in the parkinson at-risk syndrome study. *Mov. Disord.* 27, 406–412. doi: 10.1002/mds.24892

Smith, J. J., Sorensen, A. G., and Thrall, J. H. (2003). Biomarkers in imaging: realizing radiology's future. *Radiology* 227, 633–638. doi: 10.1148/radiol.2273020518

Smith, M. G., Witte, M., Rocha, S., and Basner, M. (2019). Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. *BMC Med. Res. Methodol.* 19, 1–13. doi: 10.1186/s12874-019-0868-8

Song, L., Shokri, R., and Mittal, P. (2019). "Privacy risks of securing machine learning models against adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY), 241–257. doi: 10.1145/3319535.3354211

Stoker, T. B., and Barker, R. A. (2020). Recent developments in the treatment of Parkinson's disease. *F1000Res.* 9, 11. doi: 10.12688/f1000research.25634.1

Stolze, H., Kuhtz-Buschbeck, J. P., Drücke, H., Jöhnk, K., Illert, M., and Deuschl, G. (2001). Comparative analysis of the gait disorder of normal pressure hydrocephalus and Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 70, 289–297. doi: 10.1136/jnnp.70.3.289

Strother, S. C. (2006). Evaluating fmri preprocessing pipelines. *IEEE Eng. Med. Biol Mag.* 25, 27–41. doi: 10.1109/MEMB.2006.1607667

Sulzer, D., Cassidy, C., Horga, G., Kang, U. J., Fahn, S., Casella, L., et al. (2018). Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease. *NPJ Parkinsons Dis.* 4, 11. doi: 10.1038/s41531-018-0047-3

Sveinbjornsdottir, S. (2016). The clinical symptoms of Parkinson's disease. *J. Neurochem.* 139, 318–324. doi: 10.1111/jnc.13691

Tahmasian, M., Bettray, L. M., van Eimeren, T., Drzezga, A., Timmermann, L., Eickhoff, C. R., et al. (2015). A systematic review on the applications of resting-state fmri in Parkinson's disease: does dopamine replacement therapy play a role? *Cortex* 73, 80–105. doi: 10.1016/j.cortex.2015.08.005

Talai, A. S., Sedlacik, J., Boelmans, K., and Forkert, N. D. (2021). Utility of multi-modal MRI for differentiating of Parkinson's disease and progressive supranuclear palsy using machine learning. *Front. Neurol.* 12, 648548. doi: 10.3389/fneur.2021.648548

Tamburri, D. A. (2020). "Sustainable mlops: trends and challenges," in *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (Timisoara: IEEE), 17–23. doi: 10.1109/SYNASC51798.2020.00015

Tan, A. H., Lim, S.-Y., Chong, K. K., Manap, M. A. A. A., Hor, J. W., Lim, J. L., et al. (2021). Probiotics for constipation in parkinson disease: a randomized placebo-controlled study. *Neurology* 96, e772–e782. doi: 10.1212/WNL.0000000000010998

Tedeschini, B. C., Savazzi, S., Stoklasa, R., Barbieri, L., Stathopoulos, I., Nicoli, M., et al. (2022). Decentralized federated learning for healthcare networks: a case study on tumor segmentation. *IEEE Access* 10, 8693–8708. doi: 10.1109/ACCESS.2022.3141913

Thenganatt, M. A., and Jankovic, J. (2014). Parkinson disease subtypes. *JAMA Neurol.* 71, 499–504. doi: 10.1001/jamaneurol.2013.6233

Thevathasan, W., Debu, B., Aziz, T., Bloem, B. R., Blahak, C., Butson, C., et al. (2018). Pedunculopontine nucleus deep brain stimulation in Parkinson's disease: a clinical review. *Mov. Disord.* 33, 10–20. doi: 10.1002/mds.27098

Tolosa, E., Garrido, A., Scholz, S. W., and Poewe, W. (2021). Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol.* 20, 385–397. doi: 10.1016/S1474-4422(21)00030-2

Tolosa, E., Vila, M., Klein, C., and Rascol, O. (2020). Lrrk2 in parkinson disease: challenges of clinical trials. *Nat. Rev. Neurol.* 16, 97–107. doi: 10.1038/s41582-019-0301-2

Toulas, B. (2023). *Hospital Clínic de Barcelona Severely Impacted by Ransomware Attack*. Available online at: https://www.bleepingcomputer.com/news/security/hospital-cl-nic-de-barcelona-severely-impacted-by-ransomware-attack/ (accessed July 1, 2023).

van Veluw, S. J., Zwanenburg, J. J., Hendrikse, J., van der Kolk, A. G., Luijten, P. R., Biessels, G. J., et al. (2014). "High resolution imaging of cerebral small vessel disease with 7 T MRI," in *Trends Neurovascular Interventions*, eds T. Tsukahara, G. Esposito, H. J. Steiger., G. Rinkel, and L. Regli (Cham: Springer), 125–130. doi: 10.1007/978-3-319-02411-0_21

