frontiers | Frontiers in Aging Neuroscience

TYPE Review
PUBLISHED 13 February 2023
DOI 10.3389/fnagi.2023.1076657

Check for updates

OPEN ACCESS

EDITED BY
Christophe Hirtz,
Université de Montpellier,
France

REVIEWED BY
Byron Creese,
University of Exeter,
United Kingdom
Steven Gunzler,
University Hospitals Cleveland Medical Center,
United States

*CORRESPONDENCE
Lee Lancashire
✉ Lee.Lancashire@cohenbio.org

SPECIALTY SECTION
This article was submitted to Parkinson's Disease and Aging-related Movement Disorders, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 21 October 2022
ACCEPTED 16 January 2023
PUBLISHED 13 February 2023

CITATION
Gerraty RT, Provost A, Li L, Wagner E, Haas M and Lancashire L (2023) Machine learning within the Parkinson's progression markers initiative: Review of the current state of affairs.
Front. Aging Neurosci. 15:1076657.
doi: 10.3389/fnagi.2023.1076657

COPYRIGHT
© 2023 Gerraty, Provost, Li, Wagner, Haas and Lancashire. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Machine learning within the Parkinson's progression markers initiative: Review of the current state of affairs

Raphael T. Gerraty[1], Allison Provost[1], Lin Li[2], Erin Wagner[2], Magali Haas[1] and Lee Lancashire[1]*

[1]Cohen Veterans Bioscience, New York, NY, United States, [2]PharmaLex, Frederick, MD, United States

The Parkinson's Progression Markers Initiative (PPMI) has collected more than a decade's worth of longitudinal and multi-modal data from patients, healthy controls, and at-risk individuals, including imaging, clinical, cognitive, and 'omics' biospecimens. Such a rich dataset presents unprecedented opportunities for biomarker discovery, patient subtyping, and prognostic prediction, but it also poses challenges that may require the development of novel methodological approaches to solve. In this review, we provide an overview of the application of machine learning methods to analyzing data from the PPMI cohort. We find that there is significant variability in the types of data, models, and validation procedures used across studies, and that much of what makes the PPMI data set unique (multi-modal and longitudinal observations) remains underutilized in most machine learning studies. We review each of these dimensions in detail and provide recommendations for future machine learning work using data from the PPMI cohort.

KEYWORDS

machine learning, Parkinson's Disease, multi-omic analyses, PD progression, data analysis methods

Introduction

As part of the drive toward precision medicine, there has been an increased focus on the discovery of biological markers and quantitative techniques to serve as diagnostic and prognostic tools for individual patients, and for monitoring the progression or remission of disease. High-dimensional imaging data, as well as genetics, protein, and other 'omics' assays capable of querying genotypes, transcriptomes and interactomes at low cost have opened the possibility of carrying out *de novo* discovery across a vast array of biological variables. These approaches have already begun to generate promising candidate biomarkers for a variety of neurological diseases (Nalls et al., 2014; Sanders et al., 2015; Chang et al., 2017; Wray et al., 2018; Nievergelt et al., 2019), shifting the current paradigm away from candidate studies that have proved to be unreliable and irreproducible for reasons ranging from poor study design to a lack of reporting (Landis et al., 2012; Baker, 2016; McShane, 2017; Scherer, 2017; Sun et al., 2019; Ren et al., 2020) toward profiling whole systems, which will provide a broader perspective for disease understanding.

The increased use of multi-modal and high-dimensional data is readily apparent in recent Parkinson's Disease (PD) research. PD is the second most common neurodegenerative disorder, reaching up to 4% prevalence by age 80 (Pringsheim et al., 2014). While the proximal mechanisms of PD are well understood to be damage to midbrain dopaminergic neurons, and the underlying genetic causes have been discovered in some cases, most PD diagnoses are idiopathic, and little is

known about variation in patient trajectories. One prominent example of research aimed at better understanding PD is the landmark Parkinson's Progression Markers Initiative (PPMI; Marek et al., 2011, 2018), which since its inception in 2010, has collected longitudinal data across several patient cohorts and data modalities, including

- clinical measures
- brain imaging
- gene expression levels, protein abundance, and genomic variant status
- sensors and wearable devices data

with the aim of identifying biomarkers to support the development of new interventions for PD. This initiative represents an extremely rich and well-annotated dataset for studying the progression of multiple biological variables alongside clinical measures of disease severity.

Broader investigation of whole 'omic' domains creates its own set of challenges, however. Classical statistical methods for parameter estimation, classification, clustering, and controlling for false positives may perform poorly or be computationally or analytically intractable in these settings, in which observations are often high dimensional and correlated, and meaningful associations are sparse. These problems are compounded by the increasing realization that to understand complex phenotypes, multiple high-dimensional data modalities will need to be integrated.

With increasing volumes of data being collected, machine learning (ML) techniques offer up a potential solution to the above challenges and are beginning to have an impact on research and healthcare. Here, elements of clinical diagnosis and prognosis are being automated with increasing levels of complexity and accuracy. These techniques may lead to the discovery of novel biomarkers, shedding light on the determinants of disease severity for complex neurological diseases such as PD that affect multiple biological systems and whose etiology is not fully understood. While we focus on machine learning in this review, we note that this approach is not always preferrable to classical statistical methods. In cases with a small number of variables and reasonable, well-defined null hypotheses, a significance testing framework may make sense. Furthermore, as we note below, the line between classical statistics and machine learning is often a blurry one. We also note that machine learning algorithms are not a panacea, and that complex problems will generally require good theory and good data to solve. Some measurements may not completely capture the underlying construct they are designed to. To accurately and meaningfully predict psychosis symptoms, for example, a scale which accurately tracks those symptoms is an obvious requirement.

The extensive collection and analysis of longitudinal data has highlighted the heterogenous nature of PD in general, and of disease progression in particular, where people with PD exhibit varying courses of the disease over time (i.e., individuals with PD can be differentiated along numerous axes, symptom patterns etc.; Marquand et al., 2016; Feczko et al., 2019). Akin to the "curse of dimensionality" in machine learning that describes the exponential need for more data as the dimensionality of the feature space grows (Bellman, 1956), this "curse of heterogeneity" presents multiple challenges to the field with respect to the increased sample size required to power discovery efforts that define homogeneous subgroups within a heterogeneous population who share a common clinical diagnosis. Developing tools to disentangle this heterogeneity, and to therefore subtype patients in clinically meaningful ways, will be useful for trial design and clinical care, an area with high rates of failure (Arrowsmith and Miller, 2013; Harrison, 2016) partly due to a lack of insight into the underlying pathology of these disorders of the brain (Krystal and State, 2014).

In this review, we highlight the application of machine learning to the study of PD, focusing our attention on studies using the PPMI data set. In particular, we will consider an approach to be an example of machine learning if it has an overt focus on developing or testing algorithms for *prediction* of diagnosis, symptoms, or progression in unseen data; or for the automated *compression* of high dimensional patient data into lower dimensional factors or clusters. While there are analogs of each of these areas in classical statistical approaches, they represent the main focus of most machine learning research, in contrast to the common statistical goals of null hypothesis testing and parameter estimation. Our aims in this review are to provide a qualitative summary of the research that has been done using machine learning within the PPMI cohort, as well as a reference for researchers interested in continuing the exploration of the cohort to predict diagnosis, symptoms, disease subtypes, risk factors, and patient trajectories in PD.

Machine learning techniques are usually divided into supervised and unsupervised categories. Supervised models are focused on the problem of prediction: they learn a function mapping between some input variables and a target output of interest, such as diagnosis or scores on a symptoms scale. Unsupervised models are focused on compression: they learn to reduce input data so that it has fewer dimensions, based on similarities or redundancies within the input data itself. Unsupervised learning can be further divided into *clustering* or *latent variable* algorithms, depending on whether the data are reduced to categories containing similar sets of observations (often called "subtypes") or to a set of continuous scores which reflect combinations of the original measurements. Other categories of machine learning, such as semi-supervised learning or reinforcement learning, are currently rare in Parkinson's research and thus outside the scope of this review.

