Computer Methods and Programs in Biomedicine 234 (2023) 107495

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease

Muhammad Junaid[a], Sajid Ali[a], Fatma Eid[b], Shaker El-Sappagh[c,d,e], Tamer Abuhmed[c,*]

[a] Information Laboratory (InfoLab), Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[b] Technology Management, Stony Brook University, New York 11794, USA
[c] Information Laboratory (InfoLab), College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea
[d] Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt
[e] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha, 13518, Egypt

ARTICLE INFO

Article history:
Received 20 July 2022
Revised 23 February 2023
Accepted 17 March 2023

Keywords:
Parkinson's disease progression
Machine learning
Multi-modality
Feature selection and importance
Class imbalance
Explainable AI

ABSTRACT

Background and objectives: Parkinson's Disease (PD) is a devastating chronic neurological condition. Machine learning (ML) techniques have been used in the early prediction of PD progression. Fusion of heterogeneous data modalities proved its capability to improve the performance of ML models. Time series data fusion supports the tracking of the disease over time. In addition, the trustworthiness of the resulting models is improved by adding model explainability features. The literature on PD has not sufficiently explored these three points.

Methods: In this work, we proposed an ML pipeline for predicting the progression of PD that is both accurate and explainable. We explore the fusion of different combinations of five time series modalities from the Parkinson's Progression Markers Initiative (PPMI) real-world dataset, including patient characteristics, biosamples, medication history, motor, and non-motor function data. Each patient has six visits. The problem has been formulated in two ways: ❶ a three-class based progression prediction with 953 patients in each time series modality, and ❷ a four-class based progression prediction with 1,060 patients in each time series modality. The statistical features of these six visits were calculated from each modality and diverse feature selection methods were applied to select the most informative feature sets. The extracted features were used to train a set of well-known ML models including Support vector machines (SVM), random forests (RF), extra tree classifier (ETC), light gradient boosting machines (LGBM), and stochastic gradient descent (SGD). We examined a number of data-balancing strategies in the pipeline with different combinations of modalities. ML models have been optimized using the Bayesian optimizer. A comprehensive evaluation of various ML methods has been conducted, and the best models have been extended to provide different explainability features.

Results: We compare the performance of ML models before and after optimization and using and without using feature selection. In the three-class experiment and with various modality fusions, the LGBM model produced the most accurate results with a 10-fold cross-validation (10-CV) accuracy of 90.73% using non-motor function modality. RF produced the best results in the four-class experiment with various modality fusions with a 10-CV accuracy of 94.57% using non-motor modality. With the fused dataset of non-motor and motor function modalities, the LGBM model outperformed the other ML models in both the 3-class and 4-class experiments (i.e., 10-CV accuracy of 94.89% and 93.73%, respectively). Using the Shapely Additive Explanations (SHAP) framework, we employed global and instance-based explanations to explain the behavior of each ML classifier. Moreover, we extended the explainability by implementing the LIME and SHAPASH local explainers. The consistency of these explainers has been explored. The resultant classifiers were accurate, explainable, and thus medically more relevant and applicable.

* Corresponding author.
E-mail addresses: mjunaid@skku.edu (M. Junaid), sajidali@skku.edu (S. Ali), fatma.eid@stonybrook.edu (F. Eid), shaker@skku.edu (S. El-Sappagh), tamer@skku.edu (T. Abuhmed).

https://doi.org/10.1016/j.cmpb.2023.107495

Conclusions: The select modalities and feature sets were confirmed by the literature and medical experts. The various explainers suggest that the bradykinesia (NP3BRADY) feature was the most dominant and consistent. By providing thorough insights into the influence of multiple modalities on the disease risk, the suggested approach is expected to help improve the clinical knowledge of PD progression processes.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative brain illness that progresses slowly [43]. The term *neurodegenerative* indicates a condition that results in the death of brain cells. The first signs of PD arise in the enteric nervous system, lower brain stem, and olfactory tracts [74]. The brain shell and the substantia nigra are affected by PD, which progresses from the areas of the first signs to the upper sections of the brain. Injury to the upper sections of the brain affects the part of the brain that controls movement and cognition [40]. The illness is considered to start several years before the appearance of motor symptoms including the loss or reduction of smell sleep difficulties and constipation, tremor, and slowness of movement. Further, 90 percent of PD patients suffer from vocal difficulties [71]. The symptoms of the illness intensify with time, and those in severe stages may experience dementia and hallucinations [34]. As a corollary, researchers are currently searching for strategies to identify these non-motor symptoms as early as possible to limit the disease's development.

Because of its simple implementation and high accuracy, Machine Learning (ML) is increasingly being used to diagnose medical conditions [2,17,18,20,21]. ML has been used to treat PD in particular. For example, Wan et al. [90], concentrates on research conducted following the diagnosis of PD. Patients with PD were operated on by the authors. In that study, an ML-based technique was used during brain surgery for PD to identify the real area to be operated on. To evaluate the cognitive repercussions of PD, Salmanpour et al. [72] used ML approaches including LASSOLAR (Least Absolute Shrinkage and Selection Operator - Least Angle Regression), GA (Genetic Algorithm), LOLIMOT (Local Linear Model Trees), DE (Differential evolution), and NSGAII (Non-dominated sorting genetic algorithm). Further, Pedrosa et al. [62] used an ML-based algorithm, which includes *k*-nearest neighbors (K-NN) Classifier and linear Support Vector Machine (SVM), to predict the amount of tremor in Parkinson's patients. In another study, ML was used to predict the stage of PD. For example, Prashanth and Roy [64] used Ordinal Logistic Regression (OLR), AdaBoost, RUSBoost-based, SVM classifiers to establish a new PD staging system based on commonly recognized clinical scales. However, the majority of the existing studies have focused on using the popular ML method to diagnose PD during the appearance of early symptoms. For instance, Cavallo et al. [11] attempted to predict PD using motion data collected from people's upper limbs to aid in early diagnosis. The authors implanted a device (i.e., SensHand V1) into the upper limbs of the experimental subjects and asked them to execute a series of activities. Spatio-temporal and frequency data analyses were used to calculate 48 parameters with three supervised ML algorithms (SVM, Random Forest (RF), and Naive Bayes (NB)) using a dataset of healthy ($n = 30$), Idiopathic Hyposmia (IH) ($n = 30$), and PD ($n = 30$) participants. For the diagnosis of PD, Almeida et al. [6] applied several feature extraction approaches and ML algorithms. The authors found that phonation is the most practical activity for detecting PD. Classifiers such as SVM, *k*-NN, and Multilayer Perceptron (MLP) were examined in this study.

Several different techniques can be used for the progression and detection of PD, for example, Nilashi et al. [58] employed an unsupervised technique to predict PD. In that study, the authors were able to predict the Unified PD Rating Scale (UPDRS) by lowering the dimension via partial least squares approaches, clustering via a self-organizing map, and utilizing incremental support vector regression prediction. Das [15] investigated several strategies and concluded that Neural Networks (NNs) were the most effective ML methods for predicting PD, achieving 92.9% accuracy in comparison with DMneural, Decision Tree (DT), and regression. For feature weighting and classification in a PD diagnostic task, Polat [63] used fuzzy *C*-means clustering and *k*-NN. The optimal *k* value was obtained by feeding a weighted PD dataset to a *k*-NN classifier. The optimal accuracy of 97.93% is achieved with k equal to three. Further, the adaptive Artificial Bee Colony (ABC) technique was used for feature selection and optimization to diagnose PD [91]. To ameliorate the unbalanced data, the researchers used a weighted technique and non-linear kernel function mapping to achieve a model accuracy of 98.97% with 5-fold cross-validation. Using Principal Component Analysis (PCA) for dimension reduction, Fisher Discriminant Ratio (FDR) for feature selection, and SVM for classification, Singh et al. [81] obtained a high binary-class accuracy of 95% and a multi-class accuracy of 85% for PD diagnosis. Another study for the diagnosis of PD, Abdulhay et al. [1] examined gait and tremor using the PhysioNet database to report an average accuracy of 92.7%.

Although ML-based classification methods have shown considerable accuracy for the diagnosis of PD, the literature has reported two different ways methods: ❶ numerous features were employed in the analysis, thus increasing the cost of the diagnosis [1,6,91], and ❷ it was even more difficult to extract few features [58,63,81]. Further, research investigations often have minimal impact on clinical practice for the following reasons: First, the detection and diagnosis of PD progression were explored independently [77]. However, detection and diagnosis are dependent on each other. In addition to identifying the presence of Parkinson's disease, it is also critically important to identify the progression of PD from one stage to another. Identifying patients at substantial risk of progressing helps patients and their caregivers take appropriate precautions. As a consequence, PD diagnosis and progression involve correlated processes [59]. Second, the majority of investigations, particularly neuroimaging, rely on a single modality [44,92]. Other than MRI, handwriting tasks [41], motion [11], or speech data [74]; has been used to diagnose PD using ML approaches. Meanwhile, PD is a multimodal illness [65]. Multiple symptoms of PD can be used to accurately detect and identify the progression. Third, current research has mostly focused on improving the performance of complicated ML models while ignoring their explainability [50]. As a result, despite the fact that significant breakthroughs have been achieved in prediction, ML-based models are unlikely to be accepted in a medical setting [19,37,66]. Doctors find it difficult to evaluate ML models and they believe that such models are untrustworthy. Fourth, patients with PD go through several stages throughout their lives. Time is a significant variable in the development of the disease [75]. However, few works in the literature have considered PD progression prediction as a time-series problem. Therefore, in the current study, we propose a framework for PD progression prediction based on time-series data that supports explainability for the predicted output.

Glossary

PD	Parkinson's disease
AD	Alzheimer's disease
DBS	Deep brain stimulation
PPMI	Parkinson's progression Markers initiative
SVM	Support vector Machine
ANN	Artificial neural networks
RF	Random Forest
LR	Linear Regression
DT	Decision Tree
ET	Extra Tree
NB	Naive Bayes
LGBM	Light Gradient Boosted Machine
SGD	Stochastic gradient descent
ADB	Adaptive Boosting
k-NN	*k*-Nearest Neighbour
SMOTENN	Synthetic Minority Oversampling technique Edited Nearest Neighbor
SC	Subject characteristics
BS	BIO-Specimen
MH	Medical History
M	Motor Function
NM	Non-Motor Function
M-NM	Motor Function and Non-Motor Function
M-SC	Motor Function and Subject characteristics
M-MH	Motor Function and Medical History
M-MH-SC	Motor Function, Non-Motor Function and Medical History

Contributions This research aims to propose a novel framework that exploits patients' multimodality time-series data to predict their progression and support decision explainability. The study makes the following six-fold contributions:

- ❶ We propose an end-to-end machine learning pipeline based on time series data to predict PD. The pipeline provided the possibility to explore different approaches of (i) fusion of data modalities, (ii) improving the quality of data using different techniques for missing values handling, data balancing, feature selection, data normalization, etc., (iii) optimization of different machine learning models, (iv) mapping of time series data to a non-time series data, and (v) providing XAI capabilities.
- ❷ We optimize the PD progression detection problem as two classification tasks with different complexity levels, i.e., *three-class* and *four-class* tasks according to the H&Y scale (discussed in Section 2.4). For each task, five medically relevant time series modalities are explored, including *Subject Characteristics* (SC), *Bio-Samples* (BS), *Medication History* (MH), *Motor function* (M), and *Non-Motor (MN) function* modalities. Further, different fusions of these modalities are examined to detect the most critical modalities and the specific features in each modality. These fusions provide a deeper understanding of the relationship among modalities and their contribution to the prediction. The resulting feature set leads to personalized and customized decisions for PD prediction process.
- ❸ The study has compared the performance of several widely used ML algorithms for diagnosing PD. The models were tuned using the Bayesian optimizer. In our study, multiple feature selection strategies were tested to identify the most effective feature selection method and feature list. In addition, ML models for diagnosing PD patients with three or four classes were developed and validated using real-world time series data.
- ❹ Real-world medical data are normally imbalanced. To solve this issue, variants of the over-sampling techniques such as *SVMSMOTE, BorderlineSMOTE, ADASYN (Adaptive Synthetic), SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous)*, and *SMOTEENN (Synthetic Minority Over-sampling Technique with Ensemble of Neighbors*) are compared and the optimal technique is selected by performing multiple experiments on different data modalities.
- ❺ To provide trustworthy and medically acceptable decisions, the optimized ML models are extended to provide explainable decisions. We explore different mechanisms to provide the XAI capabilities. We develop and evaluate the *SHAP* feature attribution technique to provide both global and instance-based explanations. Further, two explainers based on *LIME* and *SHAPASH*, are implemented to offer insight explanations for each classifier's decisions. The resulting models provide accurate and explainable decisions.

Organization The rest of the paper follows as outlined below: Section 2 discusses the related work in terms of PD diagnosis and progression. Section 3 highlights the materials and discusses every component of the proposed framework. Section 4 discusses the results for three-class and four-class problems in terms of the concept of explainability. Section 5 discuss in-depth the experimental analysis along with the limitation and future challenges. Section 6 concludes the paper with the future scope research.

2. Related work

There are two key concerns associated with PD. Parkinson's progression is the first, and Parkinson's stage detection is the second. We go over both of these problems in detail and discuss the related work examining these issues.