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5999–6010. Available online at: https://dl.acm.org/doi/10.5555/3295222.3295349

Vega, C. (2021). From hume to Wuhan: an epistemological journey on the problem of induction in covid-19 machine learning models and its impact upon medical research. *IEEE Access* 9, 97243–97250. doi: 10.1109/ACCESS.2021.3095222

Virreira Winter, S., Karayel, O., Strauss, M. T., Padmanabhan, S., Surface, M., Merchant, K., et al. (2021). Urinary proteome profiling for stratifying patients with familial Parkinson's disease. *EMBO Mol. Med.* 13, e13257. doi: 10.15252/emmm.202013257

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., and Capuzzo, D. (2022). Statistical stability indices for lime: obtaining reliable explanations for machine learning models. *J. Oper. Res. Soc.* 73, 91–101. doi: 10.1080/01605682.2020.1865846

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. (2021). On calibration and out-of-domain generalization. *Adv. Neural Inf. Process. Syst.* 34, 2215–2227. doi: 10.48550/arXiv.2102.10395

Wang, H., Wu, Z., Liu, Z., Cai, H., Zhu, L., Gan, C., et al. (2020). HAT: hardware-aware transformers for efficient natural language processing. *arXiv*. [preprint]. doi: 10.48550/arXiv.2005.14187

Wang, H., Wu, Z., and Xing, E. P. (2018). "Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications," in *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium* (Singapore: World Scientific), 54–65. doi: 10.1142/9789813279827_0006

Weingärtner, S., Desmond, K. L., Obuchowski, N. A., Baessler, B., Zhang, Y., Biondetti, E., et al. (2022). Development, validation, qualification, and dissemination of quantitative MR methods: overview and recommendations by the ISMRM quantitative MR study group. *Magn. Reson. Med.* 87, 1184–1206. doi: 10.1002/mrm.29084

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* 63, 826–833. doi: 10.1016/j.jclinepi.2009.11.020

Widner, K., Virmani, S., Krause, J., Nayar, J., Tiwari, R., Pedersen, E. R., et al. (2023). Lessons learned from translating ai from development to deployment in healthcare. *Nat. Med.* 29, 1304–1306. doi: 10.1038/s41591-023-02293-9

Wiens, J., Price, W. N., and Sjoding, M. W. (2020). Diagnosing bias in data-driven algorithms for healthcare. *Nat. Med.* 26, 25–26. doi: 10.1038/s41591-019-0726-6

Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., et al. (2013). Standardization of analysis sets for reporting results from adni MRI data. *Alzheimers Dement.* 9, 332–337. doi: 10.1016/j.jalz.2012.06.004

Xu, X.-W., Doi, K., Kobayashi, T., MacMahon, H., and Giger, M. L. (1997). Development of an improved CAD scheme for automated detection of lung nodules in digital chest images. *Med. Phys.* 24, 1395–1403. doi: 10.1118/1.598028

Yagis, E., DE Herrera, A. G. S., and Citi, L. (2019). "Generalization performance of deep learning models in neurodegenerative disease classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA: IEEE), 1692–1698. doi: 10.1109/BIBM47256.2019.8983088

Yoshikawa, K., Nakata, Y., Yamada, K., and Nakagawa, M. (2004). Early pathological changes in the parkinsonian brain demonstrated by diffusion tensor MRI. *J. Neurol. Neurosurg. Psychiatry* 75, 481–484. doi: 10.1136/jnnp.2003.021873

Yosinski,J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* 27, 3320–3329. Available online at: https://dl.acm.org/doi/10.5555/2969033.2969197

Zeng, X., and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *J. Exp. Theor. Artif. Intell.* 12, 1–12. doi: 10.1080/095281300146272

Zetusky, W. J., Jankovic, J., and Pirozzolo, F. J. (1985). The heterogeneity of Parkinson's disease: clinical and prognostic implications. *Neurology* 35, 522–522. doi: 10.1212/WNL.35.4.522

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 107–115. doi: 10.1145/3446776

Zhang, S., Tao, K., Wang, J., Duan, Y., Wang, B., Liu, X., et al. (2020). Substantia nigra hyperechogenicity reflects the progression of dopaminergic neurodegeneration in 6-ohda rat model of Parkinson's disease. *Front. Cell. Neurosci.* 14, 216. doi: 10.3389/fncel.2020.00216

Zhang, X., Chou, J., Liang, J., Xiao, C., Zhao, Y., Sarva, H., et al. (2019). Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Sci. Rep.* 9, 797. doi: 10.1038/s41598-018-37545-z

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: reducing gender bias amplification using corpus-level constraints. *arXiv*. [preprint]. doi: 10.48550/arXiv.1707.09457