PPMI machine learning studies

Based on a review of publications collected on the PPMI website[1] as of July 5th 2022, combined with a PubMed search for keywords "PPMI," "Parkinson's Disease," "Machine Learning" on the same day a total of 277 unique publications available in English were screened. Figure 1 shows a flow chart describing the selection and filtering of papers, in accordance with PRISMA 2020 guidelines (Page et al., 2021). Briefly, we excluded 1 commentary, 1 observational case study, 1 dissertation, 4 non-Parkinson's papers, and 6 reviews from further analysis. Of the 263 remaining papers we were able to retrieve, 149 did not meet our criteria for machine learning, because they did not focus on either prediction (supervised learning) or compression (unsupervised learning), leaving a total of 114 machine learning PPMI papers. This large reduction is due in part to the common use of the term prediction in studies where authors perform classical null hypothesis testing, asking essentially whether a particular association or difference is stronger than one would expect due to chance, given a set of assumptions about how the data were generated. This diverges from the intuitive definition of prediction, which involves at a minimum generating (and ideally validating) estimates of some aspect of yet unseen data, which is the criteria

1 https://www.ppmi-info.org/publications-presentations/publications/

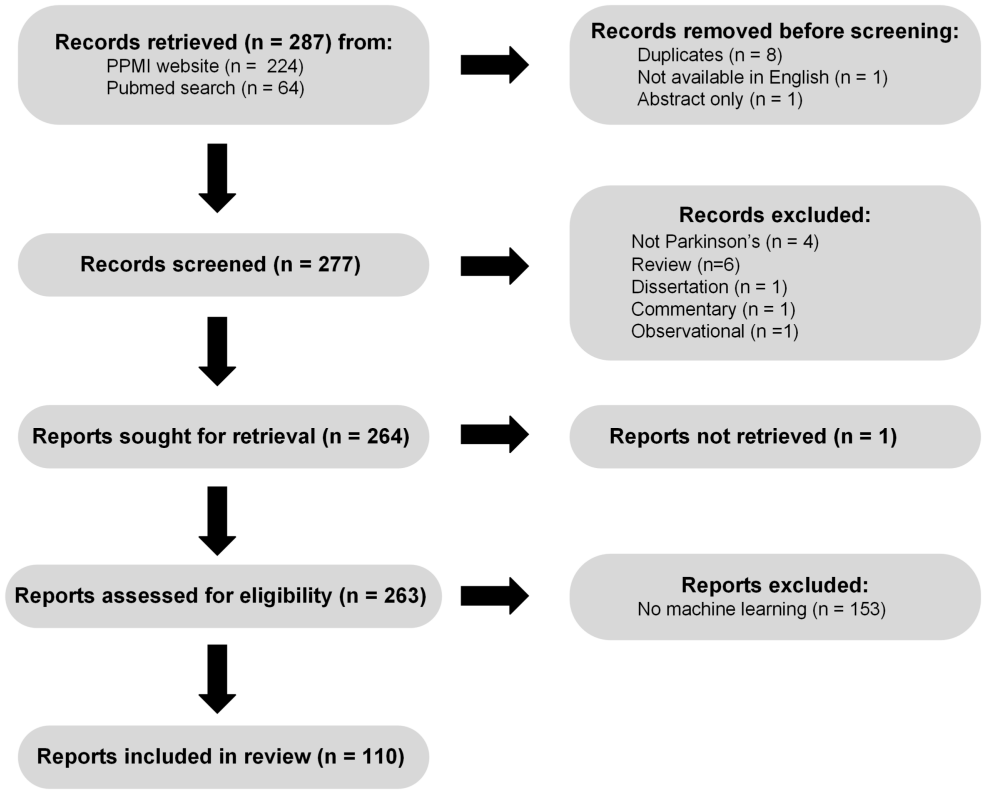

FIGURE 1
PRISMA (Page et al., 2021) flow chart illustrating the process of searching for, screening, and assessing machine learning PPMI papers.

we use here. A further 4 studies involved prediction without any model development or fitting, leaving a total of 110 papers which we consider machine learning using PPMI data. The machine learning studies we reviewed varied along several dimensions, three of which we focus on here: (1) the types of supervised or unsupervised models used, (2) the data modalities included as inputs and outputs of the models, and (3) the methods used for model validation. Broadly, most studies used supervised classification or regression methods for characterizing PD and relating different biomarkers and clinical measures to diagnosis, symptoms, and disease progression.

Studies explored a wide range of input data types, including clinical data, imaging, cerebrospinal fluid (CSF) and blood protein levels, as well as DNA and RNA data, to predict disease characteristics or identify latent dimensions or categories, but most used only clinical measures and a small subset of the available neuroimaging data. Studies also varied widely in whether and how they validated their models and performance estimates, an essential step toward making findings reproducible, generalizable, and ultimately relevant to the clinic. We will explore each of these dimensions below, summarizing previous research and making suggestions for improving future work.

Types of models

Of the 110 machine learning studies we reviewed, almost 90% (97 studies) reported results of supervised models, while only 19 studies used unsupervised methods. Of the papers reporting supervised methods, 55 attempted to predict Parkinson's diagnosis, representing a majority not only of supervised papers, but of all machine learning papers we reviewed. While early diagnosis of PD, especially in prodromal individuals exhibiting sub-threshold motor symptoms, olfactory deficits, or Rapid Eye Movement (REM) sleep disorder, is an important goal, there are numerous well-validated clinical tools for diagnosing PD at the stage in which patients are enrolled into the PPMI study. Machine learning algorithms focused on diagnosis will have little to say about the major goals of PPMI: to understand variability in patient symptoms, and especially variability in the trajectory of symptoms over time.

Significantly fewer machine learning papers made use of the longitudinal structure of PPMI data, with 26 reports predicting future symptoms from some baseline (which we call "progression prediction"). Because understanding the biological variables associated with heterogeneity in patient trajectories is a core goal of the PPMI project, we summarize these progression prediction papers in Table 1. These studies illustrate many of the characteristics of (and issues with) the broader group of papers we reviewed here, which are discussed in detail below.

Thirteen additional studies focused on predicting symptoms measured at the same time as the predictive features, while five studies focused on predicting either neuroimaging results or medication state rather than symptoms or diagnosis (Simuni et al., 2016; Freeze et al., 2018; Valmarska et al., 2018; Shu et al., 2020; Lim et al., 2021). One study reported predictive accuracy for both diagnosis and symptom levels (Soltaninejad et al., 2019). The most popular supervised models were linear regression (often with a L1-norm regularization penalty on coefficients to promote sparsity, called LASSO (Tibshirani, 1996)), support vector machines (SVM) (Cortes and Vapnik, 1995), and random forests (RFs; Breiman, 2001)

10.3389/fnagi.2023.1076657

TABLE 1 Machine learning papers predicting symptom progression using the Parkinson's Progression Markers Initiative data set.

Study	Input data					Outcome measures	Modeling approach	Validation method	Summary
	C	D	R	B	I				
1. Adams et al. (2021)	✓				✓	UPDRS III at year 4	CNN	Ten-fold CV	Trained a CNN to predict motor symptoms at year 4 from baseline raw DAT images and year 0 and 1 UPDRS III.
2. Chahine et al. (Chahine et al., 2021)	✓				✓	Diagnosis of α-synucleinopathy (aSN)	Cox hazard regression	None	Tests DAT SBR at baseline for predicting future diagnosis of PD, Lewy Body Dementia, or Multiple System Atrophy. Reports sensitivity and specificity but no validation.
3. Chen et al. (2021)	✓	✓		✓	✓	MCI diagnosis in patients with REM disorder	LASSO, Cox hazard regression	Validation set	Predicted time to MCI from baseline genetic, CSF, DAT, and clinical features.
4. Combs et al. (2021)	✓					Cognition (neuropsychological tests) at year 1	Stepwise regression	Validation set	Prediction of cognitive performance from baseline clinical and cognitive scores in both controls and patients.
5. D'Cruz et al. (2021)	✓				✓	FoG at year 2	Vertex-based shape analysis stepwise logistic regression	None (statistical tests of a subset of features in PPMI data set)	Note: Model was fit to internal cohort. Baseline clinical and MRI features used to predict FoG. Variables selected based on significance and stepwise regression. AUC reported but not validated. Statistical tests of MRI features performed on PPMI data.
6. Faghri et al. (2018)	✓					UPDRS progression sub-type	Nonnegative Matrix Factorization, Gaussian Mixture Models, Random forests	Five-fold CV Test set External validation set	Combined unsupervised and supervised methods to divide patients into progression sub-types and predict sub-type score from baseline clinical measures
7. Gramotnev et al. (2019)	✓	✓		✓	✓	Rate of MoCA change	Logistic regression with variable selection based on relative importance	Monte Carlo (after selecting significant features*)	Used baseline clinical, genetic, DAT, and CSF to predict rate of cognitive decline. Variables selected based on performance before validation*.
8. Gu et al. (2020)	✓			✓	✓	Geriatric Depression Scale at year 2	XGBoost, Stepwise logistic regression	Ten-fold CV for hyperparameter tuning Validation set	Compared methods for predicting depression severity at year 2 from clinical, CSF, and DAT measures.
9. Hayete et al. (2017)	✓	✓			✓	Rate of change in UPDRS II + III and MoCA	Dynamic Bayesian graphical model	None (Statistical tests of a subset of features in PPMI data set)	Note: Model was fit to LAB-PD data. PPMI was used for limited external validation. Predicted motor and cognitive progression mainly at years 5-7 of follow-up. Direct predictive validation was not performed, but a subset of findings was tested statistically in PPMI.
10. Jackson et al. (2021)	✓				✓	Change in UPDRS III at year 1	Ridge regression	External validation set (PPMI)	Predicted motor decline at 1 year from baseline clinical and DAT factors. Note: trained on placebo arm of clinical trial, tested on PPMI
11. Kim et al. (2019)	✓			✓	✓	Freezing of Gait (FoG from UPDRS) at year 4	Cox hazard regression	None	Predicted future freezing of gait, from baseline clinical measures, CSF, and DAT. Reports AUC but no validation.
12. Kim and Jeon (2021)	✓			✓		FoG (UPDRS) up to year 8	NA (ROC analysis of NFL)	None	Predicted gait freezing using serum NFL. Reports AUC but no validation.