2.1. PD diagnosis

The phrase *diagnosis* refers to the process of determining whether a person has PD. There has been a lot of research aiming to diagnose PD patients. For example, Segovia et al. [73], developed a technique for classifying non-PD and PD patient brain images. Their goal was to use sections of each hemisphere of the brain to diagnose PD independently. The classification rates were around 94.7% when SVM and partial least squares were used as a classifier and dimensional reduction method, respectively. Mohammed et al. [56] suggested a patient-specific dynamic feature extraction method that uses local field potential signal in a fusion with adaptive SVM to identify PD or non-PD patients by modifying the decision boundary until a feasible solution is identified. Their technique achieved a 98% classification accuracy. Using fuzzy models, Camara et al. [10] developed an autonomous real-time method for identifying resting tremor experiences in 10 Parkinson's patients. The classification step included electrophysiological data from local field potential and electrocardiography and achieved an accuracy of 98.7%.

Further, Rojas et al. [70] established a novel approach for extracting brain SPECT image features that showed 95% accuracy. Wu et al. [93] used an Auto-Regressive (AR) model to study the stochastic process of the idiopathic PD patients' stride series, which were used as features to distinguish PD stride series from non-PD cases. Their approach utilized the SVM classifier to differentiate the healthy and idiopathic PD groups, with the results 89% specificity, 72% sensitivity, and 83% Area Under the Curve (AUC) suggesting that regressive model settings might be useful for stride series classification. The clinical Decision Support Systems (CDSS)

model was suggested by [79] to examine the impact of merging patient-specific symptoms and drugs into three important functions: i.e., retrieving information, visualizing the therapy, and making suggestions on projected effective stimulation and medication doses. The authors used GNB, SVM, and RF to predict treatment outcomes. The combined ML algorithms accurately predicted 86% of the motor improvement scores one year after surgery. Parkinson individuals are non-invasively identified via gait analysis. Using gait variables, extract features using a wavelet transform. 100% categorization accuracy when combining all spatiotemporal factors. Joshi et al. [35]

Using empirical mode deformation to filter electromyograms, Dai et al. [14] proposed a novel method for identifying PD. In this method, signals were preprocessed in three phases using a novel bandpass filtering method to show that the features are linearly separable. The recommended technique was later developed as a smartphone application to be more adaptable than existing alternatives. Hirschauer et al. [29] devised another unique PD diagnostic approach that is based on continuous phonation data. The most important attributes were determined using the minimum redundancy maximum relevance approach, and the results are reported for various feature selection strategies. Their work applied the feature selection process and fed data into two kinds of neural classifiers: i.e., a standard Artificial Neural Network (ANN) and a Complex-Valued Artificial Neural Network (CVANN). The ANN achieved an accuracy of 94.28% while the CVANN achieved an accuracy of 98.12%. The expert-driven DSS model performs better in terms of accuracy than the data-driven model and is more like the judgments made by doctors. The accuracy findings show that the built models are suitable and appropriate for the purpose of recommending options to DSS users [9]. Deep brain simulations and EEG signals are vital for the study of Parkinson's disease. In this work, an automated tunable Q wavelet transform (A-TQWT) is developed to extract numerous subbands (SBs) from these signals and then analyze them using machine learning methods because manual analysis of these signals is laborious. Using EEG data from an open source dataset, the suggested technique successfully distinguishes between healthy controls and PD patients on medication and those who are not. The area under the curve for the suggested method's classification accuracy was 97% and 98.56%, respectively, while its accuracy was 96.13% and 97.65% [38]. The authors of this research noted that manual procedures are necessary for typical machine learning approaches, which are inconvenient. An automated PD detection technique dubbed PDCNNet, which combines Smoothed Pseudo Wigner-Ville Distribution (SPWVD) and convolutional neural network (CNN), is proposed to get over these restrictions. With SPWVD, the EEG signals are converted to time-frequency representation (TFR) and sent to a CNN model. Using two open databases, the suggested model successfully diagnoses Parkinson's disease with high accuracy. For one of these datasets, the accuracy of 100% and 99.97% were attained [39].

2.2. PD progression

Once the PD patients have been recognized, the next task is to determine whether each PD patient is expected to soon progress to an advanced stage of PD. There have been many studies examining the progression of PD to identify patient progression at different PD stages. For instance, Drotár et al. [16] used the SVM-RBF kernel technique to classify the stages of PD. The authors used handwriting characteristics -such as in-air movements, pressure, and control- of PD patients' hands to determine the PD stage and achieved an accuracy percentage of 81.3%. Further, Connolly et al. [13] employed SVM, LDA, and *k*-NN to determine the PD stage using a deep brain stimulation device that was implanted in 15 Parkinson's patients ahead of time and they achieved 91% accuracy. In their study, the diagnosis of PD progression was based on ML methods. The researchers used feature selection and classification to identify PD by utilizing ML approaches, such as ANN, SVM, and regression trees, and they achieved an accuracy of 93.84% using SVM. In another study [60] conducted a thorough evaluation using the recent PD detection ML algorithms; although ML encompasses a wide range of algorithms, picking the best one was not straightforward. Their work compares and identifies which of three classifiers -Multilayer Perceptron, SVM, and *k*-NN- is the most powerful and reliable for PD classification on the voice dataset.

Further, most of the studies on PD progression have concentrated on non-motor features. For example, rapid eye movement and sleep behavior disorders are both characteristic non-motor features. Prashanth et al. [65] integrated non-motor features, Cerebrospinal Fluid (CSF), and dopamine marker tests. Their work used the PPMI dataset and employed GNB, SVM, Boosted Tree, and RF to classify patients, and they achieved an accuracy of 96.4% using SVM. Al-Fatlawi et al. [4] presented the Deep Belief Network (DBN) as a tool for PD progression. The study used data collected from the UCI data repository for 195 voice recordings of 31 individuals with 16 features, and the proposed DBN achieved 94% accuracy. Abdulhay et al. [1] used tremors and gait as the basis for their novel diagnosis and they achieved an accuracy of 92.7%.

Evolutionary algorithms have also been used to develop clinically useful models to classify PD patients from healthy controls. For example, Smith et al. [83] utilized Cartesian genetic programming to evaluate human movements, the progression of PD patients from healthy controls, and the degree of dyskinesia. Animal data has been gathered using fruit flies with or without PD genetic defects, while human data has been collected through non-invasive ways utilizing commercial sensors. Ahmadlou and Adeli [3] presented Enhanced Probabilistic Neural Networks (EPNN), an ML approach that controls the spread of the Gaussian kernel by using local decision rings around training data. The proposed approach exhibited 92.5% stage classification performance when considering data from six clinical exams and functional neuroimaging data for two brain regions of interest when using the PPMI3 dataset, and an accuracy of 98.6% when classifying healthy people from PD patients. Using SPECT imaging, Illán et al. [30] proposed a computer-assisted approach for PD progression. Their suggested system consists of a whole image processing pipeline, normalization, and classification. It is intended to assist clinicians in their everyday tasks. Their study achieved an AUC of 0.96% using the SVM classifier [47]. Using EEG recordings from PD patients and healthy controls, this work suggested a deep-learning model for automated Parkinson's disease diagnosis. A 2D-CNN model was trained using these spectrograms after the EEG recordings underwent a Gabor transform to create spectrograms. Using tenfold cross-validation, the suggested model was able to obtain high classification accuracy of 99.46% for three classes (healthy controls, PD patients with and without medication, and other PD patients). Recently, Loh et al. [46] reviewed 63 publications that proposed deep learning models for automated Parkinson's disease diagnosis highlighting several approaches and modalities used in literature, including brain imaging and motion symptoms, among others.

2.3. PD progression detection scales

For PD, there are various measures used to assess patient impairment and disability; i.e., stage and severity. Using common symptoms and biomarkers, the primary scales are established to place the patients on the right stage. The most widely used scales are the Hoehn and Yahr (H&Y) scale [33] and the Unified Parkinson's Disease Rating Scale (UPDRS) [23]. The UPDRS evaluates the most relevant clinical aspects of PD to offer a thorough evaluation of disability and impairment. H&Y scale offers a comprehen-

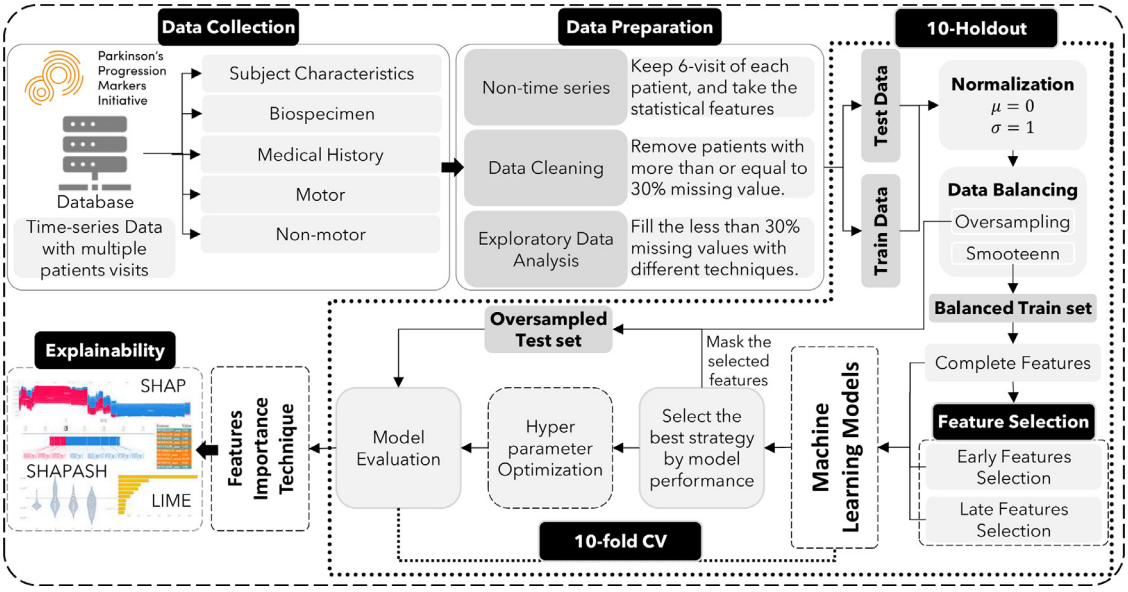

Fig. 1. Architecture of the proposed Parkinson's disease progression prediction model. Multimodality and time-series data are collected from the PPMI database. Essential and standard preprocessing steps are employed in preparing the dataset to train and test the models. In addition, to enhance the model performance, we use normalization, class imbalance, and feature selection techniques. Bayesian Optimizer is used to tune each model. Furthermore, the models are evaluated on the explanations provided by three explainers.

sive evaluation of disease development using a staging system that spans from zero (no evidence of illness) to five (severe disease) [24]. The MDS-UPDRS (Movement Disorder Society UPDRS) is an updated and improved version of the original UPDRS that includes additional items dedicated to numerous non-motor aspects of PD. therefore, it is more comprehensive than the original rating scale [33]. For the present study examining PD detection, the H&Y scale is used. The following is a concise description of the H&Y scales and the MDS-UPDRS.

2.4. Hoehn and Yarn scale

The H&Y scale is a scale that is frequently used to evaluate the overall severity of PD. H&Y scale starts with unilateral (Stage 1) to bilateral without balance problems (Stage 2), then postural instabilities (Stage 3), loss of physical dependency (Stage 4), and lastly being confined to a wheelchair or bed unless supported (Stage 5) [24]. We formulated two different research problems: ❶ The H&Y scale is used to classify PD into a three-class problem consisting of the healthy class (stage 1), the early class (stage 2), and the advance-early class (stage 3). ❷ In classifying PD into a four-class problem, the healthy class (stage 1), the early class (stage 2), the advanced-early class (stage 3), and the advanced class (stage 4) were considered. We eliminated the stages beyond the fourth class since there are few patients at stage 5. Studies have shown that these stages have a considerable relationship with the standard of living indicators and a strong association with the UPDRS [61].

2.5. MDS-UPDRS

Another scale used to determine the severity of PD is the MDS-UPDRS. The MDS-UPDRS comprehensively assesses the clinical aspects of PD (both motor and non-motor) [23]. There are 65 entries on the scale which are divided into four parts. Part ❶, which consists of 13 entries, measures non-motor aspects of everyday life. The first six entries are partly examined by the physician, while the next seven are partially evaluated by a patient questionnaire. Part ❷, which includes another 13 entries, is about motor experiences in everyday life. All the entries are fully assessed by a patient questionnaire. Part ❸ is the motor examination, which consists of 33 entries, and it is examined by a physician. Part ❹ measures motor abnormalities and consists of six entries that a physician reviews once a person with PD have begun taking medication. Research has shown that the MDS-UPDRS is a sensitive and reliable tool for evaluating the development and severity of PD [23,61].

3. Materials and methods

This section introduces an ML framework for detecting Parkinson's Disease progression. The framework consists of multiple stages including data collection, preprocessing, splitting, cross-validation, balancing, model training, hyperparameter optimization, evaluation, and explainability. Figure 1 illustrates the framework stages, and more details about these stages are included in the following Subsections.

