Neural Computing and Applications (2024) 36:19209–19230
https://doi.org/10.1007/s00521-024-10127-z

ORIGINAL ARTICLE

Explainable AI approach for early detection of Parkinson's disease using PPMI Online data

Nitisha Aggarwal[1] · Geetika Jain Saxena[2] · Sanjeev Singh[1] · Amit Pundir[2]

Received: 17 January 2023 / Accepted: 26 June 2024 / Published online: 3 August 2024

Abstract

Accurate and early disease prediction enables patients to plan and improve their quality of life in the future. Early detection of neurodegenerative diseases, such as Parkinson's disease, is a high priority and a significant challenge in which physicians must act quickly to diagnose and predict the risk of disease severity. Machine learning (ML) models combined with feature selection (FS) techniques can assist physicians in quickly diagnosing a disease. FS technique optimally subsets features to improve model performance and reduce the number of tests required for a patient, thereby speeding up diagnosis. This paper proposes an e-diagnosis approach based on ML-FS algorithms to detect Parkinson's disease using data obtained from Parkinson's Progression Markers Initiative (PPMI) Online study. Also, it can be considered patient-oriented research as it uses self-reported online collected data. The results of six FS techniques pre-applied to classification algorithms named logistic regression, random forest, support vector machine, CatBoost, extreme learning machine, and XGBoost are shown in this study. Chi-square, mutual information, and analysis of variance (ANOVA) filter-based FS methods, while sequential feature selection, Boruta, and recursive feature elimination are considered wrapper methods. The outcomes show that random forest when trained on features selected by the recursive feature elimination technique help to build an efficient and effective approach for detecting Parkinson's disease.

Keywords Parkinson's disease · Machine learning · Feature selection · SHAP · Explainable AI

1 Introduction

Parkinson's disease (PD) is a disorder of the central nervous system that immensely affects functions controlled by it. The quality of life of a large population above 60 years old is adversely affected by this progressive neurodegenerative disease globally [1]. PD is caused by the loss of neurons that produce dopamine in the substantia nigra. Due to this, the level of dopamine decreases and causes atypical brain activity. Research has established that the progression of PD has stages based on the reduction of dopaminergic neurons [2]. Resting tremors, bradykinesia, and rigidity are the most common motor symptoms of PD, and non-motor symptoms such as loss of olfactory, sleep behavior disorder, and cognitive and behavioral changes are also prevalent [3]. Treatment is effective in the early stage compared to later stages. Non-motor symptoms, recognized in the premotor or prodromal stage of PD, precede the motor symptoms in the mid or advanced stage. Currently, treatments are available only to maintain quality of life by relieving symptoms early on, and no cure is available for PD. Hence, detecting PD early is crucial to stagnating or delaying disease progression [4]. Several studies predicted that the prevalence of PD would skyrocket in the coming decades [5]. The increase in the global burden of PD is attributed primarily to the increasing number of older people worldwide. Other reasons are contributions from

✉ Amit Pundir
amitpundir@mac.du.ac.in

Nitisha Aggarwal
nitisha@south.du.ac.in

Geetika Jain Saxena
gsaxena@mac.du.ac.in

Sanjeev Singh
sanjeev@south.du.ac.in

[1] Institute of Informatics and Communication, University of Delhi South Campus, New Delhi, Delhi 110021, India

[2] Department of Electronics, Maharaja Agrasen College, University of Delhi, New Delhi, Delhi 110096, India

other long-prevailing diseases and environmental factors poised to significantly fuel the growth of PD worldwide.

Cost of health facilities, geographical reasons, and transportation limitations restrict individuals from taking benefits of clinical-based medical systems through in-person visits. For example, nearly one-third of persons with PD rarely visit a neurologist in the USA [6]. However, academic medical research centers are taking advantage of technological advances, digital medicine facilities, online healthcare systems, web-based strategies, real-time monitoring systems, and digital diagnosis decision systems to provide healthcare to many PD patients worldwide. In recent years, electronic health records have contributed significantly to understanding the diagnosis and progression of PD [7]. Technology-enabled healthcare systems helped the acquisition of data with ease through online/web-based systems and wearable health devices. Data is also collected through self-reported medical conditions from many patients using questionnaires and assisting in designing patient-oriented healthcare systems. Digital cohorts nowadays ease data acquisition. However, the incompleteness and reliability of data are still challenging. The incompleteness of datasets can be addressed to a very large extent these days by machine learning techniques. An approach involves the utilization of imputation algorithms or algorithms specifically designed to work with sparse datasets that are capable of handling missing data. Furthermore, varying levels of data completeness can be adapted by ML models by including informative and relevant features. By employing feature selection techniques, ML algorithms can prioritize the data points that contribute the most to predictive or diagnostic accuracy, effectively mitigating the effects of missing or incomplete data [8]. However, it should be acknowledged that while ML techniques can address the incompleteness of data to some extent, the reliability of data remains still a significant challenge. To address the reliability of data, a multi-faceted approach is necessary, encompassing rigorous data validation procedures, cross-validation techniques, and domain expertise. Hence, these systems need to be validated from traditional clinical in-person data studies, expert opinions, and assessments by neurologists/clinicians [9]. However, some projects such as Fox Insight (FI) [10, 11] and Parkinson’s Progression Markers Initiative Online (PPMI Online) [12, 13] are underway with a reasonable degree of agreement and well-designed questionnaires to collect and evaluate the data for PD.

Machine learning (ML) algorithms learn and extract meaningful patterns from available data to design digital diagnosis decision systems. ML has created enhanced interest in the diagnosis/prediction approach in many biomedical conditions such as diabetes [14], cardiovascular disease [15, 16], risk of coronavirus disease [17], human reproduction health including studies of embryos [18, 19], brain disorder [20], and mental health [21, 22], in addition to the diagnosis of PD. However, most studies have used data generated through invasive procedures, in-person clinic visits, and expensive protocols. It is notable that for more accurate detection of PD, more than one visit is necessary since, in the first visit investigations, pathologically confirmed PD is only 80% accurate [23]. Hence, longitudinal study of patients is vital to predicting PD accurately. Deep learning (DL) algorithms are used to analyze imaging or gait movements of PD patients, but generating such data require proper setup and is challenging when done remotely; hence, in-person visits to healthcare systems are necessary [24]. Studies also used multi-modal data to identify PD or stages of PD. However, these studies have high computational time and cost.

In the present study, data were extracted from PPMI Online Project and its characteristics were studied and compared with available contemporary datasets. This data has more than 120 characteristics that may or may not be relevant to identify PD. Hence, selected characteristics, as suggested by feature selection techniques, were used to differentiate PD from Healthy Controls (HC). More than one feature selection technique was used to get the best possible results and verify them. Data was trained on various ML classification algorithms based on bagging or boosting, statistical or feedforward learning, and deterministic or probabilistic concepts. The main objective of this study is to design an approach that eases the prediction process of PD with the least number of characteristics and reduces the need for invasive biomarkers or clinic visits to identify PD. It is well known that PD patients face difficulty even in everyday activities; hence, traveling and visiting medical research centers is a strenuous task for them. Therefore, an online-only approach can help them reduce their struggle to reach clinicians. Hence, we have attempted to design an online approach to identify prodromal PD with non-motor and subtle motor symptoms. Applying feature selection methods with classification models for PD detection is not novel, however, significant contributions are made in the current study using this approach. Firstly, the focus is on analyzing self-reporting online data, which have unique challenges and opportunities compared to traditional clinical datasets. Online data often come in large volumes and diverse formats, necessitating robust processing and analysis techniques to extract meaningful insights. Secondly, to the best of our knowledge, this study represents the first attempt to analyze PPMI Online data and statistically compare it with traditional clinical datasets. Additionally, the classification results are validated using explainable AI techniques for PD detection. By using explainable AI methods, the aim is to enhance the transparency and interpretability of the models, thereby providing insights into the underlying factors contributing to PD diagnosis. While the existence of previous studies combining FS with ML for PD detection is acknowledged, the novelty of the research lies in the application to online data and the validation using explainable AI techniques. Unfortunately, due to the approach's novelty and the dataset's challenging nature, a direct comparison with existing methods is deemed not feasible.

The utilization of online datasets by academic medical research centers to extend routine medical and preventive care for PD patients globally is investigated. With technological advancements and the growing availability of digital healthcare solutions, digital medicine facilities, online healthcare systems, web-based strategies, real-time monitoring systems, and digital diagnosis decision systems are being leveraged by medical research centers to improve online healthcare services. Motivated by the potential benefits of these online initiatives, the current study focuses on analyzing online data and statistically comparing it with well-established offline data sources. Through this comparative analysis, the aim is to establish the value and impact of online healthcare strategies in providing timely and effective healthcare services to individuals affected by PD, thereby contributing to the advancement of remote healthcare delivery. The paper proceeds to represent, in Sect. 2, the related work on PD using various techniques reported, and in the next section (Sect. 3), description of data, and statistical analysis of data given. Section 4 describes the methodology, preprocessing techniques applied to data, feature selection, and models used for training. Results and explanations of decisions taken by models are reported in Sect. 5. A comparison of different approaches to the prediction of PD is analyzed in this section. In the next section, this study has concluded with advantages and limitations. Finally, future aspects of this work are described in Sect. 6.

2 Related work

In the literature, several studies have been conducted using ML and DL algorithms on Parkinson's disease. Classification or Diagnosis of PD, present stages and progression of PD, or differential diagnosis were the main aim of these studies. Various data modalities such as neuroimaging data (MRI, SPECT, and PET) [25, 26], gait movement [27], voice recordings [28], handwriting patterns [29], biomarkers including CFS [30, 31], sleep behavior [32], RNA [33], Wearable sensor data [34], and Patient Questionnaires [35] were used to detect PD. In a recent review work Mei et al. [36] studied 209 publications, and nearly 170 publications were focused on the classification of PD from healthy control based on a single or more than one data modality. According to the study, ML-assisted clinical decision systems for the diagnosis of PD have high potential. They also suggested these systems can also be used in the early detection of PD with the adoption of novel biomarkers. In the present study, the main focus was on the literature which has physical and biomarker features and also deals with premotor and motor symptoms to screen and understand the significant symptoms. Prashanth et al. in 2014 [37] observed in their study of prodromal phase characteristics that premotor symptoms such as olfactory loss and sleep disorders predate the occurrence of first clinical features by years or even decades. They diagnosed early PD using support vector machine (SVM) and classification tree methods and yielded high accuracy of 85%. In another study, Berg et al. [38] described research criteria and probability methodology based on age and diagnostic information such as motor and non-motor clinical symptoms, clinical signs, and ancillary diagnostic tests for the diagnosis of prodromal PD using a Bayesian naive classifier with more than 80% certainty in 2015. Fereshtehnejad et al. [39] studied the independence of prodromal markers and concluded that Movement Disorder Society (MDS) prodromal criteria had 81.3% sensitivity for conversion to PD. They have also evaluated the association between prodromal PD likelihood ratios and the time to conversion in the REM sleep behavior disorder (RBD) cohorts. Mahlknecht et al. determined the predictive value of olfactory dysfunction for neurodegenerative disease in cohorts with idiopathic REM sleep behavior disorder (iRBD). Olfactory function, particularly odor identification, is a significant non-motor feature that may help to predict neurodegenerative diseases such as Lewy body disease and PD with the help of other non-motor features like sleep disorder [40].