(Continued)

TABLE 1 (Continued)

Study	Input data					Outcome measures	Modeling approach	Validation method	Summary
	C	D	R	B	I				
13. Latourelle et al. (2017)	✓	✓		✓	✓	Rates of motor and daily living symptoms (combined UPDRS II and III totals, rates estimated from linear mixed-effects models)	Reverse engineering and forward simulation (REFS) model ensemble	Five-fold CV External validation set (LABS-PD)	Large-scale prediction of symptom progression using model ensemble trained on >17,000 SNPs, clinical variables, and DAT and CSF features. <4% variance explained by biological variables. Tested results on external cohort
14. Ma et al. (2021)	✓			✓		UPDRS III at multiple years	Multiple ML models (LASSO, ridge, random forests, gradient boosting) with recursive feature elimination	CV within training set for variable selection validation set	Compared performance of multiple ML models in using each year's clinical and CSF measures to predict subsequent year's motor scores
15. Nguyen et al. (2020)	✓			✓	✓	UPDRS III total and MoCA scores at years 1, 2, 3, and 4	Deformation-based morphometry, DNN autoencoder	Five-fold CV (after selecting significant regions*)	Predicted motor and cognitive deficits from baseline MRI, CSF, and clinical scores in patients with REM disorder. Variables selected based on performance before validation*.
16. Rahmim et al. (2017)	✓				✓	UPDRS III at year 4	Random forests	LOO	Year 1 and 2 DAT, MRI, and clinical scores were used to predict UPDRS III total at year 4
17. Ren et al. (2021)	✓					Hoehn & Yahr score	Multivariate functional PCA, Cox hazard regression	External validation set (LABS-PD)	Combined unsupervised dimensionality reduction of clinical and cognitive variables with prediction of functional outcomes.
18. Rutten et al. (2017)	✓				✓	2-year change in anxiety (STAI)	Linear mixed-effects model with stepwise selection procedure	None	Stepwise selection of features for linear mixed-effects model predicting 2-year change in STAI scores from baseline clinical scores and DAT features. No validation of selected features was performed.
19. Salmanpour et al. (2019)	✓					MoCA at year 4	DNNs, LASSO, random forest, ridge, others	Monte Carlo Test set	Tested future cognitive score from a large set of feature selection and prediction algorithms. Genetic algorithm combined with local linear trees performed the best.
20. Schrag et al. (2017)	✓	✓		✓	✓	Cognition (MoCA) at 2 years	Logistic regression with variables pre-filtered based on significance	Monte Carlo Ten-fold CV validation set all after selecting significant features*	Predicted cognitive impairment with clinical scores, CSF, APOE status, and DAT. Selected variables before validation*.
21. Simuni et al. (2016)	✓			✓	✓	Time to initiation of symptomatic treatment	Random survival forest	CV	Predicted time to initiation of symptomatic therapy using random survival forests. No biological variables increased accuracy of prediction above clinical baseline.
22. Tang et al. (2019)	✓				✓	UPDRS III at year 4	DNN	LOO (after selecting significant features*)	Artificial neural network to predict future motor symptoms from imaging and clinical features. Variables selected based on performance before validation*.

(Continued)

TABLE 1 (Continued)

Study	Input data					Outcome measures	Modeling approach	Validation method	Summary
	C	D	R	B	I				
23. Tang et al. (2021)	✓			✓	✓	Cognitive decline (MoCA or neuropsychological test scores)	LASSO, Cox hazard Regression	Validation set (*t*-tests to remove variables that differ between training and validation sets*)	Prediction of cognitive decline using baseline clinical, CSF, and MRI features. Features were filtered after training based on similarity between training and test sets*
24. Tsiouris et al. (2017)	✓	✓		✓	✓	Rate of change in UPDRS total up to year 2	RIPPER (Cohen, 1995)	Ten-fold CV	Predicted change in UPDRS scores at 2-and 4-year epochs after selecting from over 600 baseline variables, including genetic, CSF, clinical and imaging features
25. Tsiouris et al. (2020)	✓	✓		✓	✓	Rate of change in UPDRS total up to year 4	Naïve Bayes, RIPPER	Ten-fold CV	Extended (Tsiouris et al., 2017) to 4 year follow up
26. Zeighami et al. (2019)	✓				✓	Change in Global Composite Outcome (Fereshtehnejad et al., 2015)	Voxel-based Morphometry, Independent Component Analysis	Ten-fold CV (after selecting significant voxels)	Tested whether baseline MRI-based atrophy marker could predict change in overall severity. Cross-cohort validation prevented data leakage.

Because of the focus of PPMI on variation in symptom trajectories, we summarize here the 26 machine learning papers from our literature search attempting to predict future symptoms. Input Data: C, clinical/demographic; D, DNA/genotype; R, RNA; B, biomarker/biospecimen; I, imaging; CV, cross-validation; LOO, leave-one-out; UPDRS, Unified Parkinson's Disease Rating Scale; DNN, deep neural network; CNN, convolutional neural network; MCI, Mild Cognitive Impairment; NFL, Neurofilament light chain; *potential data leakage. In papers containing both null hypothesis tests and measures of predictive accuracy, only variables and methods included in predictive tests are considered here.

and/or gradient boosting methods (Friedman et al., 2000; Friedman, 2001).

A smaller number of studies used unsupervised learning to generate latent variables or clusters to capture patient variability. Of the 19 studies using unsupervised methods, 11 were concerned with sub-typing Parkinson's patients using clustering models. Eleven used latent variable or dimensionality reduction methods with continuous latent factors, and 3 used both sub-typing and continuous latent variables.

Only 6 papers combined supervised and unsupervised methods. This is surprising given the stated focus of much PD research on finding subtypes which can predict differential progression across groups of Parkinson's patients. To discover latent sub-groups of patients and find predictors of future sub-group membership, it is likely that supervised and unsupervised models will need to be integrated. Notably, 3 papers combined clustering of patients into subtypes with prediction of current or future symptoms. In one example, Faghri et al. (2018) used an unsupervised combination of Nonnegative Matrix Factorization (NMF) and Gaussian Mixture Models (GMMs), respectively, to reduce dimensionality and cluster patients into sub-types, and RFs for supervised prediction of symptom levels 4 years later. Valmarska et al. (2018) explored the impact of baseline motor symptoms on disease progression. They used an unsupervised clustering approach to group the patients according to the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) part III scores (Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 2003), and developed a supervised algorithm to determine which features predicted changes in cluster assignment over time. They identified bradykinesia as the most influential attribute in their model. Zhang et al. (2019) first trained Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks to encode sequences of input clinical observations in order to predict a set of target measures, which included clinical and biological variables (such as Dopamine Transporter [DAT] scan and CSF measurements). Next, they used Dynamic Time Warping (DTW; Bellman and Kalaba, 1959) to estimate the similarity between sequences of LSTM activations for each pair of patients. Finally, Student t-distributed Stochastic Neighbor Embedding (t-SNE; Van der Maaten and Hinton, 2008) was used to compress patients into a 2-dimensional space which preserved DTW distances, and patients were divided into three subtypes using k-means clustering in the compressed space. The three clusters differed in terms of age, baseline and slope of motor and cognitive symptoms, as well as DAT scan level decline.