3.1. Data collection

This study uses the PPMI database http://www.ppmi-info.org/data, which is a multi-national, large-scale database that records the progression of PD. The database includes comprehensive patient clinical information, including medical images, clinical reports, motor and non-motor test results, biosamples, and patient history and background [51]. The study covers five essential modalities, SC, BS, MH, M, and NM, and extracts a range of data classes from the dataset with their own set of features.

The PD dataset comes in the form of time series (as patient visits) from the PPMI data. For each patient, six common visits are sorted within every 12-month interval, i.e., baseline (BL) in month 0 visit 1 in month 12, visit 2 in month 24, visit 3 in month 36, visit 4 in month 48, and visit 5 in month 60. Table S9 in the supplementary file provides more information on the patient's visits and the number of patients in each visit. Only non-time series data are accessible in the SC modality, although all other modalities offer time series observation. each patient's class is determined using the variable **NHY** based on the H&Y scale, each patient's class is selected. Table 1 shows the sensitive attributes of data and their distribution, including, age, gender, and socio-

Table 1
Parkinson's disease progression dataset insights.

Features		PAT	Male	Female	Total (Average)
PATNO		HC	202	232	217
		PD	392	256	324
Years of education (Avg)		HC	15.95	14.73	15.95
		PD	16.77	17.11	16.77
Age average		HC	62.38	60.01	62.38
		PD	64.38	63.21	64.38
Hand *8pat use both hands	left	HC	24	31	27.5
		PD	35	25	30
	right	HC	170	193	181.5
		PD	343	226	284.5
Family History of pd	0	HC	113	87	100
		PD	220	122	171
	1	HC	90	145	117.5
		PD	172	134	153

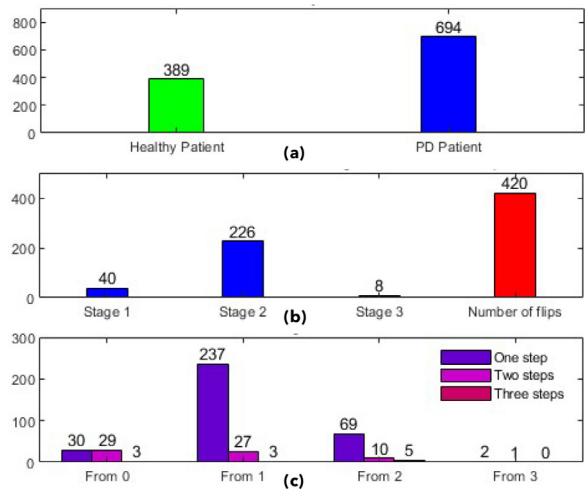

Fig. 2. Number of patients in each class. (a) The total number of healthy vs. Parkinson's patients. (b) The Total number of Parkinson's patients who remain at the same stage (blue) and who shift between stages (red). (c) The total number of patients who shift from one stage to another (single or multiple shifts). For example, there are 30 patients who shifted from stage 0 to stage 1, 29 patients have two shifts (from stage 0 to stage 2), and three patients had three shifts (from stage 0 to stage 3).

economic status. Two datasets are prepared including one for the three-class problem ($\mathbf{n} = \mathbf{953}$) and another for the four-class problem ($\mathbf{n} = \mathbf{1059}$). Figure 2 shows the different number of patients in each class and their progression 2.

3.2. Data preparation

This section discusses the data preparation steps followed prior to the training of the proposed model. The effectiveness of each modality (motor and non-motor) in detecting PD severity was evaluated and the results were shown in Figs. S1 and S2 in the supplementary file. The heterogeneous time series data was preprocessed to improve its quality through a set of steps. The following are some of the steps in the process:

- ❶ Missing data for all modalities are handled in the first stage. All features with a missing data percentage exceeding 30% are eliminated. The missing data are handled in various ways, including forward and backward filling. The median filling is used for the continuous (numerical) features, and the mode technique is used to fill categorical data.
- ❷ In the second stage, the time-series data are transformed into non-time-series data by applying statistical feature extraction on the patient's six visits. The mean, minimum, maximum, variance and standard deviation are calculated for each patient's feature.
- ❸ The dataset is then split into a training set (80%) and a testing set (20%). The training set is used to develop and validate the model using the stratified 10-fold CV technique whereas the testing set is used to determine the model generalization.
- ❹ The data are also rescaled to make them uniform. In the range [0, 1], the data are normalized. At this point, outliers are also identified, and each class's average value is used to replace them.
- ❺ Since ML models are prone to bias in the event of unbalanced datasets, the next step is data balancing. There are a variety of approaches that can be used to deal with imbalanced datasets. Commonly used methods are oversampling and undersampling of the data using the Synthetic Minority Oversampling Technique (SMOTE) and its variants [12]. After testing multiple SMOTE methods, SMOTE Edited Nearest Neighbor (SMOTENN) is used as the best fit in this study to deal with the training set's class imbalance.
- ❻ In the ML process, feature selection is a crucial aspect of model performance. This study explored the impact of various feature selection procedures in detail. We tested feature selection using three approaches: ⓐ, we employed the whole set of features and performed the experiments: ⓑ, we performed an early selection of features, i.e., we selected features in each modality before fusion on modalities; and ⓒ, we performed late feature selection, i.e., we fused all the modalities and then used the feature selection method. We applied ML methods in all three approaches and selected late selection as the best method.

3.3. Data normalization

The normalization of data is a crucial step in preprocessing for successful model training in ML. The Min-Max approach [28] is used to normalize the data by rescaling the data range to [0,1]. The following is the definition of Min-Max normalization:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times (b - a) + a \tag{1}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values respectively. In the range [a,b], the Min-Max normalization translates values x to x'.

3.4. Dataset class balancing

Since the dataset is unbalanced and to ensure the reliability of the reported results, we have addressed the imbalanced dataset using the combination of oversampling the minority class and undersampling the majority class. This has been done using SMOTENN [45] which is only applied to the training dataset, while the testing set was balanced with real patient data to mimic real-world scenarios. To avoid training bias, the balanced training set was used to train the ten ML models using a stratified 10-fold cross-validation approach.

3.5. Feature selection

The process of identifying the most important and relevant features for a prediction task is a crucial aspect of building robust machine learning models [8]. The goal of this process is to determine the smallest set of features that achieves the best-performing model. This approach reduces the risk of overfitting, removes uninformative features, and enables the classifier to focus on the most important variables [31]. Various feature selection techniques are

experimented with to achieve this and the most effective method is utilized. In this case, the recursive XGBoost and SULOV (Synthetic Unlabeled Local Outlier Voting) methods are used to reduce the number of features and identify the best set for the model.

Assuming that a data set is $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, n, x_i \in R^m, y_i \in R^m\}$, having n samples with m features. Let $\hat{y}_i$ be defined as the value predicted by the model:

$$\hat{y}_i = \sum_{i=1}^{n} \mathcal{F}_i(x_i) \tag{2}$$

where $\mathcal{F}_i$ represents an independent regression tree and $\mathcal{F}_i(x_i)$ denotes the prediction score given by the i^{th} tree to the n^{th} sample. The set of functions $\mathcal{F}_i$ in the regression tree model can be learned by minimizing the objective function as follows:

$$\mathcal{O} = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{i=1}^{n} \Omega(\mathcal{F}_i)$$

3.6. Holdout and cross validation

In the ML method adopted for prediction, both the best technique and optimized hyperparameters for the selected model are considered. The nested CV technique, with a 10-fold inner CV and 10-time outer hold-out, is used to evaluate the ML models' prediction ability and generate an unbiased estimate of their accuracy [89]. The outer hold-out process, including data splitting, normalization, class imbalance, and feature selection, is repeated ten times to ensure a robust model, and the model with the highest average accuracy is selected for further examination.

In our experiments, models learned with leave-one-out cross-validation have higher variance than those learned with regular K-fold cross-validation (see Table S13 in the Supplementary file), making leave-one-out CV an inadequate choice in terms of performance, cost, and computational time. Therefore, we proceeded with a 10-fold CV, which impacted the average performance of test results as shown in Section 4.

3.7. ML-based models and training optimization

*Models*ML models with the minimum loss during the training are selected. The trained model with the optimized parameters is then utilized for PD detection in the testing phase. To develop a prognostic model for PD progression detection, the present study assesses various supervised ML algorithms in terms of their ability to predict PD progression. High-performance ML algorithms such as SVM [32], RF, LR [5], DT [55], ET, GNB [94], LGBM, SGD [7], ADB, and *k*-NN [76] are used. These models are optimized to classify the three-stage and four-stage problems with different fusions of modalities. The top 10 ML models with the best performance across distinct modalities are selected in three ways: ❶ with the whole set of features (refer to Table S10 in the Supplementary file for CV and testing accuracy and other performance matrices), ❷ with selected features (refer to the Table S11 in the Supplementary file for CV and testing accuracy and other performance matrices), and ❸ with optimized features selected (refer to Table S12 in the Supplementary file for CV and testing accuracy and other performance matrices).

Bayesian optimizer All machine learning techniques find the best model parameters using a stratified 10-fold CV and Bayesian optimization. Each model is trained using the 10-fold CV method on an 80% training dataset. Ten times each experiment is run, and the average value and standard deviation are provided. The test dataset that was over-sampled by 20% is used to assess the generated models [96].

Let $\mathcal{K}$ denotes a model for which there is no closed-form expression. Considering the model to be expensive to evaluate. The model $\mathcal{K}$ be a well-behaved defined on a subset $\mathcal{X} \subset \mathbb{R}^d$. The goal is then to solve the following global optimization problem:

$$x^* = \arg\max_{x \in \mathcal{X}} \mathcal{K}(x)$$

By using a probabilistic model, Bayesian optimization seeks to identify the model $\mathcal{K}$ global optimum. By utilizing the model to make choices about the training set, $\mathcal{K}$, and integrating out uncertainty. At the expense of requiring extra calculations to select the next point to test [78], it provides the minimum of non-convex functions with very few evaluations [78].

Experiment workstation We performed the experiments on a workstation with 64-bit Windows 10 as the operating system and the hardware configuration of an Intel CPU core i7-9700K with 48 GB DDR4 memory.

3.8. Evaluation metrics

To measure the CV and generalization performance of the proposed model, four standard metrics are used: accuracy [Eq. (3)], precision [Eq. (4)], recall [Eq. (5)], and F1-score [Eq. (6)]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

3.9. Features importance technique

Feature importance scores are crucial in predictive modeling design, as they provide insight into the data, developed model, framework complexity, and feature selection. Therefore, modifying the feature importance can increase the effectiveness of the predictive model on any specific task. We used numerous approaches to discover the top features, including LR, RF, and XGBoost for feature importance, and then reported the best features. In **supplement file 1, S15** shows the details of the feature selection process, and **S16** shows the comparison of the models with complete features, selected features, and after optimization in different modalities.

3.10. Explainability of models

Explainable AI is a collection of tools that aid in the comprehension and interpretation of any ML model's predictions. There are two possible ways to comprehend the model: ❶ The global explanation refers to how the model makes decisions in general, while ❷ the local explanation refers to how the model makes decisions in a specific situation [49]. Attributions, visualization, example-based, game theory, knowledge extraction, and neural networks are among the many post-hoc explainability methodologies that have been proposed [68]. SHAP (Shapley Additive Explanations) [48] is a well-known and reliable approach used in the article. It is one of the best tools that assigns each feature an importance value for a particular decision.

The Shapley value explanation is expressed as a linear model, which is an innovation brought by the approach SHAP. LIME (Local

Interpretable Model-Agnostic Explanations) [69] and Shapley Values are connected by this perspective. Each SHAP value represents the positive or negative contribution of each model feature. Using SHAP value has two important advantages: ❶ SHAP values may be determined for any model, not only basic linear models. ❷ each score has a set of SHAP values unique to it [48]. The output probabilities from a particular set of samples that cover the portion of the input required to be explained are then used to create a linear model using lime. The relevance of input characteristics is then calculated using the weights of the surrogate model. SHAP specifies the explanation for an instance x as:

$$g(z') = \Phi_0 + \sum_{j=1}^{M} \Phi_j z'_j \tag{7}$$

LIME is another well-known tool [69]. Some classifiers employ user-unfriendly representations. Even if it is not the representation utilized by the classifier, Lime describes it in terms of interpretable representations [69]. The LIME model may be expressed as follows:

$$\xi(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \left\{ \mathcal{L}(f, g, w^x) + \omega(g) \right\} \tag{8}$$

where g represents the explanation model for the instance x. $\mathcal{G}$ is the family of possible explanations. $\mathcal{L}$ is the loss function, which is used to measure how close the predictions from the explanation model are to the original model. f represents the original model. w^x defines the weight between the sampled data and the original data. If the sampled data are similar to the original data, then the weight is greater, and vice versa.

$$\phi_i(f, x') = \sum_{z' \subseteq \{x'_1, \ldots, x'_n\} \setminus \{x'_i\}} \frac{|z'|!(M - |z'| - 1)!}{M!} * \left[f\left(z' \cup x'_i\right) - f\left(z'\right) \right], \tag{9}$$

End users may be better able to understand a model's decision if the most essential elements are described. Another tool, *Shapash* [80], aims to make ML more understandable and interpretable for a general audience. It provides a wide range of visuals with properly defined labels that are easy to understand. Shapash also aids data science audits by compiling essential data and model information into a single report. The RF, ET, and LGBM classifiers are used because they are the most accurate and allow for feature importance to be retrieved in terms of both a global explanation and a local explanation. Although SVM and SGD can fit sophisticated nonlinear models to data and achieve excellent performance, the resulting models are opaque, and the explainer could not support such type of classifiers [80]. As a result, for fair comparison, we chose RF, ET, and LGBM classifiers for explainability using SHAP, LIME, and Shapash explainers.