Non-motor and motor symptoms combined with other modalities of data were also studied to identify PD. For example, in the year 2016, Prashant et al. [41] also studied the combination of non-motor, cerebrospinal fluid (CSF), and striatal binding ratio (SBR) from SPECT imaging to detect preclinical PD. They yielded 96.40% accuracy using the SVM classifier. Multimodal study of clinical information such as motor and non-motor characteristics, biospecimen results, and neuroimaging outputs was used by Zhang et al. [42] for the classification of PD subtypes using long short-term memory (LSTM). They have also found that the progression rate of motor and non-motor symptoms is independent of each other. In another study, Mabrouk et al. studied non-motor and motor symptoms with image-derived features using five ML algorithms and achieved the highest accuracy of 82.2% [43]. Severson et al. [44] have discovered the progression and states of PD using the ML algorithm based on motor and non-motor symptoms. Through the study, the heterogeneous nature of PD has been accounted for especially in its early course. Extreme gradient boosting (XGBoost) and logistic regression techniques were used for the prediction of clinically significant depression and also identified depressive disturbances in Parkinson's disease (dPD) as the most significant determinant of quality of life in 2020, Gu et al. [45]. In another study, Leger et al. [46] performed two binary classifications, one was between early PD versus HC and another was early PD versus SWEDD (scans without evidence of dopamine deficit). General additive (GAM) model yielded 0.92 AUC score for PD versus HC, while performance for PD versus SWEDD was lower for all models used in the study (XGBoost got AUC 0.86). Non-motor clinical and biomarker features are less effective in discriminating early PD from SWEDD as compared to HC from PD. In 2022, Martinez-Eguiluz et al. [47] evaluated 9 ML algorithms for discriminating PD from HC using non-motor clinical PD features and achieved the best performance of SVM with 86.3% accuracy. With the advancement of technology, it can be possible to remotely diagnose and monitor PD. Prince et al. used mobile health data and multisource ensemble learning with convolutional neural networks (CNNs) and achieved an accuracy of 82.0%. However, they have not studied inter-source relationships between various features and data can have noise and irrelevant information [48]. Other than structural data, the study of medical reports using natural language processing (NLP) can also be used to diagnose stages of PD further. Frasca et al. [49] have proposed a technique to categorize the patients using the correlation between the biomedical information of medical reports collected during the visits. Doc2Vec, latent Semantic Analysis, and Text2Vec techniques were used to correlate the data of each patient's medical report. Despite various techniques used to detect PD, there is no standard approach that can online identify the prodromal phase of PD with only motor and non-motor symptoms. Therefore, in this study, we have designed an approach that detects PD with minimum questionnaires and eases the life of the patients. We applied state-of-the-art methods for feature selection and model training to optimize the performance of the proposed approach. Both filter (ANOVA, mutual information (MI), and chi-square) and wrapper methods (sequential feature selection, recursive feature elimination, and Boruta) used separately to analyze their impact on feature selection. Additionally, various classification algorithms including CatBoost, extreme learning machine (ELM), logistic regression (LR), random forest (RF), SVM, and XGBoost were utilized to examine their effectiveness and interactions with the feature selection techniques. This methodology allowed for a thorough investigation into the effects of different algorithms and techniques on each other, providing valuable insights into optimal model performance and feature relevance. These models cover a range of techniques such as ensemble methods (bagging with RF and boosting with XGBoost), probabilistic modeling (LR and SVM can be used for probability estimation), and regression (ELM and LR can perform regression tasks). Furthermore, to enhance the robustness and interpretability of the findings, the results were validated using explainable AI techniques. By employing explainable AI methods, insights into the decision-making process of the model are aimed to be provided, thereby increasing confidence in the reliability and validity of the results.

3 Dataset

PPMI Online is part of the PPMI study that aims to understand the risk and predictive factors for PD. It is an observational study that started on July 28, 2021, to expand the original PPMI study. Original PPMI [12] is an ongoing, longitudinal study started in 2011 that aims to identify clinical, imaging, genetic, and physical features of the patients for the progression of PD. Initially, for this work, data was gathered from de novo-diagnosed PD from 33 participating sites worldwide. PPMI questionnaires are developed in collaboration with multidisciplinary teams of researchers, clinicians, and experts in Parkinson's disease. Questionnaires undergo rigorous validation processes to ensure their reliability, validity, and relevance to the study population. The PPMI clinical protocol involves detailed assessments of motor and non-motor symptoms, cognitive function, and quality of life using standardized scales such as the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [50], Montreal Cognitive Assessment (MoCA) [51], and others. The data was downloaded on 25 May 2022, with 22,953 participants (including both PD and healthy) enrolled for the PPMI Online study. 18,580 participants have visited screening and contributed data longitudinally and cross-sectionally in one or more online meetings using a web-based platform. The inclusion criteria for PDs:

1. Male or female 18 years or above
2. Diagnosis of Parkinson's disease

Same is true for healthy controls in terms of age, and gender, who have not received a diagnosis of Parkinson's disease. PPMI Online study is designed with 10 characteristics of questionnaires each characteristic has one or more study assessments listed in Table 1.

These self-reported questionnaires include demographic information, socioeconomic status, family history related to Parkinson's disease, motor and non-motor aspects of the participants. PPMI Online also includes medical history

Table 1 Details of characteristics and study assessments of PPMI Online Project

Characteristics	Study assessments
Cognition	Cognitive change
Family history-online	Family history of Parkinson’s Disease:1st Degree Relatives
Medical-online	Assessment of constipation (Online)
	Caffeine consumption (Online)
	COVID-19 history (Online)
	Head injuries (Online)
	Health history: annually (Online)
	Health history: quarterly (Online)
	History of falls: baseline (Online)
	History of falls: surveillance (Online)
	Medication history (Online)
	Smoking history (Online)
Motor/MDS-UPDRS-online	MDS-UPDRS Part I:non-motor aspects
	MDS-UPDRS Part II: motor aspects
	Participant motor function questionnaire
Neurobehavioral tests-online	Geriatric depression scale (Online)
	Parkinson anxiety scale (Online)
Neuropsychological tests	Penn Parkinson’s daily activities questionnaire-15
Patient status	Age of Parkinson’s disease diagnosis (Online)
	High interest questions for Non-PD cohort (Online)
	High interest questions for PD cohort (Online)
	Hyposmia 1Qx from remote (Online)
	Parkinson’s disease history: return study visit for Non-PD cohort (Online)
	Parkinson’s disease history: return study visit for PD cohort (Online)
	Participant-visit information (Online)
Sleep disorder tests	Epworth sleepiness scale (Online)
	Parkinson’s disease sleep scale: PDSS-2 (Online)
	PPMI RBD sleep questionnaire (Online)
	RBD1Q: postuma acting out dreams (Online)
Subject enrollment	Participant enrollment status
Subject demographics	Race and ethnicity (Online)
	Registration information (Online)
	Socioeconomic status (Online)

like injury on the head, diagnosis of diabetes, depression, Erectile dysfunction, COVID-19, and habitual behavior such as caffeine consumption and smoking frequency.

3.1 Demographic and physical features

Demographic characteristics, the score of motor experience of daily living (UPDRS-II) [52], and the geriatric depression score (GDS) [53] of both groups (PD and HC) are reported in Table 2. PPMI Online is an online self-reported dataset, hence to validate the characteristics of this dataset it is necessary to compare them with other recent research studies of PD. PPMI dataset is also accessible with PPMI Online dataset which is a longitudinal observational research study assessed with in-person visits of cohorts. For statistical characteristics comparison, 197 HC and 490 PD cohorts were considered from the PPMI dataset. PPMI Online PD patients (67.41 years) were significantly older than PPMI PD patients (62 years). While HC is older for PPMI (61.30 years) as of PPMI Online (56.68). Healthy Female percentage in PPMI Online (63.37%) was higher as compared to PPMI (34.52%) while the proportion of PD Females is comparable for both datasets (39.48 and 35.61%). In the literature [11], Fox Insight (FI) study also collected online data for PD cohorts only. In the present experiment, Tables 1 and 3 from the literature [11] are also compared for the reported datasets with PPMI Online dataset. The FI dataset has a comparable mean age of 65.77 years with PPMI Online PD cohorts. Both datasets lack equalities of gender, racial, and ethnic diversity for PD

Table 2 Demographic and other characteristics of PPMI Online Dataset

Project	PPMI Online		PPMI		FI
Features	PD	HC	PD	HC	PD
Age (years)	67.41 (9.09)	56.68 (13.55)	62 (9.8)	61.3 (11.1)	65.77 (9.49)
Ranges	(23–97)	(18–91)	(29–85)	(30–83)	(18–98)
Male (%)	60.52	36.63	64.39	65.48	54.44
Female (%)	39.48	63.37	35.61	34.52	45.56
UPDRS-II Score	11.23 (7.99)	1.69 (3.39)	5.9(4.3)	NA	12.11 (8.26)
Ranges	(0–52)	(0–24)	(0–52)		(0–52)
GDS	4.17 (3.64)	1.14 (3.27)	2.5(2.5)	1.3 (2.2)	NA
Family History (1st degree) (%)	17.06	53.89	13.09	48.02	22.32
Education (years)	16.45 (3.73)	16.71(3.74)	15.4(3.2)	16.3 (2.8)	NA
Race					
White	96.24	96.69	92.42	91.13	96.85
Non-White	03.76	03.31	7.58	8.87	03.15
Hispanic or Latino	03.50	03.82	2.13	3.28	3.81
Non-Hispanic	96.50	96.18	97.87	96.72	96.19

Table 3 Percentage of PD cohort with difficulty assesses by UPDRS-II

Project	PPMI Online	PPMI	FI
UPDRS Part II Patient Questionnaire: motor aspects of experiences of daily living (M-EDL)	PD cohorts with difficulty (%)		
Speech	57.55	33.89	58.48
Saliva and drooling	48.95	36.49	51.36
Chewing and swallowing	37.36	13.03	35.43
Eating tasks	45.68	28.20	48.01
Dressing	57.75	40.52	60.76
Hygiene	42.61	25.59	43.24
Handwriting	69.41	61.85	70.82
Doing hobbies and other activities	64.99	36.73	62.93
Turning in bed	50.63	25.59	62.76
Tremor	51.98	86.02	79.71
Getting out of bed, car or deep chair	54.59	38.63	71.13
Walking and balance	59.53	36.73	69.57
Freezing	27.56	04.74	31.28

cohorts. It is well established that characteristics of PD also depend upon gender, race, and ethnicity; hence, it is important to consider all these aspects into account while designing a medical application approach to avoid discriminatory outcomes and sub-optimal results [54]. In PPMI Online, the proportion of healthy females is significantly higher, while the proportion of females suffering from PD is still under-represented. Although right now, whites are dominating this study, perhaps in the near future PPMI Online aims to expand internationally [55].

3.2 Motor aspects of experiences of daily living

UPDRS-II has been calculated by questionnaire using the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale-Part-II (UPDRS-II) which is used to assess the relationship between quality of life and difficulties of PD patients. Mean UPDRS-II score for PPMI Online was 11.23 calculated for PD cohort, FI reported 12.11 and for PPMI score was 5.9 only. Apart from the cumulative UPDRS-II score, the proportion (reporting greater than 0) of each characteristic from Motor Aspects of Experiences of Daily Living for PD cohort is reported in Table 3. In the questionnaire, 58.48% of the PD cohort from the FI study and 57.55% PD cohort from PPMI Online accepted problem in the speech and hence this feature is comparable for both datasets. However, the PD cohort from FI faced a significantly higher struggle than the PD cohort from PPMI Online dataset. For example, 62.67% of PD cohort from the FI study experienced difficulty turning in bed while only 50.63% were from PPMI Online dataset. Other motor aspects of experiences of daily living such as Tremors (79.71%), Getting Out of Bed, Car, or Deep Chair (71.13%), and Walking and Balance (69.57%) are also more difficult for the PD cohort from FI study, as reported in Table 3. PD cohorts from PPMI dataset faced tremors very frequently; nearly 86.02% of participants had tremors. Problems in Chewing and Swallowing (13.07% vs 37.36%) and Freezing (4.47% vs 27.56%) are rare in the participants of PPMI dataset as compared to PPMI Online dataset. Additionally, 35.43% and 31.28% of participants of the FI dataset have difficulty in Chewing and Swallowing and Freezing, respectively, which is similar to the PPMI Online dataset.

3.3 Cognitive instrumental activities of daily living

Penn Parkinson’s Daily Activities Questionnaire-15 [56] is also asked to assess the cognitive instrumental activities of daily living in PPMI Online study. Features were scored based on PD patient’s difficulty in performing Daily Activities on the following scale: “none,” “a little,” “somewhat,” “a lot,” and “cannot do.” Each characteristic is scored 4–0 (total score range = 0–60) with higher scores indicating better Daily Activities functions. In Table 4, percentage of PD and HC cohort with difficulty (cumulative proportion for 3–0, as score 4 represents no difficulty) was reported for each characteristic. Problem in cognitive instrumental activities of daily living is more in PD patients as compared to HC. No other referenced comparator cohort had a Penn Parkinson’s Daily Activities Questionnaire-15 at the time of data download.