Integrating supervised and unsupervised models should be an increasing focus of future work in order to ensure that we define subtypes or other latent variables that are potentially useful. To this end, research should combine the unsupervised discovery of latent factors or subtypes that explain heterogeneity in patient characteristics with supervised learning to predict latent scores from baseline symptoms and/or to predict future symptoms from latent scores.

Modalities

Clinical data

Unsurprisingly, almost all machine learning studies (96/110) made use of clinical or cognitive variables, including patients' diagnosis as well as motor, cognitive, psychiatric, and functional/daily-living symptoms.

These variables were commonly used as model inputs, targets for prediction (for supervised models), or both. The most common scale used was the UPDRS, which measures patients across all these domains. Other clinical variables included measures of dopaminergic therapy (Valmarska et al., 2018; Weintraub et al., 2020; Lim et al., 2021; Severson et al., 2021) and rapid eye movement (REM) sleep behavior disorder (RBD; Prashanth et al., 2014; Tsiouris et al., 2020; Chen et al., 2021). Seventeen studies used only clinical and demographic variables. For example, the three hybrid clustering-supervised-learning papers we described above used only clinical data as inputs to the clustering model.

Leveraging recent advances in deep learning, de Jong et al. (2019) presented a new method (VaDER) that combined an LSTM and a variational autoencoder (Kingma and Welling, 2013) to model multivariate clinical time series data. This method produced clusters with clinically divergent disease progression profiles. The identified clusters differed in gender, Schwab and England score and symptoms of depression. While they did not externally validate or assess the stability of these clusters (see *Validation and Data Leakage*, below), they did test for statistical differences between clusters in brain areas including the caudate. In the supervised domain, Prashanth et al. (2014) developed a SVM model for predicting PD diagnosis from olfactory deficits (measured by the University of Pennsylvania Smell Identification Test, UPSIT) as well as REM sleep disorder symptoms (measured by the REM Sleep Behavior Disorder Screening Questionnaire, RBDSQ), which reached a sensitivity of 88% and a specificity of 78% on an unseen validation set.

Neuroimaging markers

The next most used modality was neuroimaging, with 79 studies using some form of brain imaging data. PPMI neuroimaging includes structural and resting-state functional MRI, as well as Diffusion Tensor Imaging (DTI) data, and DAT scans. DAT imaging, a SPECT technique which measures dopamine transporter binding, was the most common neuroimaging modality in the machine learning papers we reviewed, specifically measured in the basal ganglia, most often using Striatal Binding Ratio (SBR).

DAT imaging has been extensively utilized as an input biomarker variable across the studies we evaluated. One study demonstrated that the inclusion of DAT features in addition to demographics and clinical measures improved prediction of UPDRS scores at 4 years post-baseline (Rahmim et al., 2017). It may be that DAT imaging represents a distinct domain, or possibly an endophenotype of PD progression. We observed that across these studies, DAT scan features were mainly limited to SBR, although extraction of novel image features has also been performed (Wenzel et al., 2019; Adams et al., 2021; Zhou and Tagare, 2021).

The information DAT imaging yields regarding dopaminergic tone may make it a key component in models of PD. However, it is worth noting that DAT level measurements are currently used in the consensus diagnosis of PD for PPMI subjects. That is, participants are not classified as having PD unless they exhibit a DAT scan deficit, even if they have motor abnormalities associated with the disorder. Because of this, subjects in the PPMI database labeled PD are guaranteed to have lower DAT levels in the basal ganglia than healthy controls or other groups. This is an interesting example of circularity in machine learning evaluation, which we will discuss in more detail below. We urge caution in the use of DAT levels in machine learning models containing PD diagnosis as a variable, especially in supervised models designed to predict diagnosis.

There were other examples utilizing imaging features from structural, functional, or DTI MRI scans (Badea et al., 2017; Peng et al., 2017; Amoroso et al., 2018; Singh et al., 2018; Won et al., 2019; Chan et al., 2022). One study by Uribe et al. used unsupervised hierarchical cluster analysis of structural MRI data to produce subtypes of untreated PD patients based on cortical atrophy patterns at baseline (Uribe et al., 2018). Here, two subtypes were identified, both of which were characterized by cortical thinning when compared to healthy controls. One of the subgroups showed more pronounced cortical atrophy and exhibited lower cognition scores. In a supervised learning study, Zhang et al. (2018) used a Graph Neural Network to classify patients vs. controls using DTI data.

However, such studies were relatively rare. One potential reason for this is that while other sources of data are provided in tabular form and require little preprocessing, MRI images require extensive processing and feature engineering before they can be incorporated into most machine learning models [convolution neural networks, which can work with raw or preprocessed images directly, are a notable exception and have been used in some PPMI studies (Yagis et al., 2021)]. Opportunities thus exist to further explore the utility of imaging markers. However, caution is warranted, and MRI preprocessing steps should be validated and standardized moving forward. Testing different DTI preprocessing pipelines on PPMI data, for example, Mishra et al. (2019) concluded that results are heavily dependent on the choice of preprocessing and analysis.

Genetic and other biomarkers

Relatively few machine learning studies used information from DNA (18 studies), RNA (5 studies), or CSF-based biomarker (27 studies) measurements. Forty-one studies made use of at least one such biomarker, while only 10 used multiple (see Figure 2 and *Multi-modal markers*, below).

Using CSF measures as well as non-motor clinical scores, Leger et al. (2020) compared several different supervised machine learning techniques to classify participants as PD, healthy controls, or subjects without evidence of dopamine deficiency (SWEDD). Gradient boosting decision trees (XGBoost, Chen et al., 2015) performed the best at distinguishing these classes, with an area-under-the-receiver-operating-curve (AUC) of 0.92 for PD patients versus healthy controls and 0.82 for patients versus SWEDDs. CSF alpha-synuclein was an important feature for distinguishing PD patients from controls.

Weintraub et al. (2022) built a model predicting impulse control disorders in Parkinson's patients, combining DNA from a genome wide association study (GWAS), genes previously implicated in impulse control disorders, and clinical variables to separate patients into low and high impulse control groups, with moderate accuracy (AUC of 0.72). Cope et al. (2021) developed a novel algorithm to generate a knowledge graph based on prior literature and known protein co-expression patterns, which is used to select DNA features and multiplicative interactions between features for use in standard machine learning algorithms. This method outperformed standard polygenetic risk scores in detecting PD (AUC = 0.85), highlighting potential interactions between known genetic risk loci and pathways.

RNA was the most infrequently used modality in the machine learning papers we reviewed. In one rare example using blood-based RNA sequencing, Pantaleo et al. (2022) used XGBoost models to classify subjects as PD patients or healthy controls from baseline gene expression data. The approach reached moderate accuracy (AUC of 0.72) and