4. Results

Based on single and varied fusions of modalities, we reported the performance of the top five models: RF, LGBM, ET, SVC, and SGD. We tested 10-ML models with single-modality (i.e., SC, BS, MH, M, and NM) and analyzed them for both three-class and four-class contexts. The methods we used for data fusion are averaging and combining data from multiple modalities. The main idea of the used data fusion mechanism is to explore the performance of combining different modalities using and without using an extra feature selection step. We tested the performance of different ML models with each modality using and without using feature select step in the pipeline. Based on the performance of single modalities, we fused the best two modalities and tested the performance of different ML models using and without using feature selection step. This process is continued until we reached the best number of modalities that achieved the best results. The results of the experiment revealed that the motor modality outperformed all other modalities in terms of performance across models. As a result, the fusion of modalities in tandem with motor modality has been reported: ❶ two-modality combinations such as M-NM, M-SC and M-MH, and ❷ three-modality combinations (with the best performance of the M-MH tandem) such as **M-MH-MH** and M-MH-SC.

Several experiments are carried out on the entire features (after the combination of modalities) dataset, along with feature selection for single-modality (early selection) and multi-modality (late selection). For the three instances described, well-known ML algorithms are examined. The experimental results are presented as mean ± SD (Standard Deviation). The performance of each ML classifier is compared based solely on the accuracy metric, since it reflects other measurement matrices. Readers may check the accompanying tables for further information on the other performance measures.

Experiment 1 The experiment is designed to evaluate 10 prominent ML classifiers with default hyperparameters and entire features (n=311). The average performance of the classifiers is listed in Table S1 in the Supplementary file over a 10-holdout for the three-class problem. In attaining 82.9% CV accuracy, the ETC algorithms surpassed all other models. Throughout training, the ETC model (SD = 2.0%) remains stable; however, the RF model achieves 0.68% better accuracy than ETC with the same stability during testing. The Gaussian NB model performs the worst with 56.94% CV accuracy and 57.67% testing accuracy. As presented in Table S2 in the Supplementary file, this pattern has persisted for the four-class problem. The best training accuracy is obtained by the ETC model (85.31%), whereas the maximum testing accuracy is attained by the LGBM classifier (74.88%).

Experiment 2 This experiment investigates the role of feature selection in the training of ML models with default hyperparameters. For feature selection, two strategies are used: ❶ early selection, and ❷ late selection. We examined the four distinct feature selection methods in both approaches by relying on the motor modality alone. Table S3 (for three-class) and Table S4 (for four-class) in the Supplementary file demonstrate the early selection strategy using various feature selection techniques. The recursive XGboost and SULOV approach outperformed the other feature selection strategies in both the three-class and four-class tasks across the ML models. With a 10-holdout test, the ETC ML algorithm achieved a maximum CV accuracy of 86% (SD = 2.0%) and a testing accuracy of 78.2% (SD = 0.0) in the three-class case. Similarly, for the four-class case, the ETC model outscored all other ML models with a CV accuracy of 85.8% (SD = 1.2%), while the LGBM classifier showed the greatest testing accuracy of 74.2% (SD = 1.5%).

On the other hand, the late selection strategy attained a maximum testing accuracy of 79.2% for a three-class experiment utilizing the RF classifier. Table S5 (for three-class) and Table S6 (for four-class) in the Supplementary file demonstrate the late selection strategy using various feature selection techniques. In the four-class test, the RF algorithm has a maximal CV accuracy of 86.8% (SD = 1.0%) and the highest testing accuracy of 78.1% (SD = 0.4%) using the DT classifier. Further, both feature selection strategies outperformed ML models that had been trained and validated on the whole feature set of the dataset as shown in Table 2. These experimental results suggest that the late selection strategy is preferable for attaining maximal testing accuracy. Compared to the entire features, for the three-class problem, the late selection strategy increased testing accuracy by 0.3% (with SD = 0.0), while for the four-class problem, it improved the testing accuracy by 3.22 percent (with SD = 0.0). Therefore, we used the late feature selection strategy for the rest of the experiments by leveraging the recursive XGboost, and SULOV approaches.

Experiment 3 In this experiment we investigated the significance of hyperparameter adjustment in improving the performance of the ML models. This phase is performed after the late feature se-

Table 2
Summary of the recursive XGboost and SULOV approach on the whole feature, early, and late feature selection strategies. The top row in each example represents CV accuracy whereas the bottom row represents testing accuracy.

	Three-class	Four-class
Whole Feature	82.90 ± 0.02 (ETC)	85.31 ± 0.01 (ETC)
	75.90 ± 0.02 (RF)	74.88 ± 0.03 (LGBM)
Early Selection	86.00 ± 0.02 (ETC)	85.80 ± 0.01 (ETC)
	78.20 ± 0.00 (ETC)	74.20 ± 0.01 (LGBM)
Late Selection	85.90 ± 0.01 (ETC)	86.80 ± 0.01 (RF)
	79.20 ± 0.00 (RF)	78.10 ± 0.00 (DT)

lection stage in a manner that depend on the results of the preceding experiment. The Bayesian optimization method is used to tune the hyperparameters of every ML classifier. As can be seen in Table S7 in the Supplementary file for the three-class case, the performance of the ML classifiers has increased significantly. The LGBM classifier attained the maximum CV accuracy of 94.73% (SD = 0.9%), whereas the SVM classifier achieved the best testing accuracy of 87.61% (SD = 0.7%). For the four-classe case, Table S8 in the Supplementary file shows that the LGBM model achieved 91.93% CV accuracy, while SGD achieved the highest testing accuracy of 89.30%. The top five most accurate models (RF, LGBM, ETC, SVC, and SGD) from Table S7 & S8 in the Supplementary file are selected for further investigation in the following experiments.

4.1. Determining the best classifier for single-modality and multi-modality for three-class

We now conduct experiments to determine the optimum feature selection strategy utilizing the motor modality alone. Late selection with recursive XGboost and SULOV method yields the best results in both three-class and four-class problems. The training set adopts SMOTENN imbalance to eliminate model biases, whereas the testing set utilizes random oversampling. In addition, all the models are subjected to Bayesian optimization to enhance their performance. With these settings, we investigate all modalities and their fusion across the ML classifier for the three-class problem.

Single-modality LGBM classifier achieves better training results than the other considered ML-based models. For LGBM, the accuracies for SC, BS, MH, M, and NM are 73.09%, 74.65%, 83.57%, 89.66%, and 90.73%, respectively, as presented in Table 3. Across all ML classifiers, the NM modality outperforms all other modalities (SC, BS, MH, and M). In all the experimental runs, the NM modality obtained the highest CV and testing accuracy in the single-modality test. LGBM had the maximal CV accuracy of all the ML models utilizing the NM modality, at 90.73%, while SVC had the best testing accuracy at 81.94%. Meanwhile, SGD fared poorly, with a minimum CV accuracy of 74.33% and a testing accuracy of 69.48%. The ET model underperformed for the SC modality, whereas SGD scored inadequately for the BS, MH, and M modalities overall. Further, RF is the best model overall with respective testing accuracies of 65.78%, 70.12%, and 72.21% for SC, MH, and M. In terms of BS modality, the ET model has the best testing accuracy of 69.41%. SGD had the lowest accuracies of 62.21%, 65.26%, 63.25%, 65.34% and 69.48% with SC, BS, MH, M, and NM modalities, respectively.

Multli-modality We state that the NM modality plays a crucial role in improving the performance of all ML models. We employed NM in conjunction with other modalities to solve a multi-modality challenge since it surpasses all other modalities and exhibits the best CV training and testing accuracy of all classifiers. We excluded the BS modality from the multi-modality tests since it is non-time series data. We conducted two sets of multi-modality experiments: ❶ two-modality, and ❷ three-modality. When it comes to two-modality, the NM-M combination outperforms all others, as presented in Table 3. Compared to other ML-based models, the LGBM classifier, with NM-M modalities, produces the best training accuracy of 94.89%. However, SVC outperformed LGBM with NM-M modalities, obtaining 10.68% higher testing accuracy. LGBM outperformed the other ML classifiers in other combinations, such as NM-SC and NM-MH, with respective CV accuracies of 93.86% and 91.77%. In the two-modality studies, SGD performs inefficiently.

According to our findings, the NM modality, when combined with the M-modality, improves the performance of all ML algorithms. We used NM-M in conjunction with the other two modalities–MH and SC–to address a three-modality problem. As can be seen in Table 3 in all ML classifiers, combining three modalities reduces the CV and testing accuracy. For example, LGBM attained the highest CV accuracy of 94.89% when combining two modalities (NM-M); however, when combining the MH modality and the SC modality (one at a time), it reduces to roughly 4% and 2% respectively. When merging two modalities (NM-M), SVC scored the highest testing accuracy of 86.10%, but when integrating MH and SC modality (one at a time), it drops to around 4% and 2%, respectively.

Comparison Figure 3 shows a comparison of the top five ML models utilizing the single or multimodality that provides the best CV results, as well as the best accurate classifier using the entire features, early selected features, and optimized late selected features across different modalities and their combinations. Figure 3(a) depicts the performance matrices of five ML models with optimization using only the NM modality. With an accuracy of 90.73% and a recall of 89.53%, the LGBM model performed well. However, the RF model performed better than the LGBM in terms of precision and F1-score. The LGBM model may be seen as having a step increase in performance when compared between entire features, early selected features, and optimized late selected features throughout the modalities, as shown in Fig. 3(b). In general, the model performance increased dramatically when it was switched from using the entire features without optimization to using feature selection with optimization. For example, the LGBM model produced a CV accuracy of 51.36% with entire features using SC modality, which has improved to 73.09% following late feature selection optimization.

In terms of multi-modality, the lowest performance model is SGD, which used the NM-M modality to attain CV accuracy of 78.49%, precision of 81.85%, recall of 78.60%, and F1-score of 74.04%, as shown in Fig. 3(c). When compared to SGD utilizing the NM-M modality, the LGBM model demonstrated enhanced CV model performance with improvements in accuracy of over +16%, precision of over +5.5%, recall of over +15%, and F1-score of over +6.5%. Figure 3(d) illustrates the top ML classifier–i.e., LGBM– to compare the two-modality and three-modality scenarios. The LGBM model performance is improved in any combination, whether it be two-modality or three-modality. When the whole features are used without optimization, the LGBM classifier employing the M-SC modality obtained a CV accuracy of 86.70%. When using a feature selection approach, it improved by 2.5%. When utilizing a late feature selection technique and then optimizing using a Bayesian optimizer, it further improved by 4.6%.

Critical analysis The statistical analysis is conducted to examine the behavior of the NM modality, which is a crucial aspect of enhancing the performance of all ML models. The Friedman test is used to analyze whether the models' median values differed significantly. If the difference in mean rank exceeds the critical distance of 2.728, then the difference between models is considered to be significant. The statistical difference is shown in Fig. 4(a) based on the Friedman test's model median values. LGBM has a mean rank of 1.10 and a median absolute deviation of approximately 3.0. RF while the corresponding values for SVC are 2.0 and 2.5, respectively. However, the mean rank of the ML models ET and SGD is

Table 3
Performance comparison between single-modality and multi-modality ML models for three-class problem. The bold font style is used to emphasize the highest performing ML model. The results of a 10-holdout and 10-fold CV average accuracy and testing accuracy are presented with the Standard Deviation (SD) i.e., mean ± SD. Following the late feature selection strategy, the hyperparameters of ML models are tuned using the Bayesian Optimizer.