4 Methodology

In this study, an approach of e-diagnosis to detect PD was proposed (Fig. 1). The questionnaire data from the PPMI Online dataset were preprocessed to obtain features relevant to the context of PD. Multiple FS techniques were then applied to obtain the most significant attributes. Furthermore, multiple classifiers were trained and evaluated on these selected features. The machine-made decisions to select optimal subsets of features and ML algorithms were validated using the interpretations provided by SHAP.

4.1 Preprocessing of data

After unifying the questionnaire, data were preprocessed by dropping features that have more than 70% missing values like hypertension or high blood pressure diagnosis, age of high blood pressure diagnosis, and head injury during life, from this study. A total of 120 features including cohort (PD or not PD) have been considered from the different questionnaires of online screening and online meeting-1 (*OL01*) in the current study. A questionnaire is related to the return study of the participants, in that participants have to reconfirm their current status of PD or Non-PD also available in the data. The response to this questionnaire, 43 previously claimed PD patients have been either not sure or not diagnosed with PD, while 75 healthy controls have been diagnosed with PD since enrollment. Data of all these individuals were removed from this experiment to avoid any confusion. Hence, the total number of samples is reduced to 18,462 (HC = 12,811, PD = 5651). The missing values of the remaining features were imputed using the K-nearest neighbor (KNN) imputation algorithm for both classes separately after that data were rescaled from standard scaling to avoid the effect of different scales in the data. A comparison of datasets with imputation and without imputation trained on RF is reported in Appendix 1.

Data were divided into two parts—90% of data was used for feature selection, training, and validation and 10% of the data was kept safe to evaluate the model. Training dataset was split into ten folds, of which nine were used for training the models and one for validation.

In the online meeting-2 (OL02), questionnaires related to smoking and Epworth Sleepiness Scale were also included with basic questionnaires including cognitive change, motor, and non-motor features. The number of participants reduced in the second online visit and approximately 8400 participants responded. To establish the impact of smoking and the Epworth Sleepiness Scale on PD patients, a separate dataset has been created by including these two features with the previous dataset on *Patient Id*. The number of participants was 7403 after removing missing value instances. The study of this data is reported in Appendix 2.

To bring transparency to the ML model’s decision without affecting model efficiency, Shapley Additive Explanations (SHAP) algorithm [57] has been used. SHAP algorithm works on Shapley values which have a strong mathematical foundation. It explains the impact of each feature on the output variable, helps to estimate the significant features, and clarifies the decision made by the algorithm using the XGBoost model.

Table 4 Percentage of participants with problems in daily cognitive activities

Penn Parkinson’s daily activities questionnaire-15	PD cohorts with problem (%)	HC cohorts with problem (%)
Difficulty reading the newspaper or magazine	28.42	13.84
Difficulty keeping track of the time	16.15	8.34
Difficulty counting money	12.83	5.05
Difficulty reading complex instructions	34.65	18.55
Difficulty handling an unfamiliar problem	36.91	26.81
Difficulty explaining how to do something involving several steps to another person	47.64	23.31
Difficulty remembering a list of 4 or 5 errands without writing it down	74.41	67.30
Difficulty using a map to tell where to go	26.12	21.29
Difficulty remembering new information like phone numbers or simple instructions	66.10	59.13
Difficulty doing more than one thing at a time	64.82	46.03
Difficulty learning to use new gadgets or machines around the house	50.14	37.18
Difficulty understanding your personal financial affairs	25.29	19.39
Difficulty maintaining or completing a train of thought	57.60	42.99
Difficulty discussing a TV show, book, movie, or current events	42.08	19.09
Difficulty remembering what day and month it is	29.71	17.43

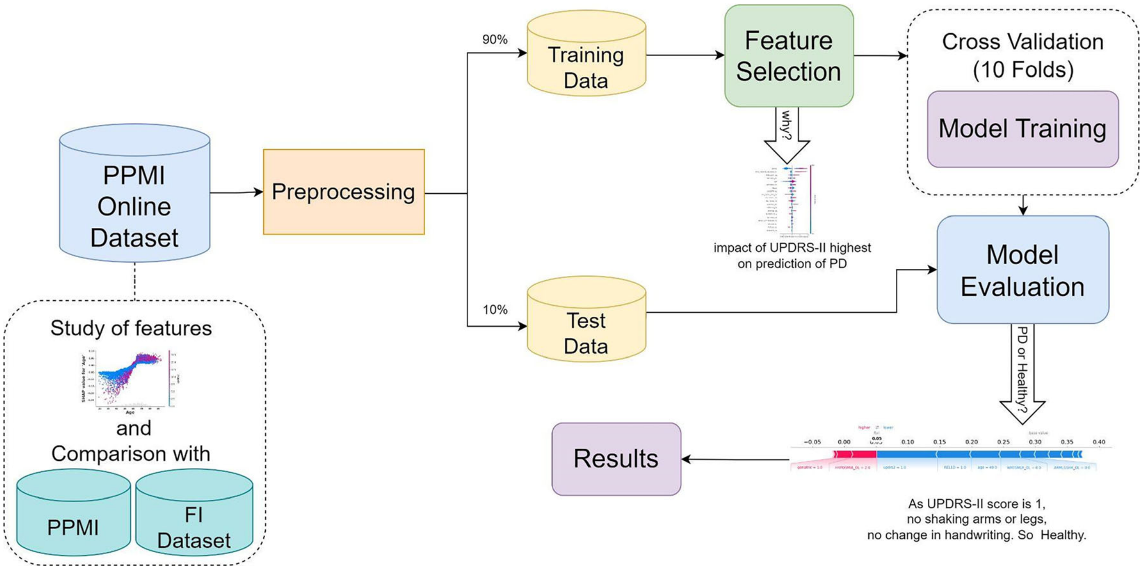
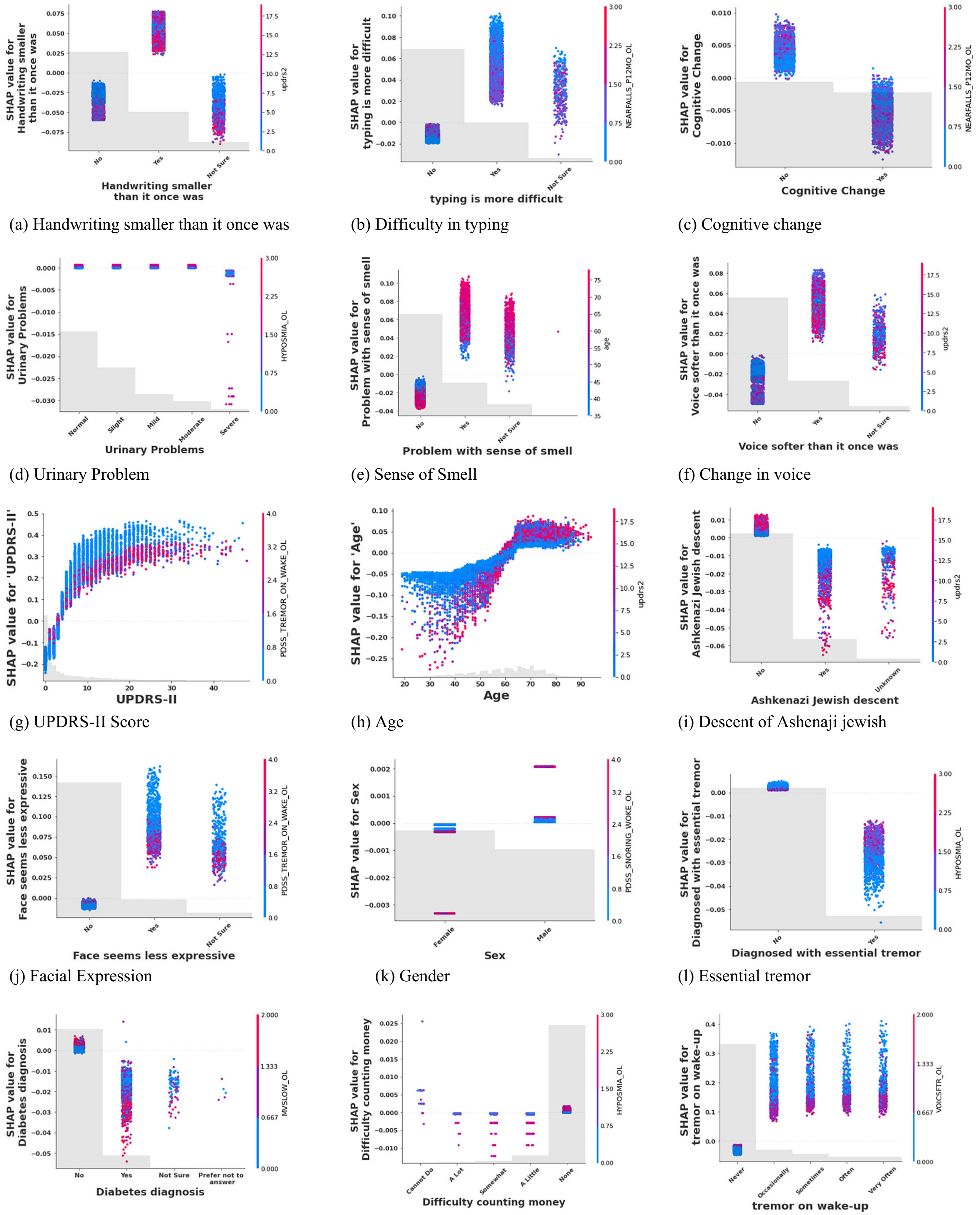


Fig. 1 Proposed approach to detect Parkinson’s disease

4.2 Interpretability of features with SHAP values

The relationship between feature and SHAP values is also visualized using scatter plots in Fig. 2. Randomly 15 features are selected to understand the behavior of features. SHAP value represents the impact of the feature’s value on the output of the model for that individual instance’s prediction. In our study, the units are logarithms of the odds of prediction of PD. The SHAP module also involves another feature that the plotted feature interacts most with automatically and if the interaction effect is present between two features, it is well visible with a vertical pattern of coloring. The light gray area in the background is a histogram that shows the distribution of values on the plot. Each dot represents an individual participant (row). In Fig. 2a, participants who have a change in handwriting have more probability of having PD as compared to those who do not have any change. Hence, this feature can be a



(a) Handwriting smaller than it once was
(b) Difficulty in typing
(c) Cognitive change

(d) Urinary Problem
(e) Sense of Smell
(f) Change in voice

(g) UPDRS-II Score
(h) Age
(i) Descent of Ashenaji jewish

(j) Facial Expression
(k) Gender
(l) Essential tremor

(m) Diabetes
(n) Difficulty in counting money
(o) tremor on wake-up

Fig. 2 Scatter plot of features and their SHAP values

significant feature because it has clearly distinguished SHAP values for the diagnosis of PD. In Fig. 2b for most of the participants, typing is not at all difficult; however, if the typing is difficult, it can contribute to predicting PD. In Fig. 2d, if a person has a severe urinary problem and also has some difficulty in the sense of smell, then this interaction helps in the prediction of PD. In Fig. 2e, it is visible that the problem in sense of smell is more in the higher age. Figure 2f shows that difficulty in speech (UPDRS-II score) and softness of voice are correlated. The value of UPDRS-II score significantly helps in the prediction of PD (Fig. 2g). In Fig. 2h, older participants with high UPDRS-II value have more chances to suffer from PD as compared to young participants with lower scores. Females very often snore have less chance to have PD, while males with the same symptom have higher chances (Fig. 2k). Tremor on wake-up is also a contributing factor for predicting PD shown in Fig. 2o.

4.3 Feature selection techniques

It is not necessary that each feature in the dataset contains some relevant information. Feature selection techniques were used in the current study to understand the significance of features and distinguish the important features from the dataset. Using feature selection techniques, it is easy to obtain and interpret information from selected features. In addition, this avoids overfitting of ML algorithms, and requirement of computational resources is also reduced.