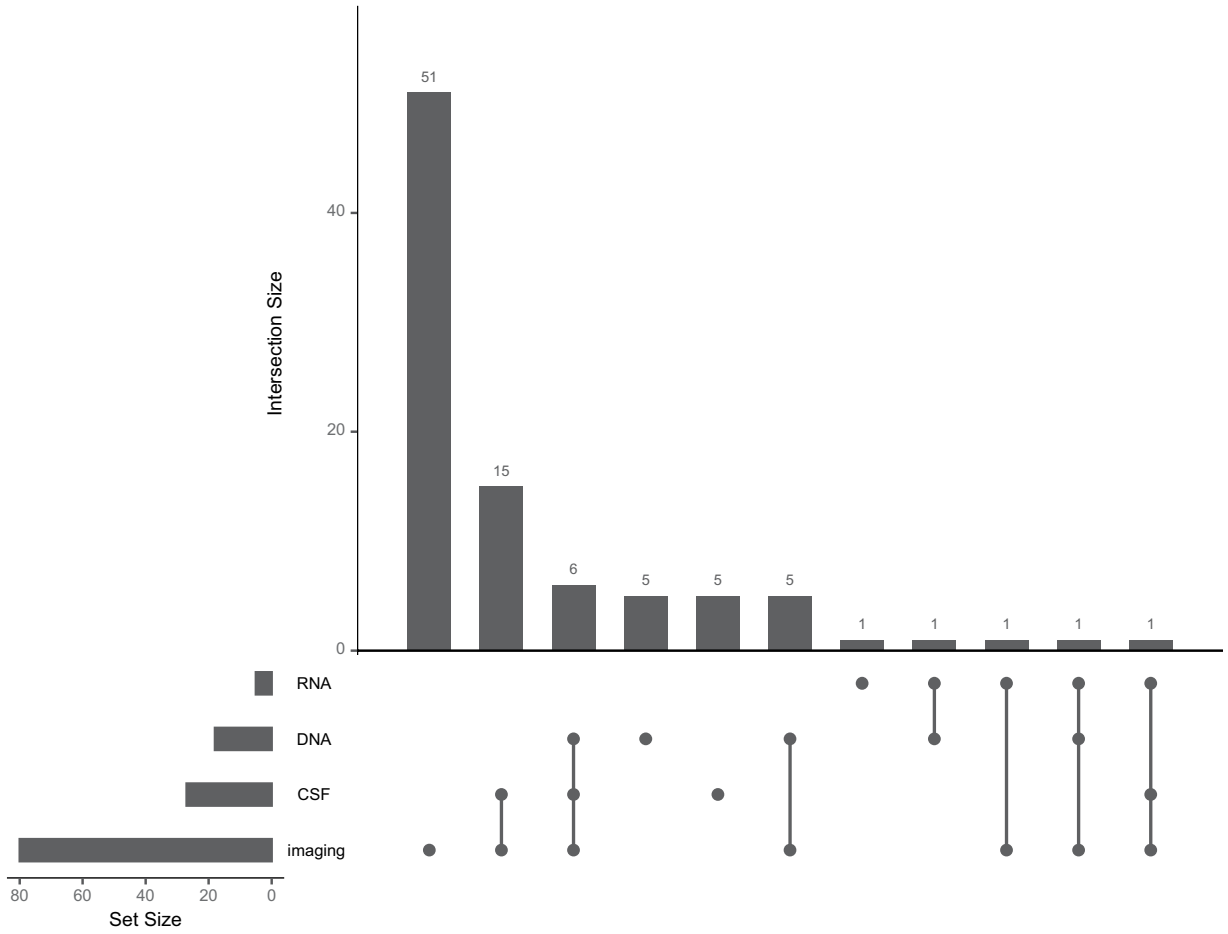

FIGURE 2
The use of multiple biological modalities is rare in machine learning studies of Parkinson's Progression Markers Initiative data. UpSet plot (Lex et al., 2014; Conway et al., 2017) showing frequencies of different combinations of imaging, CSF, DNA, and RNA data in the machine learning studies we reviewed. Each set of connected dots represents a combination of modalities (bottom). Intersection size is plotted as the number of studies for a particular modality combination (top). The set size is plotted as the overall number of studies using each modality (left). Ninety-two papers used at least one biological modality, including Dopamine Transport imaging striatal binding ratio scores. Thirty studies (out of 110 total) combined multiple biological modalities.

generated a large number of potential genes and expression pathways to be validated and explored in future work.

The limited number of papers using DNA, RNA, and CSF protein measures highlights the fact that these non-clinical data have been under-utilized to date in PPMI machine learning reports. These data modalities represent a potentially untapped resource and should be a major focus of future work.

Multi-modal markers

One of the most promising aspects of PPMI is the potential to combine multiple modalities to gain a broader perspective on biological mechanisms of patient heterogeneity. A majority of the studies we reviewed made some use of multiple modalities, although many of these used a single biological modality to predict clinical diagnosis. Many others combined clinical scores with DAT SBR alone. Relatively few studies have taken advantage of the breadth of biological data made available as part of PPMI. Of the 41 studies which made use of either DNA, RNA, or CSF measures, for example, no study used all three. In general, few studies (30/110) integrated multiple biological modalities (imaging, DNA, RNA, or CSF). Figure 2 shows the frequency of each biological modality as well as of multimodal combinations in the literature we reviewed. As for neuroimaging, most multimodal studies with brain imaging used DAT scan measurements in the basal ganglia, with fewer using structural or functional MRI or DTI.

In one example of multimodal analysis of symptom progression, Tsiouris et al. (2017) developed a supervised algorithm for selecting and combining subsets of features based on their discriminative ability, which they combined with Naïve Bayes (Domingos and Pazzani, 1997) and Repeated Incremental Pruning to Produce Error Reduction (RIPPER; Cohen, 1995) classifiers. They trained their feature selection algorithm and classifier to predict which category of rapid progression (measured as the quantile of the slope of UPDRS I-III total score) patients would fall into at 2-and 4-year epochs, using around 600 baseline features including selected SNPs, CSF-and blood-based biomarkers, clinical scores, and DAT and MRI features.

In another study or motor symptom progression, Latourelle et al. (2017) used a Bayesian machine learning based approach known as Reverse Engineering and Forward Simulation to form an ensemble of supervised models to predict symptom trajectory from genetic (53 pre-specified SNPs, PCs as well as a relatively large number (>17 K) of SNPs after pruning for linkage disequilibrium) and baseline CSF and clinical variables. Results showed baseline motor function, sex, age, DNA, and CSF biomarkers to be predictors

distinguishing slow and fast progressors (27% of variance explained in PD patients in leave-one-out cross-validation). Notably, the contributions of biological variables were relatively small, with all SNPs explaining <3% of variance in progression and CSF measures explaining <0.5%. They also tested the model ensemble in an external non-PPMI cohort (LABS-PD; Ravina et al., 2009), where it showed reduced, but significant, performance (9% of progression variance explained). Notably, there was no overlap between the biological features reported as important in this model and the Tsiouris et al. model described above, indicating that there is much more work to be done in studying the stability of machine learning findings in PPMI.

Multimodal models were also used to predict progression outside of the motor domain. In an example focused on cognitive decline, Schrag et al. (2017) built a logistic regression model containing features from DAT imaging, Montreal Cognitive Assessment (MoCA), UPSIT, and CSF Abeta/Tau ratio as predictors of cognitive impairment in PD 2 years from baseline. They reported robust performance determined by AUC of ~0.8, higher when including biological variables compared to using age alone. However, this study seems to have filtered features based on their ability to differentiate cognitively impaired subjects before performing cross-validation or splitting data into training and test sets (see *Validation and data leakage*, below), which calls the accuracy of these performance estimates into question.

Unsupervised learning of multiple modalities can be difficult, in part because high-dimensional modalities can dominate contributions to latent dimensions or classes. Using an approach known as similarity network fusion (SNF), Markello et al. (2021) were able to combine imaging, clinical, and CSF data modalities available in the PPMI cohort. They showed that this approach better captures similarities between patients across modalities with different dimensions than standard approaches of concatenating data across modalities. The fused similarity networks can then be decomposed *via* clustering or continuous latent factor algorithms. Interestingly, they tested both approaches in this paper, comparing diffusion embedding (continuous axes) with spectral clustering (discrete subtypes), two methods for unsupervised learning on graphs. They found that a continuous latent space provided the more parsimonious fit to the data.

One issue which is often overlooked when using biological modalities for prediction is how performance compares to a baseline model with only clinical data. In one cautionary example, Simuni et al. (2016) looked to identify clinical and biological predictors of time to initiation of treatment using RFs. They found that baseline clinical measures were most strongly predictive, and that the inclusion of biomarker data such as P-tau and DAT uptake did not improve predictive performance of the model above this baseline.

For the most part, machine learning papers have so far failed to make use of the uniquely multimodal data available from the PPMI project. Future work should focus on validating algorithms for combining multiple high dimensional modalities and testing whether biological data can improve predictions of patient trajectories above baseline clinical scores. However, the combination of multiple high dimensional sources of data will create problems of its own.

Validation and data leakage

Because machine learning approaches often use powerful algorithms with more tunable parameters than observations, overfitting is a serious concern, and it is essential that findings be validated by estimating stability and performance on unseen data. This is often done with one or more of the following techniques: Monte Carlo sampling, in which training and test sets are repeatedly randomly sampled; k-fold cross-validation, in which observations are divided into k groups, which are then cycled through with each group serving as the test set for one round of training; or maintaining a separate validation set to test performance only after all modeling decisions have been made. In the case of PPMI, sometimes other PD studies with overlapping observations are used as an external validation set, providing an even more meaningful test of model generalizability. In any case, it is essential that parameters-including hyperparameters such as which features or observations to include, what optimization algorithms to use, or even which particular model out of multiple possibilities is selected-be chosen only after isolating a test set from the fitting process.