Model	Dataset	Cross-validation				Testing			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF	SC	70.37 ± 0.08	72.37 ± 0.08	75.51 ± 0.06	69.13 ± 0.09	65.78 ± 0.02	63.98 ± 0.04	65.78 ± 0.02	59.68 ± 0.03
	BS	72.68 ± 0.10	71.41 ± 0.11	79.04 ± 0.09	71.99 ± 0.10	67.53 ± 0.05	61.65 ± 0.06	65.02 ± 0.05	61.47 ± 0.05
	MH	80.96 ± 0.10	82.3 ± 0.090	78.40 ± 0.09	77.88 ± 0.09	70.12 ± 0.06	69.17 ± 0.06	70.12 ± 0.06	63.54 ± 0.05
	M	83.05 ± 0.12	84.39 ± 0.09	80.51 ± 0.09	79.97 ± 0.09	72.21 ± 0.06	71.26 ± 0.08	72.21 ± 0.06	65.63 ± 0.07
	NM	**88.12 ± 0.15**	**89.46 ± 0.09**	**80.30 ± 0.09**	**85.04 ± 0.09**	**81.61 ± 0.36**	**75.40 ± 0.36**	**81.61 ± 0.36**	**69.77 ± 0.35**
	NM-M	**92.28 ± 0.20**	**93.62 ± 0.13**	**84.46 ± 0.13**	**89.20 ± 0.13**	**85.78 ± 0.40**	**79.56 ± 0.40**	**85.78 ± 0.40**	**73.93 ± 0.39**
	NM-SC	91.25 ± 0.17	92.59 ± 0.11	83.43 ± 0.11	88.17 ± 0.11	84.74 ± 0.38	78.53 ± 0.38	84.74 ± 0.38	72.90 ± 0.37
	NM-MH	89.16 ± 0.17	90.50 ± 0.11	81.34 ± 0.11	86.08 ± 0.11	82.65 ± 0.38	76.44 ± 0.38	82.65 ± 0.38	70.81 ± 0.37
	NM-M-MH	88.23 ± 0.18	85.29 ± 0.13	88.25 ± 0.18	87.23 ± 0.16	86.26 ± 0.18	88.23 ± 0.29	85.23 ± 0.18	89.24 ± 0.19
	NM-M-SC	90.24 ± 0.18	91.58 ± 0.11	82.42 ± 0.11	87.16 ± 0.11	83.74 ± 0.38	77.52 ± 0.38	83.74 ± 0.38	71.89 ± 0.37
LGBM	SC	73.09 ± 0.07	75.75 ± 0.06	78.13 ± 0.05	72.31 ± 0.07	64.75 ± 0.02	60.28 ± 0.06	64.75 ± 0.02	57.75 ± 0.04
	BS	74.65 ± 0.08	78.61 ± 0.07	78.48 ± 0.06	71.43 ± 0.10	66.65 ± 0.05	64.80 ± 0.07	68.71 ± 2.04	60.47 ± 0.03
	MH	83.57 ± 0.09	76.11 ± 0.09	82.37 ± 0.11	74.69 ± 0.12	65.03 ± 0.07	68.84 ± 0.06	65.03 ± 0.05	67.92 ± 0.03
	M	89.66 ± 0.09	78.22 ± 0.11	84.46 ± 0.11	76.78 ± 0.12	71.14 ± 0.09	70.93 ± 0.06	67.14 ± 0.07	70.01 ± 0.03
	NM	**90.73 ± 0.09**	**83.27 ± 0.14**	**89.53 ± 0.11**	**76.59 ± 0.17**	**71.26 ± 0.42**	**80.33 ± 0.41**	**71.26 ± 0.35**	**79.41 ± 0.38**
	NM-M	**94.89 ± 0.13**	**87.43 ± 0.18**	**93.69 ± 0.15**	**80.75 ± 0.21**	**75.42 ± 0.46**	**84.49 ± 0.45**	**75.42 ± 0.39**	**83.57 ± 0.42**
	NM-SC	93.86 ± 0.11	86.40 ± 0.16	92.66 ± 0.13	79.72 ± 0.19	74.39 ± 0.44	83.46 ± 0.43	74.39 ± 0.37	82.54 ± 0.40
	NM-MH	91.77 ± 0.11	84.31 ± 0.16	90.57 ± 0.13	77.63 ± 0.19	72.30 ± 0.44	81.37 ± 0.43	72.30 ± 0.37	80.45 ± 0.40
	NM-M-MH	90.05 ± 0.12	87.40 ± 0.12	91.66 ± 0.15	80.72 ± 0.17	78.39 ± 0.14	86.46 ± 0.23	75.39 ± 0.27	81.44 ± 0.20
	NM-M-SC	92.85 ± 0.11	85.39 ± 0.16	91.65 ± 0.13	78.71 ± 0.19	73.38 ± 0.44	82.45 ± 0.43	73.38 ± 0.37	81.53 ± 0.40
ET	SC	66.72 ± 0.07	69.14 ± 0.08	73.03 ± 0.06	64.86 ± 0.08	65.36 ± 0.03	62.73 ± 0.05	65.36 ± 0.03	60.56 ± 0.04
	BS	66.65 ± 0.08	70.02 ± 0.10	72.88 ± 0.08	67.75 ± 0.09	69.41 ± 0.05	61.01 ± 0.06	67.86 ± 0.03	62.30 ± 0.06
	MH	76.79 ± 0.09	79.06 ± 0.09	75.69 ± 0.09	76.45 ± 0.10	67.72 ± 0.08	65.49 ± 0.08	70.72 ± 0.08	69.03 ± 0.03
	M	78.90 ± 0.09	81.15 ± 0.11	77.78 ± 0.11	78.56 ± 0.10	69.81 ± 0.08	67.58 ± 0.08	72.81 ± 0.08	71.14 ± 0.03
	NM	**83.95 ± 0.14**	**80.96 ± 0.09**	**77.59 ± 0.14**	**78.35 ± 0.10**	**79.21 ± 0.38**	**71.72 ± 0.38**	**76.95 ± 0.43**	**75.26 ± 0.33**
	NM-M	**88.11 ± 0.18**	**85.12 ± 0.13**	**81.75 ± 0.18**	**82.51 ± 0.14**	**83.38 ± 0.42**	**75.89 ± 0.43**	**81.12 ± 0.47**	**79.42 ± 0.38**
	NM-SC	87.08 ± 0.16	84.09 ± 0.11	80.72 ± 0.16	81.48 ± 0.12	82.34 ± 0.40	74.85 ± 0.40	80.08 ± 0.45	78.39 ± 0.35
	NM-MH	84.99 ± 0.16	82.07 ± 0.11	78.63 ± 0.16	79.39 ± 0.12	80.25 ± 0.40	72.76 ± 0.40	77.99 ± 0.45	76.30 ± 0.35
	NM-M-MH	85.23 ± 0.17	84.23 ± 0.14	82.23 ± 0.17	81.23 ± 0.11	83.23 ± 0.24	81.23 ± 0.37	85.20 ± 0.23	79.21 ± 0.19
	NM-M-SC	86.07 ± 0.16	83.08 ± 0.11	79.71 ± 0.16	80.47 ± 0.12	81.34 ± 0.40	73.85 ± 0.41	79.08 ± 0.45	77.38 ± 0.36
SVC	SC	66.60 ± 0.09	67.84 ± 0.10	69.22 ± 0.08	66.65 ± 0.10	59.97 ± 0.04	59.92 ± 0.04	59.97 ± 0.04	59.36 ± 0.04
	BS	69.17 ± 0.11	68.56 ± 0.12	68.73 ± 0.12	69.47 ± 0.13	63.45 ± 0.04	59.26 ± 0.06	61.45 ± 2.06	61.85 ± 0.06
	MH	80.21 ± 0.12	71.84 ± 0.10	76.72 ± 0.12	70.12 ± 0.12	70.45 ± 0.07	70.57 ± 0.06	70.45 ± 0.05	60.66 ± 0.02
	M	86.32 ± 0.12	73.93 ± 0.10	78.83 ± 0.12	72.21 ± 0.12	76.54 ± 0.07	72.66 ± 0.06	72.56 ± 0.05	62.75 ± 0.02
	NM	**87.37 ± 0.12**	**73.74 ± 0.10**	**83.88 ± 0.12**	**77.28 ± 0.17**	**81.94 ± 0.37**	**82.06 ± 0.36**	**81.94 ± 0.40**	**66.89 ± 0.37**
	NM-M	**91.53 ± 0.16**	**77.90 ± 0.14**	**88.04 ± 0.16**	**81.44 ± 0.21**	**86.10 ± 0.41**	**86.22 ± 0.40**	**86.10 ± 0.44**	**71.06 ± 0.41**
	NM-SC	90.50 ± 0.14	76.87 ± 0.12	87.01 ± 0.14	80.41 ± 0.19	85.07 ± 0.39	85.19 ± 0.38	85.07 ± 0.42	70.02 ± 0.39
	NM-MH	88.41 ± 0.14	74.78 ± 0.12	84.52 ± 0.14	78.32 ± 0.19	82.98 ± 0.39	83.10 ± 0.38	82.98 ± 0.42	67.93 ± 0.39
	NM-M-MH	89.10 ± 0.19	76.05 ± 0.14	85.04 ± 0.13	80.31 ± 0.14	82.22 ± 0.39	81.06 ± 0.32	85.90 ± 0.39	64.06 ± 0.39
	NM-M-SC	89.49 ± 0.14	75.86 ± 0.12	86.05 ± 0.14	79.40 ± 0.19	84.06 ± 0.39	84.18 ± 0.38	84.06 ± 0.42	69.02 ± 0.39
SGD	SC	65.18 ± 0.09	66.26 ± 0.10	69.13 ± 0.07	64.37 ± 0.09	62.21 ± 0.03	61.14 ± 0.03	62.31 ± 0.03	60.06 ± 0.03
	BS	65.25 ± 0.11	66.20 ± 0.11	69.99 ± 0.10	64.54 ± 0.12	65.26 ± 0.05	61.44 ± 0.07	63.71 ± 0.05	61.55 ± 0.07
	MH	67.17 ± 0.10	70.53 ± 0.09	67.28 ± 0.11	62.72 ± 0.10	63.25 ± 0.06	62.27 ± 0.06	63.25 ± 0.06	61.76 ± 0.03
	M	69.26 ± 0.10	72.62 ± 0.11	69.37 ± 0.11	64.81 ± 0.10	65.34 ± 0.08	64.36 ± 0.06	65.34 ± 0.06	63.85 ± 0.03
	NM	**74.33 ± 0.10**	**77.69 ± 0.09**	**74.44 ± 0.11**	**72.88 ± 0.10**	**69.48 ± 0.41**	**73.76 ± 0.06**	**69.48 ± 0.41**	**67.99 ± 0.33**
	NM-M	**78.49 ± 0.14**	**81.85 ± 0.13**	**78.60 ± 0.15**	**74.04 ± 0.14**	**73.64 ± 0.45**	**77.92 ± 0.10**	**73.64 ± 0.45**	**72.15 ± 0.37**
	NM-SC	77.46 ± 0.12	80.82 ± 0.11	77.57 ± 0.13	73.01 ± 0.12	72.61 ± 0.43	76.89 ± 0.08	72.61 ± 0.43	71.12 ± 0.35
	NM-MH	75.37 ± 0.12	78.73 ± 0.11	75.48 ± 0.13	70.92 ± 0.12	70.52 ± 0.43	74.80 ± 0.08	70.52 ± 0.43	69.03 ± 0.35
	NM-M-MH	84.06 ± 0.39	82.03 ± 0.39	85.013 ± 0.39	82.13 ± 0.39	70.66 ± 0.31	73.6 ± 0.44	70.61 ± 0.13	71.63 ± 0.43
	NM-M-SC	76.45 ± 0.12	79.81 ± 0.11	76.56 ± 0.13	72.03 ± 0.12	71.68 ± 0.43	75.88 ± 0.08	71.60 ± 0.43	70.11 ± 0.35

3.90 and 5.0, respectively, which are both higher than the critical distance (2.728). Therefore, the median values for the NM modality are significantly different for the models. The multimodality–the NM-M modality in this example–is subjected to the same analysis. LGBM (rank = 1.0) and RF (rank = 2.20) have lower mean ranks than the critical distance, as illustrated in Fig. 4(b). Other models have higher ranks than the critical distance, such as SVC (rank = 2.80), ET (rank = 4.0), and SGD (rank = 5.0).

Explainability Only models with a balance of accuracy and explainability should be used in the medical industry. The ML models should address the following queries by providing enough information about the link between the input feature to the model and the final predictions. Which features are the most important in the prediction? Why are the patients classified as a specific class/category on the medical records? Consequently, the study's inclusion of explainability in the methodology seems to be more intuitive and clinically acceptable. Due to the low explainability of the existing methods, PD diagnosis prediction remains an extremely difficult problem despite the performance improvements that extant studies have achieved [66]. In the present work we have implemented complex black-box models such as ET, LGBM, and RF. We used both global and local explainers to demonstrate how to provide the explainability of the ML models.

Three explainers, SHAPASH, SHAP, and LIME, are used to provide a pool of potential explanations for the model's results. Every explanation comes from a separate explainer that estimates the high relevance score to the most notable features to various modalities and returns with additional information regarding the explainer's trustworthiness in terms of accuracy. From various perspectives, including visualization, and feature significance (local and global), the explainers provide insights into the driving elements of the ML model's decision.