Filter methods such as ANOVA [58], chi-square [59], and MI [60] have been used to assess the statistical significance of the features. ANOVA is a statistical test that computes the variance among groups. The variance of groups can be explained or unexplained. The features independent of the target variable are removed from the dataset. This test is used when input variables are continuous and the output variable is categorical in a classification problem. Mutual information is the measure of the mutual dependence between the dependent and the independent variable. If the two random variables are independent, the value of MI is equal to zero. Chi-square test is an interdependence test to measure the dependence between categorical feature variables and a categorical target variable. These methods are fast and their computational cost is low even so it is not necessary they suggest a stable subset because the interaction between the features and with classifiers are not considered by these methods.

In wrapper methods, ML algorithms are used to search optimized feature subsets using all possible combinations of features based on the evaluation condition. Wrapper methods identify an optimal subset of features because of their exhaustive nature. In the present study, three wrapper methods are used namely Boruta [61], Sequential Forward Selection (SFS), and Recursive Feature Elimination (RFE). Boruta Algorithm is also a statistically grounded technique and works well without any user-defined threshold. It defines the threshold by adding shadow features to the dataset. Then, RF classifier trains on this extended dataset to get important features. Although it can handle interactions between variables yet it does not deal with highly correlated features. SFS is a greedy search algorithm started with a null feature set and seeks the optimal feature subset by iteratively picking features based on the classifier performance. It is observed that by adding new features, previously selected features may become obsolete. RFE is also a greedy search algorithm that takes a full set of features and iteratively removes the worst-performing feature. It then ranks the features based on the sequence of their elimination. The main advantage of this algorithm is that it keeps only those features that are most relevant in predicting the target variable.

4.4 Machine learning algorithms

ML and DL models have significant results for the automatic diagnosis of diseases. In this experiment, six ML algorithms namely SVM, RF, LR, XGBoost, CatBoost, and ELM were used to analyze the PPMI Online dataset. Prediction of PD and healthy control is a binary classification problem. SVM, LR, ELM, and RF were implemented from the sklearn library in Python. XGBoost and CatBoost were implanted from their respective official sites.

Random forest is one of the most powerful ML algorithms. It has its roots in ensemble machine learning algorithms where many classifiers have been parallelly trained on the randomly selected subset of features (bootstrapped samples). In the final step, the decision was made based on the decision of all the classifiers by majority voting or by the applying average rule. The pooling of base classifiers compensates for the weakness of individual base classifiers and most of the time results are accurate. The random forest has decision trees as base classifiers which are grown to the maximum of their size and unpruned. Ideally, all these decision trees should be less correlated to produce an accurate classification result.

SVM is based on statistical learning algorithms used for both classification and regression. In classification, the main goal of this algorithm is to design a higher dimensional hyperplane that can segregate data between correct categories. The dimension of the hyperplane depends upon the number of vectors/features. SVM is effective in high-dimensional space and also memory efficient.

Logistic regression is also borrowed from the field of statistics that estimates the probability of an event using independent variables of data. It is a simple and efficient algorithm for binary classification problems; however, it can be extended into multiple classifications. It achieves very good results on the data which have linearly separable classes. LR also performs extremely robustly and accurately for datasets that have nonlinear relationships between independent and dependent variables.

Extreme learning machine is a feedforward network that uses Moore–Penrose generalized inverse to set its weights instead of gradient-based backpropagation. ELM is much faster than previously designed networks because the backpropagation technique makes the learning process slower. Research revealed ELM outperformed classical machine learning algorithms in training time and accuracy. It has many application domains such as medicine, robotics, chemistry, economy, and transportation [62].

Both XGBoost and CatBoost are gradient-boosting algorithms. Gradient boosting is an ensemble learning algorithm that is designed to generate the prediction through a set of interconnected several classifiers. In XGBoost and CatBoost, classifiers are arranged in a sequential manner where each next classifier learns from the previous classifier's errors. Random forest is also an ensemble learning algorithm where base classifiers are designed in parallel and their decisions are independent of each other. However, with the gradient boosting algorithm each new classifier makes the decision based on the predecessor classifier. CatBoost classifier is specially designed for categorical features while in the case of XGBoost it is necessary to encode them before training.

All classifiers except ELM are optimized using the Grid Search optimization technique. In this experiment, tenfold cross-validation has been used for the evaluation of trained machine-learning algorithms. In tenfold cross-validation, trained data was split into 10 parts, where 9 parts were used for model training and one part was for testing the trained model. Mean accuracy of all the folds has been calculated for the evaluation of models. Metrics from different quadrants of the confusion matrix like Precision, specificity, F1-Score, and sensitivity are also reported. It is important to consider the True positive rate in the medical dataset, hence sensitivity is a crucial metric in this study. ROC-AUC score is also calculated to check that models are learning from both classes. We have reported all these metrics on the test dataset (10% of the entire dataset) also which kept separated after preprocessing.

5 Results and discussion

5.1 Result with preprocessed full features dataset

Preprocessed dataset with full features used to train machine learning algorithms SVM RF, LR, XGBoost, CatBoost, and ELM. In Table 5, highest mean accuracy was achieved with the CatBoost classifier 96.73 using the tenfold cross-validation technique on the training dataset. Although the dataset has class imbalance, still most of the classifiers have learned from both classes and got AUC scores ranging from 89.27 to 94.91%. Result is appealing yet it is very difficult to collect nearly 115 attributes from individual participants, and hence, feature selection has been performed.

5.2 Result with filter method

5.2.1 ANOVA

To select the best k features from ANOVA feature selection method, a loop has been designed that ranges from 1 to all features of the dataset. After each iteration, the model was trained with the resulting feature subset and evaluated on a previously separated test dataset. Through this experiment, the best accuracy has been achieved at 93.11% with 52 features by random forest classifier. Therefore, performance metrics were calculated on this 52 features subset, reported in Table 6 using SVM RF, LR, XGBoost, CatBoost, and ELM Classifiers. CatBoost classifier outperforms all other classifiers and achieved the highest mean accuracy of 94.22% (shown in bold in Table 6).

To understand the importance of selected features, Fig. 3 is plotted for the features selected by ANOVA. Feature *updrs2* (UPDRS-II) is the most significant feature and is directly proportional to the experience of difficulties increasing in daily living. The second most significant feature *DFCLTYTYPE_OL* (more difficulty typing, using mouse, or touchscreen) is also related to difficult experiences in ordinary tasks and has been asked in the Motor Function Questionnaire. The third most important feature *PDSS_TREMOR_ON_WAKE_OL* (experience tremor on wake-up) comes from the sleep study of the participant which asked questions about sleep quality.

Absolute SHAP values are also plotted for the selected features set in Fig. 4, which shows the importance of a

Table 5 Mean performance metrics with full feature dataset

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	**96.73 ± 0.56**	**95.86**	**93.69**	**94.64**	**0.94**
ELM	95.48 ± 0.58	94.15	92.54	94.93	0.94
LR	94.34 ± 0.51	92.58	90.13	93.92	0.91
RF	95.42 ± 0.68	93.16	92.37	94.91	0.93
SVM	94.16 ± 0.72	92.42	89.85	93.29	0.91
XGBoost	91.72 ± 0.10	89.59	88.53	89.27	0.89

Table 6 Mean performance metrics with feature selected by ANOVA

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	**94.22 ± 0.51**	**93.19**	**93.19**	**93.43**	**0.93**
ELM	93.28 ± 0.63	91.93	92.28	92.89	0.92
LR	91.13 ± 0.53	90.31	89.46	91.12	0.89
RF	93.47 ± 0.65	91.95	92.29	92.87	0.92
SVM	91.53 ± 0.47	90.63	89.14	90.84	0.89
XGBoost	89.65 ± 0.80	87.52	88.45	89.31	0.87

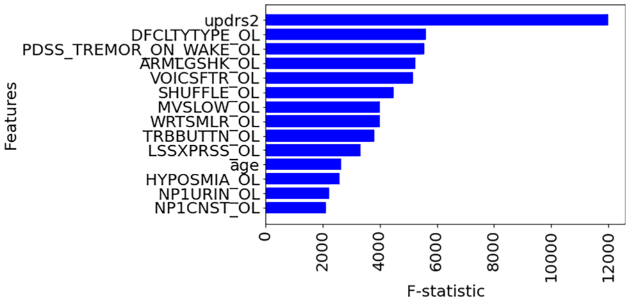


Fig. 3 *F*-statistic of top ranked-14 features by ANOVA

feature in predicting the Parkinson's. The mean absolute effect of features represented here, however, does not explain in a positive or negative way. Again UPDRS-II is the most significant feature, while in spite of the *DFCLTYTYPE_OL* feature, *PDSS_TREMOR_ON_WAKE_OL* feature becomes the second most important feature according to SHAP values. In the last bar, cumulative importance of less important features (38 features) is plotted, and their collective contributions are even less as compared to the most important feature (UPDRS-II). Hence, it may be possible that features selected by ANOVA are not the optimized feature set.

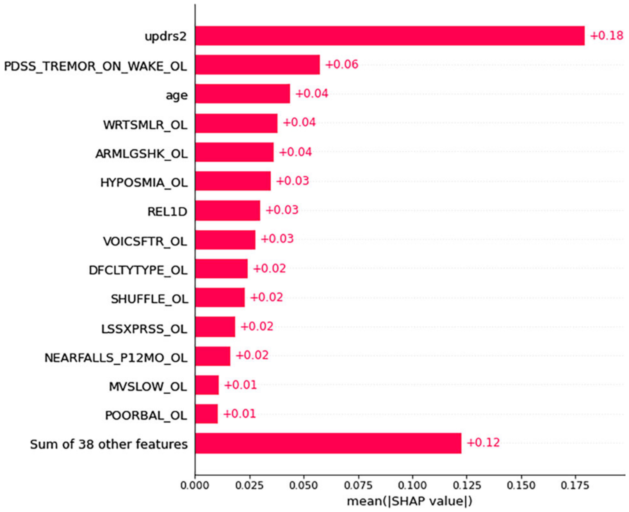


Fig. 4 Mean SHAP value of features selected by ANOVA

5.2.2 Mutual information

Same experiment was conducted with mutual information and achieved 93.28% accuracy with 64 top-ranked features on the test dataset reported in Table 7. Hence, 64 features were selected to train the various classifiers. Highest mean accuracy achieved through tenfold was 94.63 with RF classifier. F1 scores (0.93) also suggested both CatBoost and RF have classified each instance whether it belongs to HC or PD class accurately (the highest values are shown in bold).

Weight of the selected features by MI algorithm is given in Fig. 5. (For visual convenience only best 14 features are plotted.) MI has also ranked UPDRS-II as the most significant feature, while it considered *WRTSMLR_OL* (smaller handwriting from previous) as the second most important feature. Again CatBoost classifier has performed better than all other classifiers and has yielded 94.43% accuracy. Both ELM and RF classifiers have performed similarly and achieved nearly 93% accuracy. Although MI has selected more features than ANOVA, still it does not increase the accuracy. Both ANOVA and MI have selected almost the same features; however, feature ranking is different even so achieve similar accuracy as shown in Fig. 6.