Surprisingly, 13 supervised studies of the 97 we reviewed described a measure of predictive accuracy but showed no evidence of having a validation procedure for evaluating it. There are more subtle issues in validating predictive models than simply ignoring validation, however. In many cases, validation is performed in such a manner that information has been "leaked" from test set observations to the model fitting procedure, thus invalidating the validation process (Yagis et al., 2021). In one common form of data leakage, features are selected based on statistical tests (e.g., differentiating PD from healthy controls), before splitting the data into training or test sets to train a classifier to differentiate PD from healthy controls. This is circular, because the variables have already been selected because they distinguish PD from healthy controls across both training and test subjects. Twenty more of the studies we reviewed (in addition to those with no validation method at all) showed evidence of such data leakage.

Overall, the lack of attention paid to model validation and clear separation of training and test sets is concerning. In addition to the 33% of supervised studies described above, in others it was unclear from methods precisely how models were evaluated: at what point in the analysis pipeline training and test sets were separated, or even whether this occurred at all. Without clear descriptions of this process, papers are not reproducible and evaluating their results is challenging.

In unsupervised settings, even less attention is paid to model validation. How exactly to validate unsupervised models can be a difficult question, since there are no labels or target variables to predict. The answer will depend on the exact model being used, and detailed treatment is outside the scope of this review. However, latent factor models with continuous axes can in general be treated like regression models in terms of evaluation and validated the same way (e.g., variance explained or mean squared error on held out data; Owen and Perry, 2009). In principle, clustering methods can be treated similarly and evaluated on held-out data, since many clustering models can be considered special cases of continuous dimensionality reduction models (Fu and Perry, 2020). This seems to be rarely done in practice, and of the 19 studies using unsupervised machine learning we reviewed here, 13 failed to provide any prediction-based validation of their unsupervised learning models. Other studies varied in how they tested stability or validity of their models, with only 4 studies explicitly using held-out data to evaluate.

As noted above, external validation can provide a more stringent test of model generalizability than held-out data. After all, we would

like our findings to apply to broader populations of patients than those meeting the criteria for enrollment in a specific study. Datasets such as PARS (Parkinson's Associated Risk Study), Penn-Udall (Morris K Udall Parkinson's Disease Research Center of Excellence cohort), and 23andMe were used for external validation by a small number of studies reviewed here. These datasets should be extensively evaluated to understand the similarities and differences with PPMI and their suitability as independent replication/validation datasets, while considering known confounders such as age, sex, and other demographics (Van Den Eeden et al., 2003; Pringsheim et al., 2014; Simuni et al., 2016; Hayete et al., 2017; Latourelle et al., 2017; Rutten et al., 2017; Schrag et al., 2017); which measurements were collected; and differences in preprocessing of biospecimen and imaging data.

Future papers need to present a clear strategy for estimating validity and generalizability of findings. Authors should report how evaluation procedures were kept separate from model training in a clear and reproducible manner. In addition, papers should justify the evaluation metrics chosen for a particular study (sensitivity, specificity, AUC, recall, etc.) based on the costs and benefits of different types of model errors and on any potential issues with specific metrics in a particular setting (e.g., severely imbalanced classes).

Discussion

One striking aspect of the machine learning PPMI literature we reviewed here is the lack of overlap between findings across studies. This is perhaps unsurprising given how much studies differed in their input features, target variables, and modeling approaches. Based on our review of this variability, we make the following recommendations to improve future PPMI machine learning research:

- One of the major goals of PPMI is to understand why different PD patients have different disease trajectories. Machine learning studies should place more focus on predicting variability in patient symptoms rather than distinguishing PD patients from healthy controls, especially variability in future symptom progression from baseline measurements.
- There is increasing interest in sub-typing or phenotyping PD patients, but we need to ensure that our latent categories or dimensions are consistent, predictable, and useful. Much more work is needed in testing the stability across different unsupervised methods (see Alexander et al., 2021 for an example in Alzheimer's research), and for validating proposed sub-types. Algorithms should also be developed and validated for combining sub-typing or latent factor discovery with supervised learning of patient outcomes.
- Future work should incorporate the multiple biological modalities available to PPMI researchers. This is especially true of raw MRI and DTI images, DNA, and RNA transcriptomic data, as well as multimodal combinations of these domains. Preprocessing pipelines for each domain should be validated and standardized across studies.
- For longitudinal studies with biological modalities, studies should report the extent to which a model improves upon results using only baseline clinical and demographic information.
- More care needs to be taken to validate machine learning results in PPMI studies. Supervised and unsupervised methods need to be tested on unseen data. Measures of predictive accuracy such as AUC should not be reported on training data alone, especially after filtering variables for significance or prediction improvement. When performing validation, data leakage— often caused by preprocessing or variable selection involving both training and test data sets— needs to be more scrupulously avoided in future work. Finally, these procedures need to be articulated more clearly and in a manner that can be reproduced by other researchers. To suggest truly generalizable results, findings should eventually be validated in data sets external to PPMI.

Author contributions

LLa conceptualized the work and contributed to the review of articles and writing of manuscript. RG contributed to the review of articles and writing of manuscript. AP, EW, and LLi contributed to the review of articles. MH conceptualized the work. All authors contributed to the article and approved the submitted version.

Funding

This work was jointly funded by Cohen Veterans Bioscience (COH-0003) and a generous grant from the Michael J. Fox Foundation as part of the Parkinson's Progression Markers Initiative (PPMI) (MJFF-019075). Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmiinfo.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI is a public-private partnership funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. The list with full names of all PPMI funding partners found at www.ppmi-info.org/fundingpartners.

Acknowledgments

We would like to thank Luba Smolensky, Ken Marek, Andrew Siderowf, and Mark Frasier for helpful discussions during the initial landscaping phase.

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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adams, M. P., Rahmim, A., and Tang, J. (2021). Improved motor outcome prediction in Parkinson's disease applying deep learning to DaTscan SPECT images. *Comput. Biol. Med.* 132:104312. doi: 10.1016/j.compbiomed.2021.104312

Alexander, N., Alexander, D. C., Barkhof, F., and Denaxas, S. (2021). Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med. Inform. Decis. Mak.* 21, 1–13. doi: 10.1186/s12911-021-01693-6

Amoroso, N., La Rocca, M., Monaco, A., Bellotti, R., and Tangaro, S. (2018). Complex networks reveal early MRI markers of Parkinson's disease. *Med. Image Anal.* 48, 12–24. doi: 10.1016/j.media.2018.05.004

Arrowsmith, J., and Miller, P. (2013). Trial watch: phase II and phase III attrition rates 2011-2012. *Nat. Rev. Drug Discov.* 12:569. doi: 10.1038/nrd4090

Badea, L., Onu, M., Wu, T., Roceanu, A., and Bajenaru, O. (2017). Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLoS One* 12:e0188196. doi: 10.1371/journal.pone.0188196

Baker, M. (2016). Reproducibility crisis. *Nature* 533, 353–366.

Bellman, R. (1956). Dynamic programming and lagrange multipliers. *Proc. Natl. Acad. Sci. U. S. A.* 42, 767–769. doi: 10.1073/pnas.42.10.767

Bellman, R., and Kalaba, R. (1959). On adaptive control processes. *IRE Trans. Autom. Control.* 4, 1–9. doi: 10.1109/TAC.1959.1104847

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chahine, L. M., Brumm, M. C., Caspell-Garcia, C., Oertel, W., Mollenhauer, B., Amara, A., et al. (2021). Dopamine transporter imaging predicts clinically-defined α-synucleinopathy in REM sleep behavior disorder. *Ann. Clin. Trans. Neurol.* 8, 201–212. doi: 10.1002/acn3.51269

Chan, Y. H., Wang, C., Soh, W. K., and Rajapakse, J. C. (2022). Combining neuroimaging and omics datasets for disease classification using graph neural networks. *Front. Neurosci.*:605. doi: 10.3389/fnins.2022.866666

Chang, D., Nalls, M. A., Hallgrímsdóttir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., et al. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* 49, 1511–1516. doi: 10.1038/ng.3955

Chen, T., He, Tong, Benesty, Michael, Khotilovich, Vadim, Tang, Yuan, Cho, Hyunsu, et al. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2.

Chen, F., Li, Y., Ye, G., Zhou, L., Bian, X., and Liu, J. (2021). Development and validation of a prognostic model for cognitive impairment in Parkinson's disease with REM sleep behavior disorder. *Front. Aging Neurosci.*:416. doi: 10.3389/fnagi.2021.703158

Cohen, W.W. (1995). *Fast effective rule induction*. In A. Prieditis and S. Russell (Eds.), Proceedings of the 12th International Conference on Machine Learning (ML-95), Lake Tahoe, California, (pp. 115–123). Morgan Kaufmann, San Mateo, CA.