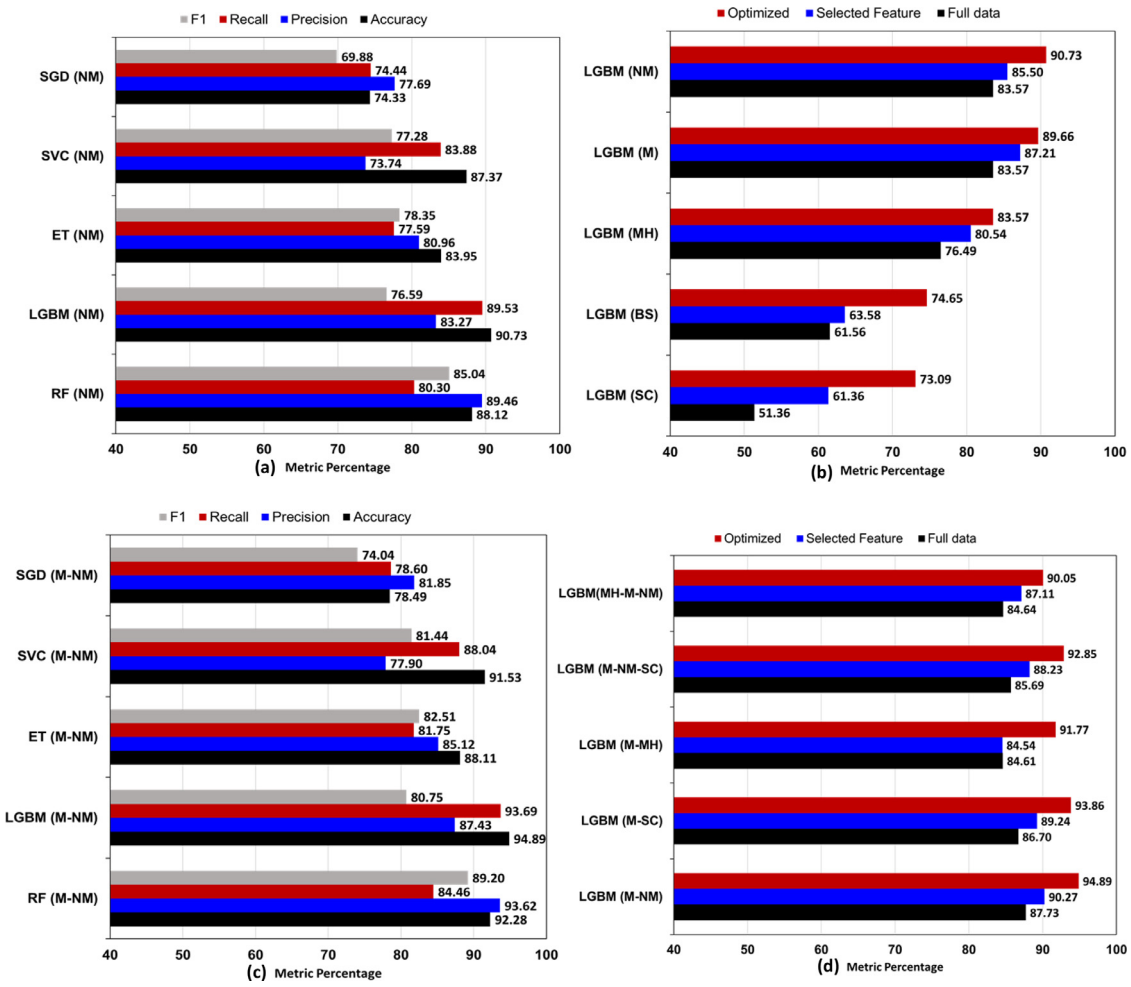

Fig. 3. Performance of the top five ML classifiers for the three-class problem. (a) comparison of the NM modality between the ML models after late feature selection optimization. (b) The best model (LGBM) classifier is compared between the different modality when whole features are selected (black), an early feature selection criterion is applied (blue), and late feature selection is applied followed by Bayesian optimization (red). (c) Two-modality (NM-M) is compared between the ML models after late feature selection optimization. (d) The best model (LGBM) classifier is compared between the combination of two-modality and three-modality when whole features are selected (black), an early feature selection criteria is implemented (blue), and late feature selection is applied followed by Bayesian optimization (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

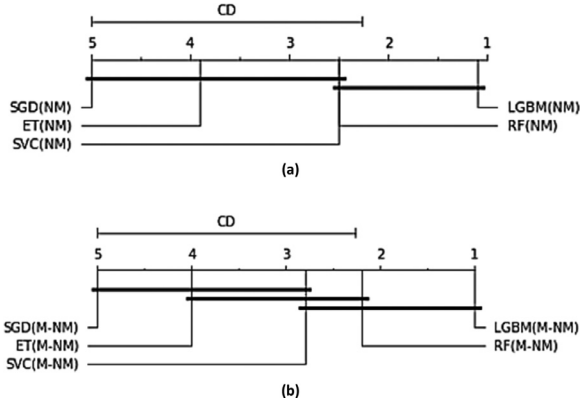

Fig. 4. Critical diagram for the top five ML classifier based on the Friedman test for three-class classification problem using the NM modality. The critical diagram compares different models for: (a) the single-modality and (b) the multi-modality.

❶ *ETC Model:* The feature importance plot is created by stacking the influence of a feature on the classes. Figure 5 shows a comparison between global and local feature importance. SHAP is used to provide a global explanation of the classes. As shown in Fig. 5(a), **NP3BRADY** is the crucial feature for predicting if a patient will be healthy or at an early stage, whereas it has less significance for predicting advanced stages. In addition, SHAP ranks the features according to their importance. In the rank, NP3PRSPR, NP3TTAPR, and NP3FTAPR all include important features.

We need to assess the significance of features for individual scenarios after examining the global feature importance of the employed feature set (i.e., local feature importance). We used SHAPASH and LIME for the local feature importance. SHAPASH identifies the features that are most important in predicting class 2 (i.e., early-stage), and these are illustrated in Fig. 5(b). According to SHAPASH, the top features are NP3BRADY and NP3PRSPR, which is consistent with the SHAP explainer's global feature ranking to predict PD early stage. Further, the LIME explainer assigns a low value to features in predicting the advanced stage, similar to how SHAP assigns a low value to features in predicting the advanced stage globally, as shown in Fig. 5(c). The consistency of SHAP's global and local explanation of LIME and SHAPASH in predicting the class is demonstrated.

❷ *LGBM Model:* In contrast to ETC, the global feature importance plot generated by SHAP via the LGBM ML technique shows a substantial difference as illustrated in Fig. 6(a). SHAP, like ETC, shows that the **NP3BRADY** feature has the highest ranking. Meanwhile, the NP3BRADY feature contribution is better for predicting

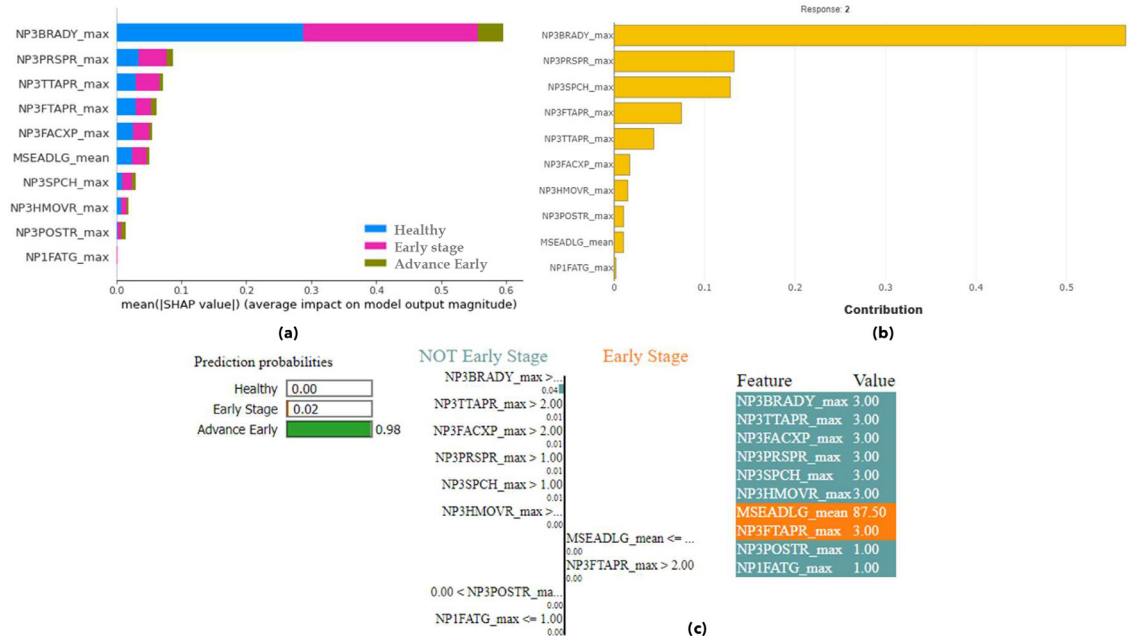

Fig. 5. The explainability of the ET ML classifier for predicting the three-class case of PD by (a) SHAP -which shows the global feature importance by the class; (b) SHAPASH, which shows the local feature importance by a single instance; and (c) LIME, which showed the local feature importance by a single instance.

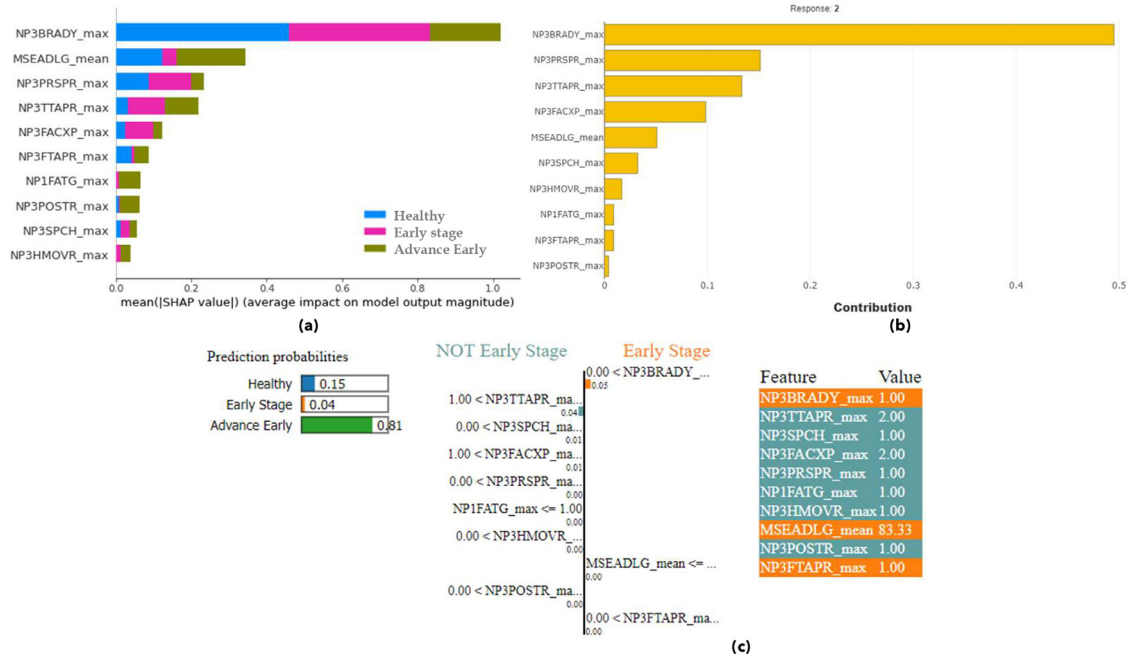

Fig. 6. Explainability of the LGBM ML classifier for predicting the three-class case of PD by (a) SHAP, which shows the global feature importance by the class; (b) SHAPASH, which shows the local feature importance by a single instance; and (c) LIME, which shows the local feature importance by a single instance.

the advanced early stage than predicting healthy or early patients. The next-best feature, MSEADLG, predicts the advanced stage better than the healthy or early stages. In addition, the NP1FATG and NP3POSTR features have a significant influence on determining the advanced stage in the rank list.

In terms of the local explanation, SHAPASH and LIME have the same consistency features. SHAPASH identifies the features that are most important in predicting class 2 which are shown in Fig. 6(b) (i.e., early-stage). According to SHAPASH, the most important features are **NP3BRADY** and **NP3PRSPR** with the maximum statistical features in the time-series data. Relative to the model, ETC and LGBM list the same top features when utilizing SHAPASH. When comparing explainers within the same classifier, SHAP and SHPASH have **NP3BRADY** in common. Further, the LIME explainer assigns a slightly higher value to features in predicting the advanced stage, similar to how SHAP assigns high weight to features in predicting the advanced stage as shown in Fig. 6(c). SHAP, LIME, and SHAPASH are shown to be consistent in predicting the class across the models.