5.2.3 Chi-square test

Chi-square has suggested 101 features and achieved the highest accuracy of 92.69% using Random Forest Classifier on the test dataset. Chi-square does not select features properly as shown in Fig. 6, as with 20 features it achieved accuracies below 80%. From all these 101 features, SVM RF, LR, XGBoost, CatBoost, and ELM Classifiers have been trained, and mean accuracy and other metrics are reported in Table 8. Despite a large number of features, any of the models did not learn much and hence it can be concluded that features selected by chi-square algorithms

Table 7 Mean performance metrics with feature selected by MI

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	94.43 ± 0.47	93.44	93.42	93.56	0.93
ELM	93.70 ± 0.60	92.44	92.76	92.95	0.92
LR	91.84 ± 0.60	91.10	89.41	91.41	0.90
RF	**94.63 ± 0.57**	**93.36**	**93.66**	**93.98**	**0.93**
SVM	91.89 ± 0.52	91.05	89.61	91.06	0.90
XGBoost	89.77 ± 0.89	87.58	88.78	88.48	0.88

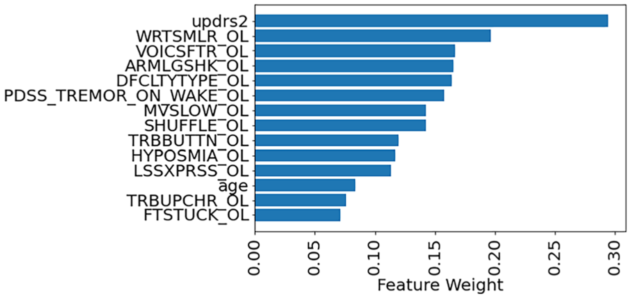


Fig. 5 Weight of selected features by MI algorithm

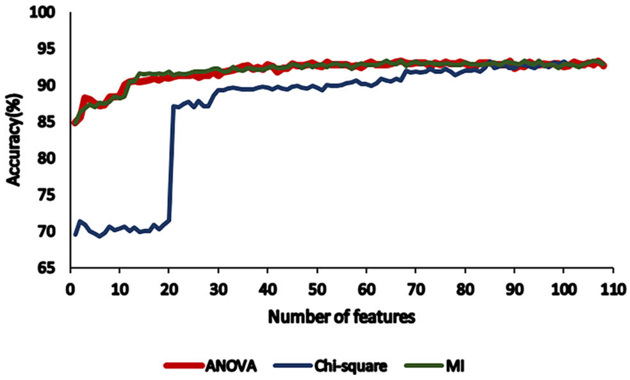


Fig. 6 Accuracy (%) versus number of features for filter methods

are less relevant to predicting the Parkinson's. It can be concluded that CatBoost has performed better with a large number of feature subsets (shown in bold) and yielded better results than other classifiers.

5.3 Result with wrapper method

Filter methods are fast and easy to use even so they may or may not provide an optimized feature subset which dilutes the objective of this study. Hence, wrapper methods are also used to get the best feature subset with a minimum no. of features.

5.3.1 Boruta (RF)

Boruta identified 20 significant features using RF classifier. This algorithm has the ability to capture all features that are relevant to the target variable without suggesting any threshold value. All the above-mentioned six classifiers were trained using the selected feature subset, and results are reported in Table 9. From the result, it is clearly visible that there is no significant difference in the result as compared to the results of filter methods except for RF classifiers. However, the number of features is reduced to almost one-fourth of the features selected by the filter method without any decline in mean accuracy. RF has outperformed all other classifiers and achieved 95.19% mean accuracy on the training dataset (shown in bold in Table 9).

SHAP model was trained to understand the importance of features. Figure 7 shows the positive and negative relationship of the feature with the target variable (PD or Not). The *x*-axis represents the Shapley value and *y*-axis represents features. Features are sorted by the mean SHAP value, i.e., most significant features are at the top. All the observations are visible with respective feature values. The values of the feature are represented by the color of the point. The UPDRS-II score is the most significant feature and has a large distribution of the Shapley values while features like *TRBBUTTN_OL* (trouble in buttoning buttons) and *FTSTUCK_OL* (Feet seem to get stuck to the floor) have almost zero impact on predicting PD. The lower value of UPDRS-II score has a negative relationship with PD and higher score has a positive relationship. An ordinal feature like *PDSS_TREMOR_ON_WAKE_OL* (tremor on wake-up) if the value is zero has a negative impact; however, as the frequency of tremor increases, impact on model output also increases. From Fig. 7, it can be concluded that a few features from the subset selected by Boruta have very small impact on the model's output. Hence, more optimized feature subsets need to be searched using some exhaustive search methods.

Table 8 Mean performance metrics with feature selected by chi-square

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	**94.87 ± 0.50**	**94.02**	**93.87**	**94.67**	**0.93**
ELM	93.51 ± 0.62	92.21	92.53	93.99	0.92
LR	92.34 ± 0.52	91.59	90.13	92.94	0.90
RF	93.49 ± 0.68	92.25	92.47	93.94	0.91
SVM	92.17 ± 0.69	91.43	89.87	92.29	0.90
XGBoost	89.81 ± 0.88	87.65	88.73	88.37	0.88

Table 9 Mean performance metrics with feature selected by Boruta

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	94.61 ± 0.40	93.40	93.56	94.01	0.93
ELM	93.80 ± 0.56	92.29	92.85	93.55	0.92
LR	91.77 ± 0.46	89.71	90.20	91.15	0.89
RF	**95.19 ± 0.54**	**94.38**	**94.07**	**94.56**	**0.94**
SVM	91.59 ± 0.46	89.39	88.16	91.29	0.89
XGBoost	90.74 ± 0.89	88.57	88.69	90.32	0.88

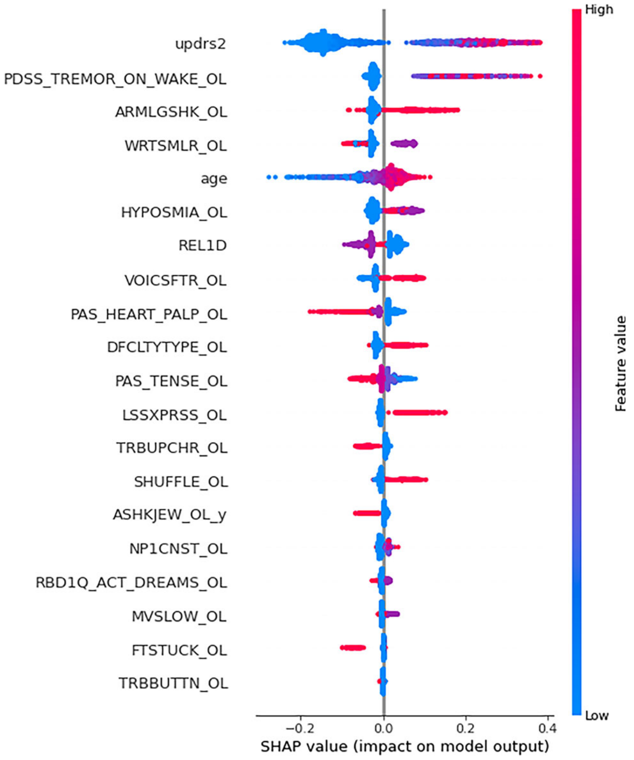


Fig. 7 Impact of feature value on prediction of PD for training dataset

5.3.2 Sequential feature selection (SFS)

The same experiment has been designed for SFS as designed for ANOVA and MI methods to select the *k* number of features. Since wrapper methods perform a greedy search approach to get the best feature subset, they are slow compared to the Filter method. However, the results are very promising so the computational cost can be neglected. SFS recommended 12 features, and the mean accuracy on this subset was similar to the full feature set in Table 10 (highest values are shown in bold).

Almost all the features picked by wrapper methods are the same. Both RFE and SFS suggested motor aspects of daily living are the most important symptom**.** SHAP algorithm is also used to predict PD with the SFS identified feature set with the test set. A few instances were taken as examples, the probability of PD is calculated as shown in Fig. 8, the base value is the average of the model output over the dataset, and for this study, it is 0.30. Base value implies that if for any instance all the feature values are missing it will give the prediction as 0.30. Feature values are mentioned below the arrow and $f(x)$ is the predicted value for the current instance. Pink arrows represent features that proposed the model toward PD and blue represents healthy. The size of the arrow represents the impact of that feature. Considering the decision threshold as 0.50 represents values equal to or above 0.50 was classified as PD and below 0.5 as healthy. In Fig. 8a, one instance (called A for convenience) has been tested on the SHAP model, output has correctly predicted PD by model, and the value of UPDRS-II score has the highest impact and is supported by other symptoms like tremors on wake-up, shaking of Arm or Legs with old age on predicting PD. In another example, it is a real negative case, person (B) is healthy and correctly predicted by the model shown in Fig. 8b. Here UPDRS-II score is very low, which implies no difficulty in daily life experience. Although the participant has a First-degree relative diagnosed with PD, still with other features interaction it contributes to a negative impact on prediction.

Table 10 Mean performance metrics with feature selected by SFS

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	94.57 ± 0.61	93.99	93.18	93.58	0.92
ELM	93.96 ± 0.59	93.24	93.54	92.69	0.92
LR	92.53 ± 0.61	91.65	91.18	92.12	0.91
RF	**94.92 ± 0.55**	**94.16**	**93.53**	**93.76**	**0.93**
SVM	92.37 ± 0.58	92.16	91.33	91.46	0.90
XGBoost	92.26 ± 0.98	92.09	91.89	91.02	0.90

For better understanding, it is crucial to know the false negative and positive cases also. So that it can analyze why the model is predicting wrong for some scenarios. In the first example, in Fig. 9a, a younger participant (C) has a high value of UPDRS-II score and sometimes his heart beats fast. The model intends to classify it as negative (healthy); however, it was positive. In the second example, in Fig. 9b, an older participant (D) did not have a high UPDRS-II score intended as a PD patient; however, it was negative. Other factors such as tremors on waking up and getting up at night to pass the urine could explain the diagnosis of PD or not. One more instance has been taken named as E. He was a middle-aged participant, who had slight difficulty in the daily experiences of life and reported shaking of arms or legs also yet the model has induced E to be healthy as shown in Fig. 9c.

5.3.3 Recursive feature elimination (RFE)

One more wrapper method is used to select the feature subset. RFE has achieved 95.48% accuracy only with 14 features in Table 11 (highest results are shown in bold). The most important symptoms are the motor aspect of experience in daily living, experiencing tremors on walking. From the experiment, it came to be noticed that Handwriting smaller than before can be a significant symptom of PD. Relation between selected features and with cohort was visualized using a heatmap in Appendix 3.

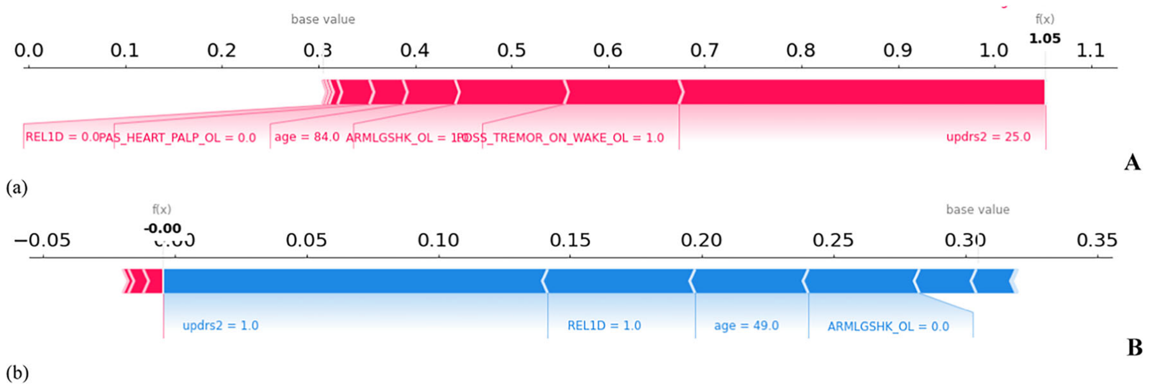


Fig. 8 Correct prediction of real positive (**a**) and real negative (**b**) instances (A and B, respectively) by SHAP model train with SFS feature subset