Combs, H. L., Wyman-Chick, K. A., Erickson, L. O., and York, M. K. (2021). Development of standardized regression-based formulas to assess meaningful cognitive change in early Parkinson's disease. *Arch. Clin. Neuropsychol.* 36, 734–745. doi: 10.1093/arclin/acaa104

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. (Accessed September 15, 2017) doi: 10.1093/bioinformatics/btx364

Cope, J. L., Baukmann, H. A., Klinger, J. E., CNJ, R., Böttinger, E. P., Konigorski, S., et al. (2021). Interaction-based feature selection algorithm outperforms polygenic risk score in predicting Parkinson's disease status. *Front. Genet.* 12:744557. doi: 10.3389/fgene.2021.744557

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

D'Cruz, N., Vervoort, G., Chalavi, S., Dijkstra, B. W., Gilat, M., and Nieuwboer, A. (2021). Thalamic morphology predicts the onset of freezing of gait in Parkinson's disease. *NPJ Parkinson's Dis.* 7, 1–10. doi: 10.1038/s41531-021-00163-0

de Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., et al. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Giga Sci.* 8:giz 134. doi: 10.1093/gigascience/giz134

Domingos, P., and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* 29, 103–130. doi: 10.1023/A:1007413511361

Faghri, F., Hashemi, S. H., Leonard, H., Scholz, S. W., Campbell, R. H., Nalls, M. A., et al. (2018). Predicting onset, progression, and clinical subtypes of Parkinson disease using machine learning. *bio Rxiv*:338913. doi: 10.1101/338913

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* 23, 584–601. doi: 10.1016/j.tics.2019.03.009

Fereshtehnejad, S.-M., Romenets, S. R., Anang, J. B., Latreille, V., Gagnon, J. F., and Postuma, R. B. (2015). New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. *JAMA Neurol.* 72, 863–873. doi: 10.1001/jamaneurol.2015.0703

Freeze, B., Acosta, D., Pandya, S., Zhao, Y., and Raj, A. (2018). Regional expression of genes mediating trans-synaptic alpha-synuclein transfer predicts regional atrophy in Parkinson disease. *Neuro Image Clin.* 18, 456–466. doi: 10.1016/j.nicl.2018.01.009

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* 28, 337–407. doi: 10.1214/aos/1016218223

Fu, W., and Perry, P. O. (2020). Estimating the number of clusters using cross-validation. *J. Comput. Graph. Stat.* 29, 162–173. doi: 10.1080/10618600.2019.1647846

Gramotnev, G., Gramotnev, D. K., and Gramotnev, A. (2019). Parkinson's disease prognostic scores for progression of cognitive decline. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-020-76437-z

Gu, S.-C., Zhou, J., Yuan, C. X., and Ye, Q. (2020). Personalized prediction of depression in patients with newly diagnosed Parkinson's disease: a prospective cohort study. *J. Affect. Disord.* 268, 118–126. doi: 10.1016/j.jad.2020.02.046

Harrison, R. K. (2016). Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug Discov.* 15, 817–818. doi: 10.1038/nrd.2016.184

Hayete, B., Wuest, D., Laramie, J., McDonagh, P., Church, B., Eberly, S., et al. (2017). A Bayesian mathematical model of motor and cognitive outcomes in Parkinson's disease. *PLoS One* 12:e0178982. doi: 10.1371/journal.pone.0178982

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Jackson, H., Anzures-Cabrera, J., Taylor, K. I., and Pagano, G.PASADENA Investigators and Prasinezumab Study Group (2021). Hoehn and Yahr stage and striatal Dat-SPECT uptake are predictors of Parkinson's disease motor progression. *Front. Neurosci.*:1595. doi: 10.3389/fnins.2021.765765

Kim, R., and Jeon, B. (2021). Serum neurofilament light chain predicts future freezing of gait in Parkinson's disease. *Parkinsonism Relat. Disord.* 91, 102–104. doi: 10.1016/j.parkreldis.2021.08.015

Kim, R., Lee, J., Kim, H. J., Kim, A., Jang, M., Jeon, B., et al. (2019). CSF β-amyloid 42 and risk of freezing of gait in early Parkinson disease. *Neurology* 92, e40–e47. doi: 10.1212/WNL.0000000000006692

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv* 1312:6114. doi: 10.48550/arXiv.1312.6114

Krystal, J. H., and State, M. W. (2014). Psychiatric disorders: diagnosis to therapy. *Cells* 157, 201–214. doi: 10.1016/j.cell.2014.02.042

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 490, 187–191. doi: 10.1038/nature11556

Latourelle, J. C., Beste, M. T., Hadzi, T. C., Miller, R. E., Oppenheim, J. N., Valko, M. P., et al. (2017). Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol.* 16, 908–916. doi: 10.1016/S1474-4422(17)30328-9

Leger, C., Herbert, M., and DeSouza, J. F. (2020). Non-motor clinical and biomarker predictors enable high cross-validated accuracy detection of early PD but lesser cross-validated accuracy detection of scans without evidence of dopaminergic deficit. *Front. Neurol.* 11:364. doi: 10.3389/fneur.2020.00364

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Lim, J., Ji, C. X., Oberst, M., Blecker, S., Horwitz, L., and Sontag, D. (2021). Finding regions of heterogeneity in decision-making via expected conditional covariance. *Adv. Neural Inf. Proces. Syst.* 34, 15328–15343. doi: 10.48550/arXiv.2110.14508

Ma, L.-Y., Tian, Y., Pan, C.-R., Chen, Z.-L., Ling, Y., Ren, K., et al. (2021). Motor progression in early-stage Parkinson's disease: a clinical prediction model and the role of cerebrospinal fluid biomarkers. *Front. Aging Neurosci.* 12:627199. doi: 10.3389/fnagi.2020.627199

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., et al. (2018). The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort. *Ann. Clin. Trans. Neurol.* 5, 1460–1477. doi: 10.1002/acn3.644

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

Markello, R. D., Shafiei, G., Tremblay, C., Postuma, R. B., Dagher, A., and Misic, B. (2021). Multimodal phenotypic axes of Parkinson's disease. *NPJ Parkinson's Dis.* 7, 1–12. doi: 10.1038/s41531-020-00144-9

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biol. Psychiatry* 1, 433–447. doi: 10.1016/j.bpsc.2016.04.002

McShane, L. (2017). In pursuit of greater reproducibility and credibility of early clinical biomarker research. *Clin. Transl. Sci.* 10:58. doi: 10.1111/cts.12449

Mishra, V. R., Sreenivasan, K. R., Zhuang, X., Yang, Z., Cordes, D., and Walsh, R. R. (2019). Influence of analytic techniques on comparing DTI-derived measurements in early stage Parkinson's disease. *Heliyon* 5:e01481. doi: 10.1016/j.heliyon.2019.e01481

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

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* 46, 989–993. doi: 10.1038/ng.3043

Nguyen, A. A., et al. (2020). Dynamical role of pivotal brain regions in Parkinson symptomatology uncovered with deep learning. *Brain Sci.* 10:73. doi: 10.3390/brainsci10020073

Nievergelt, C. M., Maihofer, A. X., Klengel, T., Atkinson, E. G., Chen, C. Y., Choi, K. W., et al. (2019). International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nat. Commun.* 10, 1–16. doi: 10.1038/s41467-019-12576-w

Owen, A. B., and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* 3, 564–594. doi: 10.1214/08-AOAS227

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372:n160. doi: 10.1136/bmj.n160

Pantaleo, E., Monaco, A., Amoroso, N., Lombardi, A., Bellantuono, L., Urso, D., et al. (2022). A machine learning approach to Parkinson's disease blood transcriptomics. *Genes* 13:727. doi: 10.3390/genes13050727

Peng, B., Wang, S., Zhou, Z., Liu, Y., Tong, B., Zhang, T., et al. (2017). A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease. *Neurosci. Lett.* 651, 88–94. doi: 10.1016/j.neulet.2017.04.034

Prashanth, R., Roy, S. D., Mandal, P. K., and Ghosh, S. (2014). Parkinson's disease detection using olfactory loss and REM sleep disorder features. *Ann. Int. Conf. IEEE Eng. Med. Biol. Soc* 2014, 5764–5767. doi: 10.1109/EMBC.2014.6944937

Pringsheim, T., Jette, N., Frolkis, A., and Steeves, T. D. (2014). The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord.* 29, 1583–1590. doi: 10.1002/mds.25945

Rahmim, A., Maia, P. D., Gao, X., F Damasceno, P., and Raj, A. (2017). Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images. *Neuro Image Clin.* 16, 539–544.