❸ *RF Model:* The SHAP, SHPASH, and LIME explanations are compared to the preceding two models (ETC and LGBM) with the RF model. NP3BRADY, like ETC and LGBM, is at the top of the SHAP explainer's global feature importance list in predicting the three-class PD patients as shown in Fig. 7(a). The NP3BRADY feature

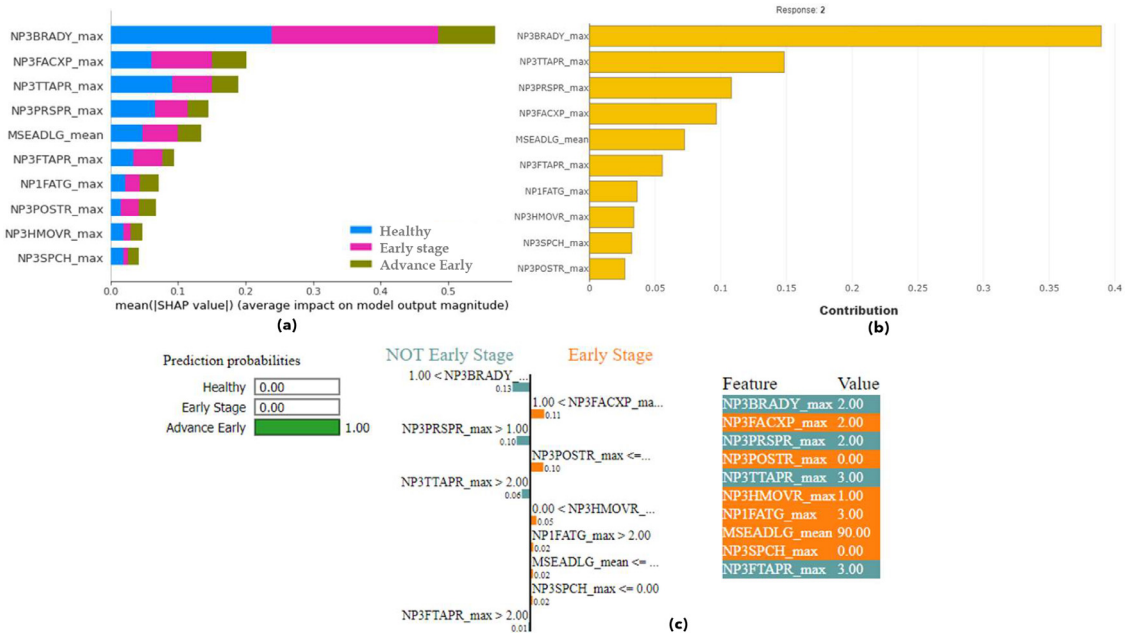

Fig. 7. Explainability of the RF ML classifier for predicting the three-class case of PD by (a) SHAP which shows the global feature importance by the class; (b) SHAPASH which shows the local feature importance by a single instance; and (c) LIME which the local feature importance by a single instance.

helps predict healthy and early-stage disease; however, the feature does play a significant role in advanced stage prediction. SHAP for RF models, unlike ETC and LGBM, lists the NP3FACXP feature as the second most important feature.

For the local explanation of the models, SHPASH and LIME explainers are used for the RF model in the same way as ETC and LGBM for fair feature comparison. The RF model is provided with an example of the early-stage label (response 2). The SHAPASH explainer generates the list of important features that contains NP3BRADY with the maximum statistical attribute as shown in Fig. 7(b). The features NP3TTAPR and NP3PRSPR respectively contribute 15% and 10%. Meanwhile, the LIME explanation illustrates the feature significance score with the probability of prediction. As shown in Fig. 7(c), LIME prediction exhibits a 100% probability of correctly classifying the input instance. The NP3BRADY feature makes the largest contribution to the prediction. following the investigation, it is determined that NP3BRADY is the most important feature in predicting the correct class, as is tested on three models (ETC, LGBM, and RF) and assessed using three distinct explainers (SHAP, SHAPASH, and LIME), including global and local explainers.

4.2. Determining the best classifier for single-modality and multi-modality for four-class

case Similar experiments are conducted for the four-class issue. We study all modalities and their fusion across the ML classifier for the four-class scenario using identical parameters as those in the previous section.

Single-modality We first discuss the performance of ML algorithms for predicting PD based on each of the five single modalities. Compared to other ML-based models, the RF classifier produces the best training results. As presented in Table 4, the RF accuracies for SC, BS, MH, M, and NM are 72.68%, 87.05%, 88.27%, 74.81%, and 94.57%, respectively. The NM modality performs better than all other modalities (SC, BS, MH, and M) across all ML classifiers. The NM modality had the greatest CV and testing accuracies in the single-modality test in all the experimental runs. SVC had the greatest testing accuracy at 80.10%, whereas RF had the highest CV accuracy of all the ML models using the NM modality, at 94.57%. Meanwhile, SGD performed poorly, with a minimum CV accuracy of 79.26% and the ET classifier scored a testing accuracy of 72.48%. Overall, the ET model fared poorly for the SC modality, while SGD performed poorly for the BS, MH, and M modalities. Further, with a test accuracy of 65.78% using the SC modality, RF is the best model overall. The SVC model has the greatest testing accuracy in the BS and MH modalities with respective values of 73.76% and 73.80%,. The LGBM model offers the greatest testing accuracy of 69.20% for the M modality. With the SC modality, SVC had the lowest testing accuracy of 63.45%, while ET had the minimum testing accuracy of 66.18% with the MH modality. Lastly, SGD had the smallest testing accuracies with the BS and M modalities, at 68.37% and 59.25%, respectively.

Multi-modality The NM modality, as can be seen in the experimental results, plays a critical role in boosting the performance of all ML models. Because NM outperforms all other modalities and has the highest CV training and testing accuracy of any classifier, we used it in combination with other modalities to address the multi-modality problem. The BS modality was eliminated from the multi-modality testing because it is non-time series data. Two types of multi-modality tests are conducted: ❶ two-modality and ❷ three-modality. Table 4 shows that, aside from the LGBM classifier, the NM-SC combination surpasses all others in terms of two-modality. The NM-M combination outperformed the other combinations in the LGBM model. The LGBM classifier with NM-M modalities offers the best training accuracy of 93.37%, outperforming all other considered ML-based models and NM-SC combinations. With NM-SC modalities, SVC surpassed RF and LGBM (with NM-M modalities), thus achieving 0.37% greater testing accuracy than RF and 12.5% greater testing accuracy than LGBM. With a CV accuracy of 90.62%, LGBM surpassed all other ML classifiers in NM-MH combinations. SGD showed overall poor performance in the two-modality investigations.

According to our results, when the NM modality is used in conjunction with the M-modality, the LGBM algorithms perform better. When NM and SC are combined, all other ML-based classifiers perform better. We employed NM-M and NM-SC combined with the other two modalities of MH and M, to handle a three-modality issue for the individual classifier. For example, since NM-SC im-

Table 4
Performance comparison between single-modality and multi-modality ML models for four-class problem. The highest performing ML model each case is bolded. The results of a 10-holdout and 10-fold CV average accuracy and testing accuracy are presented with the Standard Deviation (SD), i.e., mean ± SD. Following the late feature selection strategy, the hyperparameters of ML models are tuned using the Bayesian Optimizer.

Model	Dataset	Cross-validation					Testing			
		Accuracy	Precision	Recall	F1		Accuracy	Precision	Recall	F1
RF	SC	72.68 ± 0.10	71.41 ± 0.11	79.04 ± 0.09	71.99 ± 0.10	~	67.53 ± 0.05	61.65 ± 0.06	65.02 ± 0.05	61.47 ± 0.05
	BS	87.05 ± 0.06	84.93 ± 0.05	85.35 ± 0.05	82.84 ± 0.06		72.08 ± 0.05	72.02 ± 0.05	72.08 ± 0.05	69.97 ± 0.05
	MH	88.27 ± 0.10	84.09 ± 0.09	86.28 ± 0.09	83.76 ± 0.09		72.07 ± 0.08	71.31 ± 0.09	72.07 ± 0.08	70.28 ± 0.08
	M	74.81 ± 0.07	71.79 ± 0.07	73.55 ± 0.06	69.76 ± 0.08		65.77 ± 0.03	65.54 ± 0.04	65.77 ± 0.03	63.92 ± 0.03
	NM	**94.57 ± 6.40**	**90.39 ± 6.39**	**92.58 ± 6.39**	**90.06 ± 6.39**		**78.37 ± 6.38**	**77.61 ± 6.39**	**78.37 ± 6.38**	**76.58 ± 6.38**
	NM-M	90.76 ± 0.18	92.10 ± 0.11	82.94 ± 0.11	87.68 ± 0.11		84.26 ± 0.38	78.04 ± 0.38	84.26 ± 0.38	72.41 ± 0.37
	NM-SC	**91.58 ± 0.20**	**90.92 ± 0.14**	**84.76 ± 0.14**	**86.50 ± 0.14**		**86.07 ± 0.41**	**79.86 ± 0.41**	**86.07 ± 0.41**	**74.23 ± 0.40**
	NM-MH	88.01 ± 0.22	89.35 ± 0.16	80.19 ± 0.16	84.93 ± 0.16		81.50 ± 0.43	75.29 ± 0.43	81.50 ± 0.43	69.66 ± 0.42
	NM-SC-MH	88.01 ± 0.22	85.03 ± 0.08	88.01 ± 0.22	85.03 ± 0.08		75.39 ± 0.35	69.76 ± 0.34	75.39 ± 0.35	69.76 ± 0.34
	NM-SC-M	88.11 ± 0.15	89.45 ± 0.08	80.29 ± 0.08	85.03 ± 0.08		81.61 ± 0.35	75.39 ± 0.35	81.61 ± 0.35	69.76 ± 0.34
LGBM	SC	74.65 ± 0.08	78.61 ± 0.07	78.48 ± 0.06	71.43 ± 0.10		66.65 ± 0.05	64.8 ± 0.07	68.71 ± 2.04	60.47 ± 0.03
	BS	81.55 ± 0.04	79.30 ± 0.05	80.28 ± 0.04	77.38 ± 0.05		72.39 ± 0.02	71.96 ± 0.03	72.39 ± 0.02	71.27 ± 0.03
	MH	86.32 ± 0.11	83.11 ± 0.11	85.36 ± 0.11	81.04 ± 0.12		70.85 ± 0.05	69.91 ± 0.06	70.85 ± 0.05	68.55 ± 0.06
	M	82.16 ± 0.07	77.52 ± 0.07	80.28 ± 0.07	77.30 ± 0.08		69.20 ± 0.04	67.72 ± 0.05	68.14 ± 0.05	67.84 ± 0.05
	NM	**92.62 ± 6.41**	**89.41 ± 6.41**	**91.66 ± 6.41**	**87.34 ± 6.42**		**77.15 ± 6.35**	**76.21 ± 6.36**	**77.15 ± 6.35**	**74.85 ± 6.36**
	NM-M	**93.37 ± 0.11**	**85.91 ± 0.16**	**92.17 ± 0.13**	**79.23 ± 0.19**		**73.90 ± 0.44**	**82.97 ± 0.43**	**73.9 ± 0.37**	**82.05 ± 0.40**
	NM-SC	91.19 ± 0.14	87.73 ± 0.19	93.99 ± 0.16	81.05 ± 0.22		75.72 ± 0.47	84.79 ± 0.46	75.72 ± 0.40	83.87 ± 0.43
	NM-MH	90.62 ± 0.16	83.16 ± 0.21	89.42 ± 0.18	76.48 ± 0.24		71.15 ± 0.49	80.22 ± 0.48	71.15 ± 0.42	79.30 ± 0.45
	NM-M-MH	89.96 ± 0.18	85.03 ± 0.08	87.32 ± 0.11	85.02 ± 0.28		73.20 ± 0.57	79.22 ± 0.27	85.03 ± 0.08	73.23 ± 0.44
	NM-M-SC	90.72 ± 0.08	85.26 ± 0.13	89.52 ± 0.19	76.58 ± 0.16		71.25 ± 0.41	80.32 ± 0.4	71.25 ± 0.34	79.40 ± 0.37
ET	SC	62.98 ± 0.12	64.23 ± 0.11	69.55 ± 0.10	62.35 ± 0.11		64.59 ± 0.05	63.82 ± 0.06	66.14 ± 0.06	60.72 ± 2.06
	BS	78.33 ± 0.07	75.49 ± 0.06	77.16 ± 0.05	73.88 ± 0.06		70.56 ± 0.06	70.16 ± 0.06	70.56 ± 0.06	69.20 ± 0.05
	MH	76.37 ± 0.12	73.90 ± 0.12	74.24 ± 0.11	71.63 ± 0.12		66.18 ± 0.06	64.72 ± 0.06	66.18 ± 0.06	64.88 ± 0.06
	M	76.18 ± 0.07	72.35 ± 0.08	74.10 ± 0.07	70.21 ± 0.08		63.02 ± 0.03	62.43 ± 0.04	63.02 ± 0.03	61.97 ± 0.03
	NM	**82.67 ± 0.42**	**80.20 ± 0.42**	**80.54 ± 0.41**	**77.93 ± 0.42**		**72.48 ± 0.36**	**71.02 ± 0.36**	**72.48 ± 0.36**	**71.18 ± 0.36**
	NM-M	87.21 ± 0.12	70.92 ± 0.14	84.86 ± 0.11	77.50 ± 0.14		72.14 ± 0.38	68.52 ± 0.39	77.40 ± 0.38	80.82 ± 0.37
	NM-SC	**89.03 ± 0.15**	**72.73 ± 0.16**	**86.68 ± 0.14**	**79.32 ± 0.17**		**73.96 ± 0.41**	**70.34 ± 0.42**	**79.22 ± 0.41**	**82.63 ± 0.40**
	NM-MH	84.46 ± 0.17	68.16 ± 0.18	82.11 ± 0.16	74.75 ± 0.19		69.39 ± 0.43	65.77 ± 0.44	74.65 ± 0.43	78.06 ± 0.42
	NM-SC-MH	85.23 ± 0.18	85.03 ± 0.08	85.03 ± 0.08	76.74 ± 0.18		85.03 ± 0.18	73.74 ± 0.38	85.03 ± 0.28	79.74 ± 0.38
	NM-SC-M	86.56 ± 0.09	68.27 ± 0.11	82.21 ± 0.08	74.85 ± 0.11		69.49 ± 0.35	65.87 ± 0.36	74.75 ± 0.35	78.17 ± 0.34
SVC	SC	69.17 ± 0.11	68.56 ± 0.12	68.73 ± 0.12	69.47 ± 0.13		63.45 ± 0.04	59.26 ± 0.06	61.45 ± 2.06	61.85 ± 0.06
	BS	79.56 ± 0.07	76.80 ± 0.06	76.28 ± 0.05	73.97 ± 0.07		73.76 ± 0.80	73.21 ± 0.01	73.76 ± 0.08	72.03 ± 0.05
	MH	78.75 ± 0.12	76.96 ± 0.11	76.85 ± 0.10	73.91 ± 0.11		73.80 ± 0.05	75.12 ± 0.05	73.80 ± 0.05	72.03 ± 0.06
	M	76.09 ± 0.08	71.31 ± 0.08	73.74 ± 0.08	69.84 ± 0.09		62.69 ± 0.04	60.76 ± 0.04	63.21 ± 0.04	60.93 ± 0.04
	NM	**85.05 ± 6.42**	**83.26 ± 6.41**	**83.15 ± 6.40**	**80.21 ± 6.41**		**80.10 ± 0.35**	**81.42 ± 6.35**	**80.10 ± 6.35**	**78.33 ± 6.36**
	NM-M	90.01 ± 0.14	76.38 ± 0.12	86.52 ± 0.14	79.92 ± 0.19		84.58 ± 0.39	84.70 ± 0.38	84.58 ± 0.42	69.54 ± 0.39
	NM-SC	**91.83 ± 0.17**	**78.20 ± 0.15**	**88.34 ± 0.17**	**81.74 ± 0.22**		**86.40 ± 0.42**	**86.52 ± 0.41**	**86.40 ± 0.45**	**71.35 ± 0.42**
	NM-MH	87.26 ± 0.19	73.63 ± 0.17	83.77 ± 0.19	77.17 ± 0.24		81.83 ± 0.44	81.95 ± 0.43	81.83 ± 0.47	66.78 ± 0.44
	NM-SC-MH	85.03 ± 0.08	73.21 ± 0.01	81.17 ± 0.19	84.29 ± 0.08		83.27 ± 0.26	83.21 ± 0.01	85.13 ± 0.09	68.53 ± 0.35
	NM-SC-M	89.36 ± 0.11	73.73 ± 0.09	83.87 ± 0.11	77.27 ± 0.16		81.93 ± 0.36	82.05 ± 0.35	81.93 ± 0.39	66.89 ± 0.36
SGD	SC	65.25 ± 0.11	66.20 ± 0.11	69.99 ± 0.10	64.54 ± 0.12		65.26 ± 0.05	61.44 ± 0.07	63.71 ± 0.05	61.55 ± 0.07
	BS	73.35 ± 0.09	71.64 ± 0.10	72.82 ± 0.08	68.18 ± 0.09		68.37 ± 0.02	67.29 ± 0.01	68.37 ± 0.02	66.22 ± 0.01
	MH	72.96 ± 0.10	71.41 ± 0.10	73.14 ± 0.08	68.47 ± 0.11		71.92 ± 0.06	71.45 ± 0.06	71.92 ± 0.06	71.1 ± 0.060
	M	71.26 ± 0.08	67.78 ± 0.09	67.09 ± 0.08	64.13 ± 0.09		59.25 ± 0.04	58.74 ± 0.05	59.84 ± 0.04	56.35 ± 0.04
	NM	**79.26 ± 6.40**	**77.71 ± 6.40**	**79.44 ± 6.38**	**74.77 ± 6.41**		**78.22 ± 6.36**	**77.75 ± 6.36**	**78.22 ± 6.36**	**77.4 ± 6.360**
	NM-M	76.97 ± 0.12	80.33 ± 0.11	77.08 ± 0.13	72.52 ± 0.12		72.12 ± 0.43	76.40 ± 0.08	72.12 ± 0.43	70.63 ± 0.35
	NM-SC	**78.79 ± 0.15**	**82.15 ± 0.14**	**78.90 ± 0.16**	**74.34 ± 0.15**		**73.94 ± 0.46**	**78.22 ± 0.11**	**73.94 ± 0.46**	**72.45 ± 0.38**
	NM-MH	74.22 ± 0.17	77.58 ± 0.16	74.33 ± 0.18	69.77 ± 0.17		69.37 ± 0.48	73.65 ± 0.13	69.37 ± 0.48	67.88 ± 0.40
	NM-SC-MH	77.98 ± 0.32	75.03 ± 0.08	67.98 ± 0.32	67.98 ± 0.32		71.03 ± 0.08	69.87 ± 0.09	67.98 ± 0.32	69.87 ± 0.09
	NM-SC-M	74.32 ± 0.09	77.68 ± 0.08	74.43 ± 0.10	69.86 ± 0.09		69.47 ± 0.40	73.75 ± 0.05	69.47 ± 0.40	67.98 ± 0.32