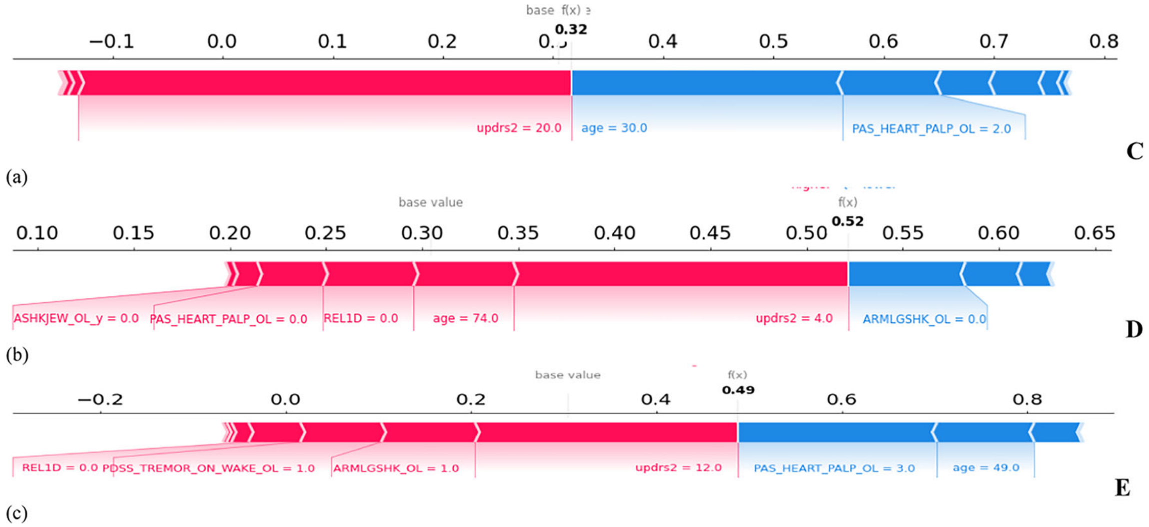


Fig. 9 Incorrect prediction of real positive (**a**, **c**) and real negative (**b**) instances (C, D, and E, respectively) by SHAP model train with SFS feature subset

Table 11 Mean performance metrics with feature selected by RFE

Model	Accuracy	Precision	Recall	AUC	F1-score
CatBoost	94.86 ± 0.76	93.51	93.71	93.62	0.90
ELM	94.38 ± 0.53	93.84	91.31	93.21	0.91
LR	92.02 ± 0.75	91.93	87.13	90.78	0.87
RF	**95.48 ± 0.65**	**94.96**	**93.41**	**93.22**	**0.93**
SVM	92.06 ± 0.46	91.89	89.33	90.01	0.98
XGBoost	91.72 ± 0.90	90.59	88.55	90.28	0.88

SHAP model was trained with the features suggested by the RFE algorithm. Same instances (which were previously taken for SFS) were taken to evaluate the effect of the new features on the models' output. Instance A and Instance B correctly predicted PD and health, respectively, shown in Fig. 10. Instance A, in Fig. 10a, has high UPDRS score which can be considered a reason to make decision as PD. This decision is supported by tremor and shaking hands and legs. While B has an UPDRS-II score of only 1 and reported no shakes in arms or legs and no changes in handwriting. However, B may feel the loss of smell and for this reason, the impact of *HYPOSAMIA_OL* was on the right side of the x-axis in Fig. 10b. Nonetheless, the model has predicted accurately B as healthy.

Again check false negative (C) and false positive (D) instances of SFS featured model for this RFE featured trained model. Here, Instance C was shifted more toward PD as shown in Fig. 11a, however, showing the wrong prediction as healthy. It is a real positive case, while for instance D (Fig. 11b) prediction as healthy is correct despite the age (74 years) of the participant and score on the Geriatric Depression Scale which is more than 5. Other factors such as no change in handwriting or voice explain the decision as healthy. When considering E, the

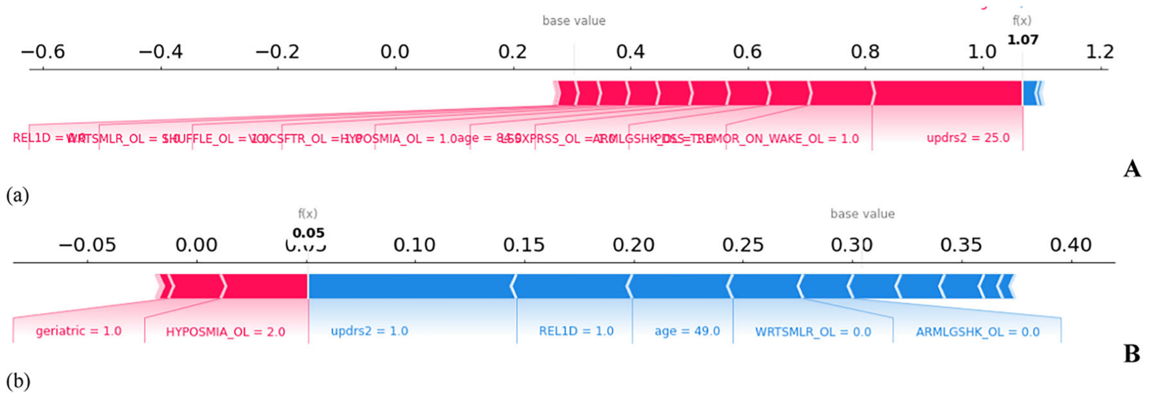


Fig. 10 Correct prediction of real positive (**a**) and real negative (**b**) instances (A and B, respectively) by SHAP model train with RFE feature subset

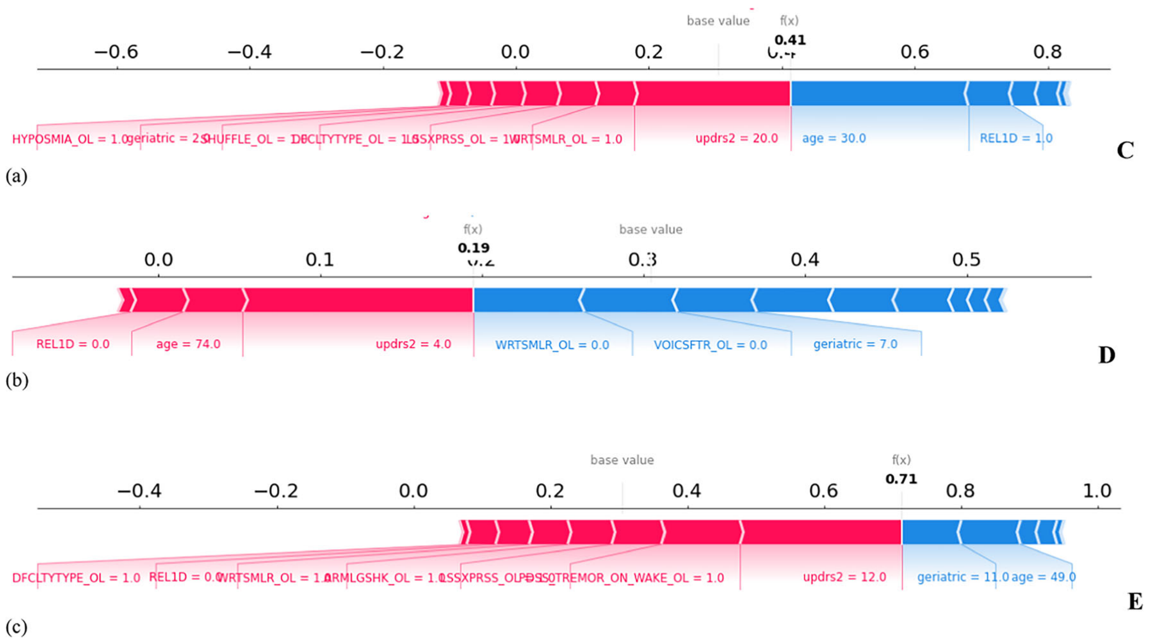


Fig. 11 Incorrect prediction of real positive (**a**) instance (C) by SHAP model trained with REF feature subset. Correct prediction of real negative (**b**) and Real positive (**c**) instances (D and E, respectively) by SHAP model trained with REF feature subset

participant has difficulty in typing, change in voice, small handwriting, shaking of arms or legs, and reported tremors on waking up. Therefore, the prediction of PD is correct. He is a real positive case, in Fig. 11c.

Hence, the features suggested by RFE algorithms are more informative as compared to all other algorithms used in the present study. The decision-making ability increased with the symptoms such as a softer voice than the previous, smaller handwriting than previous, difficulty in typing, and the problem in sense of smell.

ANOVA, MI, and chi-square (filter methods) efficiently identified features with strong statistical relationships with the target variable (PD or HC), along with helping in the reduction of dimensionality and removal of noise from the dataset for the present study. SFS, RFE, and Boruta (wrapper method) assessed feature subsets iteratively, selecting the most relevant features by evaluating their impact on model performance (classification accuracy). By comparing the results of these feature selection techniques and the full dataset (with 120 attributes), robust selection of features contributing optimally to model classification efficiency was ensured. For instance, accuracy with the Full Feature Dataset ranged from 91.27 to 96.73%. Similarly, ANOVA-selected 52 features dataset achieved similar accuracy ranges. MI algorithm selected 64 features and got 94.63% accuracy, while chi-square produced results with a larger number (101) of attribute sets. Wrapper methods selected smaller feature sets compared to filter methods; for example, Boruta identified 20 significant features with 95.19% accuracy. UPDRS-II score is the most significant feature for most algorithms. SFS recommended 12 features with mean accuracy similar to the full feature set, while RFE achieved 95.48% accuracy with only 14 features. Filter methods helped identify statistically linked features, and wrapper methods fine-tuned these traits using machine learning to create optimized sets for better classification by these models. It is concluded from the results that RF has outperformed all the classifiers for almost every feature subset. Therefore, the RF classifier was used to evaluate test data. Results with all the feature subsets are reported in Table 12. RFE and Boruta have yielded accuracy above 94%. MI and SFS have similar accuracy although there was a significant difference in the number of features in both subsets (values in bold are the highest for the parameter). The ROC-AUC curve is also plotted to understand the performance of RF on different feature subsets. It is visible in Fig. 12 that the RF classifier performs equally well on all the feature subsets.

6 Conclusion

The present study proposed an e-diagnosis approach to discriminating PD from healthy controls. It is well established from previous research that early PD symptoms can be quantified to detect PD. Here, a self-reported online dataset extracted from the PPMI Online project was studied and its characteristics were used to detect PD**.** The demographic and health status characteristics that are relevant for the detection of PD, present in this dataset. A comparative analysis of this dataset with established datasets (FI and PPMI) validated the characteristics. In addition, we studied the discriminatory quality of each feature using the Shapley Additive Explanations algorithm and the results conclude that some symptoms are more accurately capable of classifying between PD patients and HC. SVM, LR, CatBoost, ELM, XGBoost, and RF were used to classify PD patients from HC using features extracted from well-drafted questionnaires. Significant characteristics were selected using filter and wrapper methods from PPMI Online dataset to improve the classification capability of ML models. Filter methods suggested statistical relevance to a large feature subset, while wrapper methods identified optimized feature subsets using ML algorithm. ANOVA, MI, and chi-square recommended subsets 52, 64, and 101 feature, respectively. Boruta advocated for 20 features primarily composed of motor and non-motor aspects of life. SFS identified 12 features and best accuracy on this subset was 93.63% with RF classifier. RFE suggested 14 features and that UPDRS-II score is the most significant feature to predict PD. RF achieved the highest mean accuracy of 95.48% on RFE feature subsets. In the dataset, healthy controls are overrepresented and hence the performance of classifiers was also evaluated using metrics such as AUC and F1-score. It can be concluded from the results that classifiers have learned from both classes. The distribution of both the classes in train and test datasets was also maintained as in the original dataset. The proposed approach can assist clinicians to determine prodromal PD with high accuracy and also contribute to the digital health system.