Ravina, B., Tanner, C., Dieuliis, D., Eberly, S., Flagg, E., and Galpern, W. R. (2009). A longitudinal program for biomarker development in Parkinson's disease: a feasibility study. *Mov. Disord.* 24, 2081–2090. doi: 10.1002/mds.22690

Ren, A. H., Fiala, C. A., Diamandis, E. P., and Kulasingam, V. (2020). Pitfalls in cancer biomarker discovery and validation with emphasis on circulating tumor DNA. *Cancer Epidemiol. Biomark. Prev.* 29, 2568–2574. doi: 10.1158/1055-9965.EPI-20-0074

Ren, X., Lin, J., Stebbins, G. T., Goetz, C. G., and Luo, S. (2021). Prognostic modeling of Parkinson's disease progression using early longitudinal patterns of change. *Mov. Disord.* 36, 2853–2861. doi: 10.1002/mds.28730

Rutten, S., van der Ven, P. M., Weintraub, D., Pontone, G. M., Leentjens, A. F. G., Berendse, H. W., et al. (2017). Predictors of anxiety in early-stage Parkinson's disease–results from the first two years of a prospective cohort study. *Parkinsonism Relat. Disord.* 43, 49–55. doi: 10.1016/j.parkreldis.2017.06.024

Salmanpour, M. R., Shamsaei, M., Saberi, A., Setayeshi, S., Klyuzhin, I. S., Sossi, V., et al. (2019). Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. *Comput. Biol. Med.* 111:103347. doi: 10.1016/j.compbiomed.2019.103347

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., et al. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* 87, 1215–1233. doi: 10.1016/j.neuron.2015.09.016

Scherer, A. (2017). Reproducibility in biomarker research and clinical development: a global challenge. *Fut. Med.* 11, 309–312. doi: 10.2217/bmm-2017-0024

Schrag, A., Siddiqui, U. F., Anastasiou, Z., Weintraub, D., and Schott, J. M. (2017). Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. *Lancet Neurol.* 16, 66–75. doi: 10.1016/S1474-4422(16)30328-3

Severson, K. A., Chahine, L. M., Smolensky, L. A., Dhuliawala, M., Frasier, M., Ng, K., et al. (2021). Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. *Lancet Digital Health* 3, e555–e564. doi: 10.1016/S2589-7500(21)00101-1

Shu, Z., Pang, P., Wu, X., Cui, S., Xu, Y., and Zhang, M. (2020). An integrative nomogram for identifying early-stage Parkinson's disease using non-motor symptoms and white matter-based radiomics biomarkers from whole-brain MRI. *Front. Aging Neurosci.* 12:457. doi: 10.3389/fnagi.2020.548616

Simuni, T., Long, J. D., Caspell-Garcia, C., Coffey, C. S., Lasch, S., Tanner, C. M., et al. (2016). Predictors of time to initiation of symptomatic therapy in early Parkinson's disease. *Ann. Clin. Transl. Neurol.* 3, 482–494. doi: 10.1002/acn3.317

Singh, G., Samavedham, L., and Lim, E. C.Alzheimer's Disease Neuroimaging Initiative (2018). Determination of imaging biomarkers to decipher disease trajectories and differential diagnosis of neurodegenerative diseases (DIsease Tre ND). *J. Neurosci. Methods* 305, 105–116. doi: 10.1016/j.jneumeth.2018.05.009

Soltaninejad, S., Basu, A., and Cheng, I.. (2019). Automatic classification and monitoring of denovo parkinson's disease by learning demographic and clinical features. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany.

Sun, Q., Welsh, K. J., Bruns, D. E., Sacks, D. B., and Zhao, Z. (2019). Inadequate reporting of analytical characteristics of biomarkers used in clinical research: a threat to interpretation and replication of study findings. *Clin. Chem.* 65, 1554–1562. doi: 10.1373/clinchem.2019.309575

Tang, J., Yang, B., Adams, M. P., Shenkov, N. N., Klyuzhin, I. S., Fotouhi, S., et al. (2019). Artificial neural network–based prediction of outcome in Parkinson's disease patients using DaTscan SPECT imaging features. *Mol. Imaging Biol.* 21, 1165–1173. doi: 10.1007/s11307-019-01334-5

Tang, C., Zhao, X., Wu, W., Zhong, W., and Wu, X. (2021). An individualized prediction of time to cognitive impairment in Parkinson's disease: a combined multi-predictor study. *Neurosci. Lett.* 762:136149. doi: 10.1016/j.neulet.2021.136149

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288.

Tsiouris, K. M., Konitsiotis, S., Koutsouris, D. D., and Fotiadis, D. I. (2020). Prognostic factors of rapid symptoms progression in patients with newly diagnosed Parkinson's disease. *Artif. Intell. Med.* 103:101807. doi: 10.1016/j.artmed.2020.101807

Tsiouris, K.M., Rigas, G., Gatsios, D., Antonini, A., Konitsiotis, S., Koutsouris, D. D., et al. (2017). Predicting rapid progression of Parkinson's disease at baseline patients evaluation. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju.

Uribe, C., Segura, B., Baggio, H. C., Abos, A., Garcia-Diaz, A. I., and Campabadal, A. (2018). Cortical atrophy patterns in early Parkinson's disease patients using hierarchical cluster analysis. *Parkinsonism Relat. Disord.* 50, 3–9. doi: 10.1016/j.parkreldis.2018.02.006

Valmarska, A., Miljkovic, D., Konitsiotis, S., Gatsios, D., Lavrač, N., and Robnik-Šikonja, M. (2018). Symptoms and medications change patterns for Parkinson's disease patients stratification. *Artif. Intell. Med.* 91, 82–95. doi: 10.1016/j.artmed.2018.04.010

Van Den Eeden, S. K., Tanner, C. M., Bernstein, A. L., Fross, R. D., Leimpeter, A., Bloch, D. A., et al. (2003). Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *Am. J. Epidemiol.* 157, 1015–1022. doi: 10.1093/aje/kwg068

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 579–2605.

Weintraub, D., Caspell-Garcia, C., Simuni, T., Cho, H. R., Coffey, C. S., Aarsland, D., et al. (2020). Neuropsychiatric symptoms and cognitive abilities over the initial quinquennium of Parkinson disease. *Ann. Clin. Trans. Neurol.* 7, 449–461. doi: 10.1002/acn3.51022

Weintraub, D., Posavi, M., Fontanillas, P., Tropea, T. F., Mamikonyan, E., Suh, E., et al. (2022). Genetic prediction of impulse control disorders in Parkinson's disease. *Ann. Clin. Trans. Neurol.* 9, 936–949. doi: 10.1002/acn3.51569

Wenzel, M., Milletari, F., Krüger, J., Lange, C., Schenk, M., Schenk, M., et al. (2019). Automatic classification of dopamine transporter SPECT: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics. *Eur. J. Nucl. Med. Mol. Imaging* 46, 2800–2811. doi: 10.1007/s00259-019-04502-5

Won, J. H., Kim, M., Park, B.-Y., Youn, J., and Park, H. (2019). Effectiveness of imaging genetics analysis to explain degree of depression in Parkinson's disease. *PLoS One* 14:e0211699. doi: 10.1371/journal.pone.0211699

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* 50, 668–681. doi: 10.1038/s41588-018-0090-3

Yagis, E., Atnafu, S. W., Seco de Herrera, A. G., Marzi, C., Scheda, R., Giannelli, M., et al. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.* 11, 1–13. doi: 10.1038/s41598-021-01681-w

Zeighami, Y., Fereshtehnejad, S. M., Dadar, M., Collins, D. L., Postuma, R. B., and Dagher, A. (2019). Assessment of a prognostic MRI biomarker in early de novo Parkinson's disease. *Neuro Image Clin.* 24:101986. doi: 10.1016/j.nicl.2019.101986

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, 1–12. doi: 10.1038/s41598-018-37545-z

Zhang, X., Chou, J., and Wang, F.. (2018). Integrative analysis of patient health records and neuroimages via memory-based graph convolutional network. Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore.

Zhou, Y., and Tagare, H. D. (2021). Self-normalized classification of Parkinson's disease DaTscan images. *Proc. IEEE Int. Conf. Bioinformatics Biomed.* 2021, 1205–1212. doi: 10.1109/bibm52615.2021.9669820