Table 12 Performance metrics with RF classifier on test dataset

Feature subset	Accuracy	Precision	Recall	AUC	F1-score
ANOVA	92.98	92.47	91.73	92.71	0.91
Chi-square	92.55	92.16	91.24	92.18	0.91
MI	93.82	92.88	92.63	93.56	0.92
Boruta	**94.78**	**93.70**	**93.45**	**93.20**	**0.93**
SFS	93.63	92.78	92.47	93.25	0.92
RFE	94.95	93.69	93.68	93.78	0.93

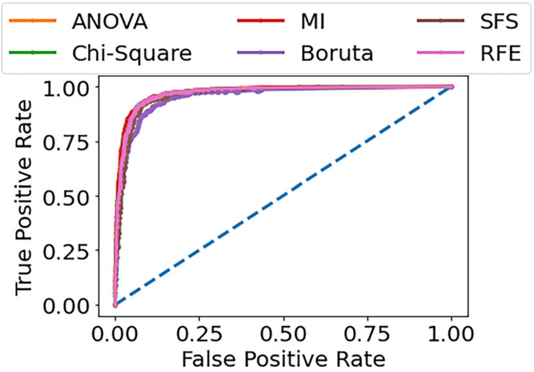


Fig. 12 ROC curves for RF classifier trained on different feature datasets

One of the main limitations of this study was that the dataset has a significantly higher representation of whites than of other races. Hence, in the future, if data will be available globally, this approach can be optimized for the generic population. Also, a study will be conducted to screen, identify subtypes of PD, and categorize the specific symptoms for each subtype. Furthermore, the progression of PD can be modeled and identify stages in the future.

Appendix 1

See Table 13.

Table 13 Comparison of performance metrics on imputed and non-imputed dataset

Dataset	Accuracy	Precision	Recall	AUC	F1-score
Without imputation (*n* = 14,210)	95.40	95.23	95.85	95.17	95.12
KNN(2)	96.82	96.03	96.52	96.88	96.14
KNN(4)	**96.97**	**96.11**	**96.10**	**96.63**	**96.55**
KNN(8)	96.43	95.81	95.41	98.51	95.96

Highest results are achieved using KNN(4) as shown in bold in the table

Data split—90% training and 10% testing

Model—Random forest

Appendix 2

PPMI Online is a longitudinal study; in this project some features were collected at the time of screening, while some were collected in different online meetings. Questionnaires related to smoking and Epworth Sleepiness Scale (ESS) were not collected before the online meeting-2 (OL02). Hence to evaluate the effect of these features we have included them in the main study. Since PPMI Online is an online project, hence it is not mandatory that every registered participant complete all the questionnaires of every online meeting. Only 8419 participants responded in OL02, after removing missing value instances only 7403 participants remained for the study. We have scaled data after unifying questionnaires. Dataset was split into two parts for training(90%) and testing(10%) purposes.

To select significant features, RFE algorithm was used. This algorithm has suggested 15 features including ESS_total(Epworth Sleepiness Score). RF classifier was trained on the training dataset only with 15 features. The model was tested using the test dataset, and results are reported in Table 14. To check the impact of the selected features on the detection of PD, SHAP model was also trained. In Fig. 13, the significance of features to identify PD was plotted. First-degree relatives have PD that was the most important feature followed by UPDRS-II score. ESS_total was also considered in the best 15 features, while features related to smoking were not selected even in the best 30 features.

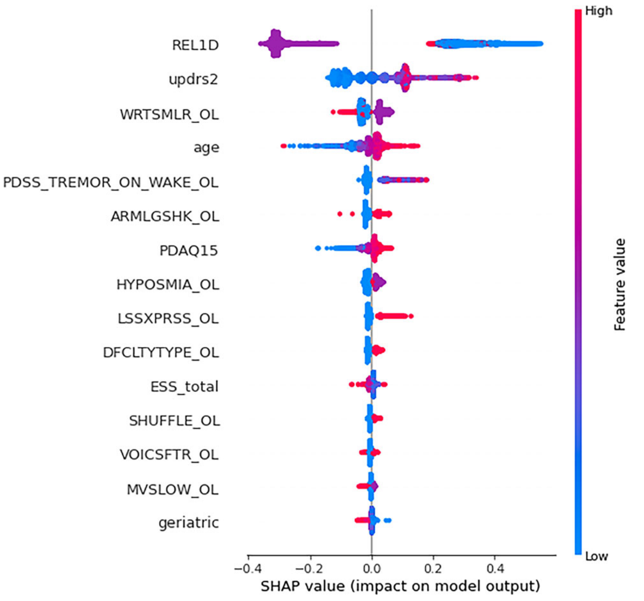


Fig. 13 Impact of feature value on prediction of PD for dataset

Appendix 3

See Fig. 14.

Table 14 Performance metrics on the dataset included second meeting features

Dataset	Accuracy	Precision	Recall	AUC	F1-score
Dataset including 0L02 features (*n* = 7403)	97.44	97.38	97.49	97.66	97.43

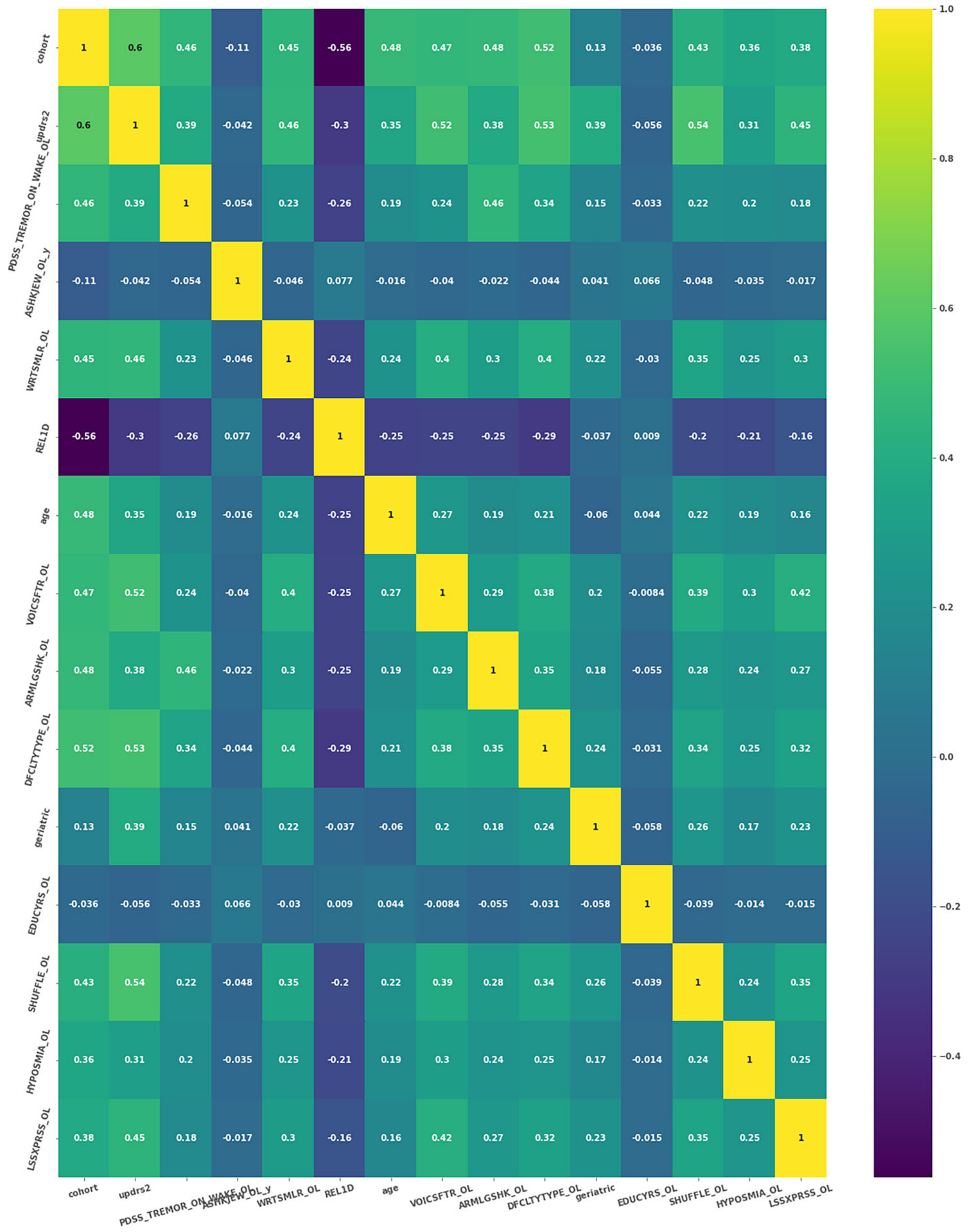


Fig. 14 Relation between features and with cohort(PD or HC) selected by RFE

Acknowledgements The authors thank Project SAMARTH, an initiative of the Ministry of Education (MoE), Government of India, at University of Delhi South Campus (UDSC), for their support.

Funding The authors have not received any grant for this work.

Data availability The datasets analyzed during the current study are available on the official website of Parkinson's Progression Markers Initiative (PPMI), https://www.ppmi-info.org/.

Declarations

Conflict of interest The authors of the study have no conflict of interest with any people or organizations.

References

1. Fahn S (2003) Description of Parkinson's disease as a clinical syndrome. Ann N Y Acad Sci 991:1–14. https://doi.org/10.1111/j.1749-6632.2003.tb07458.x
2. Kalia LV, Kalia SK, Lang AE (2015) Disease-modifying strategies for Parkinson's disease. Mov Disord 30(11):1442–1450. https://doi.org/10.1002/mds.26354
3. Jankovic J (2008) Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatr 79(4):368–376. https://doi.org/10.1136/jnnp.2007.131045
4. Tysnes OB, Storstein A (2017) Epidemiology of Parkinson's disease. J Neural Trans 124:901–905. https://doi.org/10.1007/s00702-017-1686-y
5. Wanneveich M, Moisan F, Jacqmin-Gadda H, Elbaz AJP (2018) Projections of prevalence, lifetime risk, and life expectancy of

Parkinson's disease (2010–2030) in France. Mov Disord 33:1449–1455. https://doi.org/10.1002/mds.27447

6. Willis AW, Schootman M, Evanoff BA, Perlmutter JS, Racette BA (2011) Neurologist care in Parkinson disease: a utilization, outcomes, and survival study. Neurology 77:851–857. https://doi.org/10.1212/WNL.0b013e31822c9123
7. Smolensky L, Amondikar N, Crawford K, Neu S, Kopil C, Daeschler M, Riley L, 23andMe Research Team, Brown E, Toga AW, Tanner C (2022) Fox Insight Collects Online, longitudinal patient-reported outcomes and genetic data on Parkinson's disease. https://www.medrxiv.org/content/10.1101/19002659v1. Accessed 22 Aug 2022
8. Aggarwal N, Shukla U, Saxena GJ, Rawat M, Bafila AS, Singh S, Pundir A (2023) Mean based relief: an improved feature selection method based on ReliefF. Appl Intell 53:23004–23028. https://doi.org/10.1007/s10489-023-04662-w
9. Brajovic S, Blaser DA, Zisk M, Caligtan C, Okun S, Hall M, Pamer CA (2018) Validating a framework for coding patient-reported health information to the medical dictionary for regulatory activities terminology: an evaluative study. JMIR Med Inform 6(3):e42. https://doi.org/10.2196/medinform.9878
10. Dobkin R, Amondikar N, Caspell-Garcia C, Barnes J, Bataille L, Chahine L, Katz A, Kopil C, Marras C, Melnick A, Schwieger T, Siddiqi B, Smolensky L, Standaert D, Tanner C (2019) Innovative recruitment strategies increase diversity of participation in the Fox Insight Longitudinal Cohort, Program No. S16.009. Neurology 95:S16-009
11. Chahine LM, Chin I, Caspell-Garcia C, Standaert DG, Brown E, Smolensky L, Arnedo V, Daeschler D, Riley L, Korell M, Dobkin R, Amondikar N, Gradinscak S, Shoulson I, Dean M, Kwok K, Cannon P, Marek K, Kopil C, Tanner CM, Marras C (2020) Fox insight study. comparison of an online-only Parkinson's disease research cohort to cohorts assessed in person. J Parkinsons Dis 10(2):677–691. https://doi.org/10.3233/JPD-191808
12. Parkinson Progression Marker Initiative (2011) The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol 95:629–635
13. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, Simuni T, Jennings D, Tanner CM, Trojanowski JQ, Shaw LM, Seibyl J, Schuff N, Singleton A, Kieburtz K, Toga AW, Mollenhauer B, Chahine LM, Weintraub D, Foroud T, Tosun-Turgut D, Poston K, Arnedo V, Frasier M, Sherer T, the Parkinson's Progression Markers Initiative (2018) The Parkinson's progression markers initiative (PPMI) establishing a Parkinson's disease biomarker cohort. Ann Clin Transl Neurol 5(12):1460–1477. https://doi.org/10.1002/acn3.644
14. Chang V, Bailey J, Xu QA, Sun Z (2022) Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl 35(32):16157–16173. https://doi.org/10.1007/s00521-022-07049-z
15. ÖzbayKarakuş M, Er O (2022) A comparative study on prediction of survival event of heart failure patients using machine learning algorithms. Neural Comput Appl 34:13895–13908. https://doi.org/10.1007/s00521-022-07201-9
16. Sharma K, Chawla R, Shukla U, Aggarwal N Bafila AS, Singh S, Pundir A, Saxena GJ (2022) Predictive analysis based on feature relevance estimation for survival rate post heart failure using ensemble learners. In: 2022 IEEE international conference on distributed computing and electrical circuits and electronics (ICDCECE). Ballari, India, pp 1–7. https://doi.org/10.1109/ICDCECE53908.2022.9792679
17. Chadaga K, Chakraborty C, Prabhu S, Umakanth S, Bhat V, Sampathila N (2022) Clinical and laboratory approach to diagnose COVID-19 using machine learning. Interdiscip Sci Comput Life Sci 14:452–470. https://doi.org/10.1007/s12539-021-00499-4
18 CN Barreto N, Castro GZ, Pereira RG, Pereira FAN, Reis FM, Junior WM, Cavallo IKD, Gomes KB (2022) Predicting in vitro fertilization success in the Brazilian public health system: a machine learning approach. Med Biol Eng Comput 60(7):1851–1861. https://doi.org/10.1007/s11517-022-02569-1
19. Aggarwal N, Shukla U, Saxena GJ, Kumar M, Bafila AS, Singh S, Pundir A (2023) An improved technique for risk prediction of Polycystic Ovary Syndrome (PCOS) using feature selection and machine learning. In: Shukla A, Murthy BK, Hasteer N, Van Belle JP (eds) Computational intelligence: select proceedings of InCITe 2022. Singapore: Springer Nature Singapore, pp 597–606. https://doi.org/10.1007/978-981-19-7346-8_51
20. Shukla U, Saxena GJ, Kumar M, Bafila AS, Pundir A, Singh S (2021) An improved decision support system for identification of abnormal EEG signals using a 1D convolutional neural network and Savitzky-Golay filtering. IEEE Access 9:163492–163503. https://doi.org/10.1109/ACCESS.2021.3133326
21. Thushari PD, Aggarwal N, Vajrobol V et al (2023) Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions. Soc Netw Anal Min 13:141. https://doi.org/10.1007/s13278-023-01145-1
22. Vajrobol V, Aggarwal N, Shukla U et al (2023) Explainable cross-lingual depression identification based on multi-head attention networks in Thai context. Int J Inf Tecnol. https://doi.org/10.1007/s41870-023-01512-3
23 Kurlan R, Lang AE, Marek K, Kieburtz K, Oakes D, Elliott R, Shoulson I (2009) Parkinson Study Group LABS-PD Investigators. A longitudinal program for biomarker development in Parkinson's disease: a feasibility study. Mov Disord 24(14):2081–90. https://doi.org/10.1002/mds.22690
24. Belic M, Bobic V, BadŽa M, Šolaja N, Đurić-Jovičić M, Kostic VS (2019) Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease–a review. Clin Neurol Neurosurg 184:105442. https://doi.org/10.1016/j.clineuro.2019.105442
25. Choi H, Ha S, Im HJ, Paek SH, Lee DS (2017) Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. NeuroImage Clin 16:586–594. https://doi.org/10.1016/j.nicl.2017.09.010
26. Segovia F, Górriz JM, Ramírez J, Martínez-Murcia FJ, Castillo-Barnes D (2019) Assisted diagnosis of Parkinsonism based on the striatal morphology. Int J Neural Syst 29:1950011. https://doi.org/10.1142/S0129065719500114
27. Pham TD, Yan H (2018) Tensor decomposition of gait dynamics in Parkinson's disease. IEEE Trans Bio-Med Eng 65:1820–1827. https://doi.org/10.1109/TBME.2017.2779884
28. Ma C, Ouyang J, Chen HL, Zhao XH (2014) An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach. Comput Math Methods Med. https://doi.org/10.1155/2014/985789
29. Pereira CR, Pereira DR, Rosa GH, Albuquerque VHC, Weber SAT, Hook C, Papa JP (2018) Handwritten dynamics assessment through convolutional neural networks: an application to Parkinson's disease identification. Artif Intell Med 87:67–77. https://doi.org/10.1016/j.artmed.2018.04.001
30. Váradi C, Nehéz K, Hornyák O, Viskolcz B, Bones J (2019) Serum Nglycosylation in Parkinson's disease: a novel approach for potential alterations. Molecules 24:2220. https://doi.org/10.3390/molecules24122220
31. Maass F, Michalke B, Willkommen D, Leha A, Schulte C, Tönges L et al (2020) Elemental fingerprint: reassessment of a cerebrospinal fluid biomarker for Parkinson's disease. Neurobiol Dis 134:104677. https://doi.org/10.1016/j.nbd.2019.104677
32. Iranzo A, Molinuevo JL, Santamaría J, Serradell M, Martí MJ, Valldeoriola F, Tolosa E (2006) Rapid-eye-movement sleep

behaviour disorder as an early marker for a neurodegenerative disorder: a descriptive study. Lancet Neurol 5(7):572–577. https://doi.org/10.1016/S1474-4422(06)70476-8

33. Uehara Y, Ueno SI, Amano-Takeshige H, Suzuki S, Imamichi Y, Fuzimaki M, Ota N, Murase T, Inoue T, Saiki S, Hattori N (2021) Non-invasive diagnostic tool for Parkinson's disease by sebum RNA profile with machine learning. Sci Rep 11:18550. https://doi.org/10.1038/s41598-021-98423-9

34. Mughal H, Javed AR, Rizwan M, Almadhor AS, Kryvinska N (2022) Parkinson's disease management via wearable sensors: a systematic review. IEEE Access 10:35219–35237. https://doi.org/10.1109/ACCESS.2022.3162844

35. Prashanth R, Dutta RS (2018) Early detection of Parkinson's disease through patient questionnaire and predictive modelling. Int J Med Inform 119:75–87. https://doi.org/10.1016/j.ijmedinf.2018.09.008

36. Mei J, Desrosiers C, Frasnelli J (2021) Machine learning for the diagnosis of Parkinson's disease: a review of literature. Front Aging Neurosci 13:184. https://doi.org/10.3389/fnagi.2021.633752

37. Prashanth R, Roy SD, Mandal PK, Ghosh S (2014) Parkinson's disease detection using olfactory loss and REM sleep disorder features. In: 2014 36th annual international conference of the IEEE engineering in Medicine and Biology Society. IEEE, pp 5764–5767. https://doi.org/10.1109/EMBC.2014.6944937

38. Berg D, Postuma RB, Adler CH, Bloem BR, Chan P, Dubois B, Gasser T, Goetz CG, Halliday G, Joseph L et al (2015) MDS research criteria for prodromal Parkinson's disease. Mov Disord 30(12):1600–1611. https://doi.org/10.1002/mds.26431

39. Fereshtehnejad S-M, Montplaisir JY, Pelletier A, Gagnon J-F, Berg D, Postuma RB (2017) Validation of the MDS research criteria for prodromal Parkinson's disease: longitudinal assessment in a REM sleep behavior disorder (RBD) cohort. Mov Disord 32(6):865–873. https://doi.org/10.1002/mds.26989

40. Mahlknecht P, Iranzo A, Högl B, Frauscher B, Müller C, Santamarıa J, Tolosa E, Serradell M, Mitterling T, Gschliesser V et al (2015) Olfactory dysfunction predicts early transition to a Lewy body disease in idiopathic RBD. Neurology 84(7):654–658. https://doi.org/10.1212/WNL.0000000000001265

41. Prashanth R, Roy SD, Mandal PK, Ghosh S (2016) High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. Int J Med Inform 90:13–21. https://doi.org/10.1016/j.ijmedinf.2016.03.001

42. Zhang X, Chou J, Liang J, Xiao C, Zhao Y, Sarva H, Henchcliffe C, Wang F (2019) Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. Sci Rep 9(1):797. https://doi.org/10.1038/s41598-018-37545-z

43. Mabrouk R, Chikhaoui B, Bentabet L (2018) Machine learning based classification using clinical and DaTSCAN SPECT imaging features: a study on parkinson's disease and SWEDD. IEEE Trans Radiat Plasma Med Sci 3(2):170–177

44. Severson KA, Chahine LM, Smolensky LA, Dhuliawala M, Frasier M, Ng K, Ghosh S, Hu J (2021) Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. Lancet Digit Health 3(9):e555–e564. https://doi.org/10.1016/S2589-7500(21)00101-1

45. Gu S-C, Zhou J, Yuan C-X, 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. https://doi.org/10.1016/j.jad.2020.02.046

46. Leger C, Herbert M, DeSouza JF (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

47. Martinez-Eguiluz M, Arbelaitz O, Gurrutxaga I et al (2022) Diagnostic classification of Parkinson's disease based on non-motor manifestations and machine learning strategies. Neural Comput Appl. https://doi.org/10.1007/s00521-022-07256-8

48. Prince J, Andreotti F, De Vos M (2018) Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. IEEE Trans Biomed Eng 66(5):1402–1411

49. Frasca M, Tortora G (2022) Visualizing correlations among Parkinson biomedical data through information retrieval and machine learning techniques. Multimed Tools Appl 81:14685–14703. https://doi.org/10.1007/s11042-021-10506-x

50. Colosimo C, Martínez-Martín P, Fabbrini G, Hauser RA, Merello M, Miyasaki J, Poewe W, Sampaio C, Rascol O, Stebbins GT, Schrag A, Goetz CG (2010) Task force report on scales to assess dyskinesia in Parkinson's disease: critique and recommendations. Mov Disord 25(9):1131–1142. https://doi.org/10.1002/mds.23072

51. Smith T, Gildeh N, Holmes C (2007) The Montreal Cognitive Assessment: validity and utility in a memory clinic setting. Can J Psychiatr 52(5):329–332. https://doi.org/10.1177/070674370705200508

52. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgans S, LeWitt PA, Nyenhuis D, Olanow CW, Rascol O, Schrag A, Teresi JA, van Hilten JJ, LaPelle N, Movement Disorder Society UPDRS Revision Task Force (2008) Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 23:2129–2170

53. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO (1982) Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 17(1):37–49

54. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digital Med 3(1):81. https://doi.org/10.1038/s41746-020-0288-5

55. Brown E, Alonso D, Chahine L, Coffey C, Dobkin R, Korell M, Lorenzo A et al (2022) Parkinson's Progression Markers Initiative (PPMI) Online expands biomarker research in Parkinson's disease (PD) (P9–11.006). Neurology 98(18):1529

56. Brennan L, Siderowf A, Rubright JD, Rick J, Dahodwala N, Duda JE, Hurtig H, Stern M, Xie SX, Rennert L, Karlawish J, Shea JA, Trojanowski JQ, Weintraub D (2016) Development and initial testing of the Penn Parkinson's daily activities questionnaire. Mov Disord 31(1):126–134. https://doi.org/10.1002/mds.26339

57. Rodríguez-Pérez R, Bajorath J (2019) Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J Med Chem 63(16):8761–8777

58. Kaufmann J, Schering AG (2014) Analysis of variance ANOVA. Wiley. https://doi.org/10.1002/9781118445112.stat06938

59 McHugh ML (2013) The chi-square test of independence. Biochem Med (Zagreb) 23(2):143–9. https://doi.org/10.11613/bm.2013.018

60. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550. https://doi.org/10.1109/72.298224

61 Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Stat Softw 36(11):1–13. https://doi.org/10.18637/jss.v036.i11

62. Wang J, Lu S, Wang SH, Zhang YD (2022) A review on extreme learning machine. Multimed Tools Appl 81:41611–41660. https://doi.org/10.1007/s11042-021-11007-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.