proves the performance of the RF model, we combined additional modalities with the NM-SC combination. Combining three modalities decreases the CV and testing accuracies in all ML classifiers, as presented in Table 4. For example, when combining two modalities (NM-SC), SVC achieved the best CV accuracy of 91.83%. Still, combining MH and M modality (one at a time) drops to respective values of about 7% and 2.5%. SVC obtained the maximum testing accuracy of 86.40% when combining two modalities (NM-SC), but when incorporating MH and M modality (one at a time), it reduces to roughly 3% and 4.5%, respectively. As a result, we did not investigate the four-modality combination in our study.

Comparison Figure 8 shows a comparison of the top five ML models utilizing the single or multimodality that provides the best CV results, as well as the best accurate classifier using the whole feature, early selected features, and optimized late selected features across different modality and their combination. 8(a) shows the performance matrices of the five ML models optimized solely using the NM modality. The RF model performed well, scoring 94.57% accuracy, 90.39% precision, 92.58% recall, and 90.06% F1-score. When comparing entire features, early selected features, and optimized late selected features across all modalities, the LGBM and RF models both demonstrate significant performance improvements, as shown in Fig. 8(b). The model's performance substantially improved when switching from employing entire features and no optimization to feature selection with optimization. With entire features and the NM modality, the LGBM model achieved a CV accuracy of 80.11%, which was increased to 94.57% after late features selection optimization.

The model with the lowest multi-modality performance is SGD, which achieved CV accuracy of 78.79%, precision of 82.15%, recall of 78.90%, and F1-score of 74.34% using the M-SC modality, as shown in Fig. 8(c). Compared to SGD using the M-SC modality, the RF

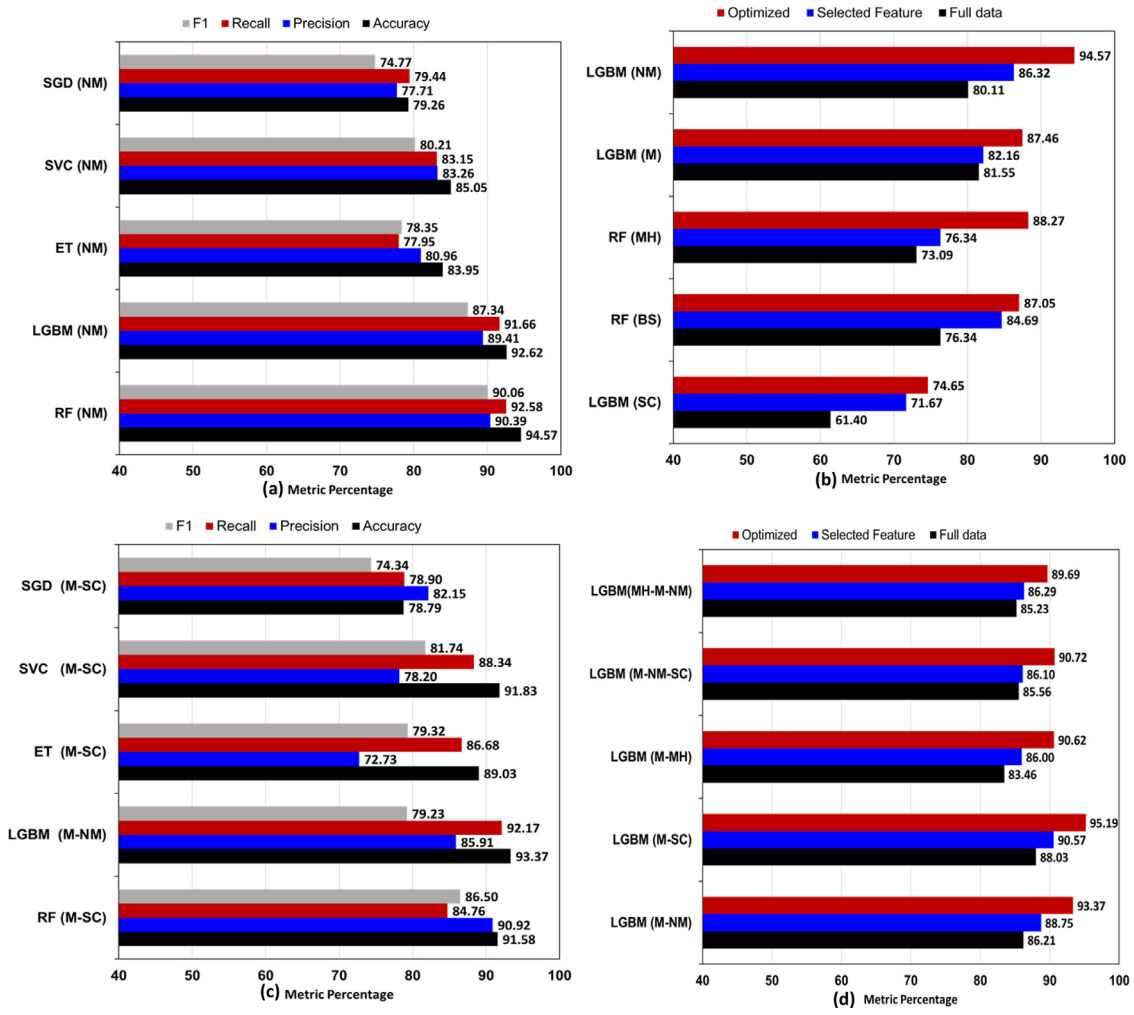

Fig. 8. Performance of the top five ML classifiers for the four-class problem. (a) Comparison of the NM modality between the ML models after late feature selection optimization. (b) Comparison of the best model (LGBM) classifier between the different modality when whole features selected (black), an early feature selection criterion is applied (blue), and late feature selection is applied followed by Bayesian optimization (red). (c) Two-modality (NM-M) is compared between the ML models after late feature selection optimization. (d) The best model (LGBM) classifier is compared between the combination of two-modality and three-modality when whole features are selected (black), an early feature selection criteria is implemented (blue), and late feature selection is applied followed by Bayesian optimization (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model outperformed the CV model with accuracy of over 91.00%, precision of over 92%, recall of approximately 85%, and F1-score of over 56%. Figure 8(d) compares the two-modality and three-modality cases using the top ML classifier with NM-M modality, i.e., LGBM. The performance of the LGBM model improves in every combination, be it a two-modality or three-modality combination. The LGBM classifier using the M-SC modality produced a CV accuracy of 88.03% when all features were employed without optimization. It increased by 2.54% when employing a feature selection strategy, and it increased by 4.62% after employing a late feature selection strategy and then optimizing using a Bayesian optimizer.

Critical analysis A statistical assessment is conducted to investigate the behavior of the NM modality, which is critical for improving the performance of all ML models. The Friedman test is used to determine whether the median values of the models varied considerably. The difference between models becomes significant when the difference in mean rank exceeds the critical distance of 2.728. In terms of the model's median Friedman test values, the statistical difference is demonstrated in Fig. 9(a). The mean ranks of RF and LGBM are 1.10 and 1.9, respectively. By contrast, the ML models ET, SVC, and SGD, have respective mean ranks of 3.10, 3.90, and 5.0, all of which exceed the critical distance (2.728). As a result, the NM modality's median values vary greatly among models. The same analysis is performed for multimodality, in this case, the NM-SC modality. As shown in Fig. 9(b), the mean ranks of LGBM

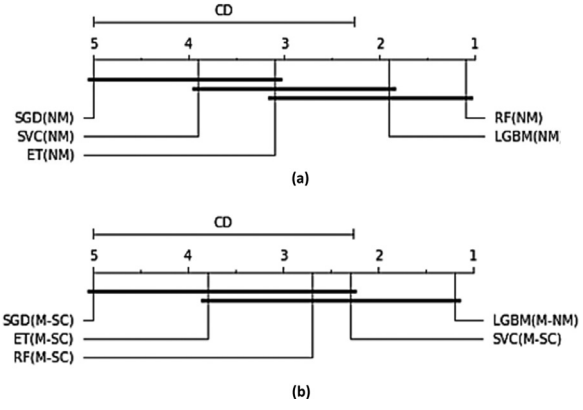

Fig. 9. Critical diagram for the top five ML classifiers based on the Friedman test for the four-class problem using the NM modality. The critical diagram compares different models for the (a) single-modality and (b) multi-modality.

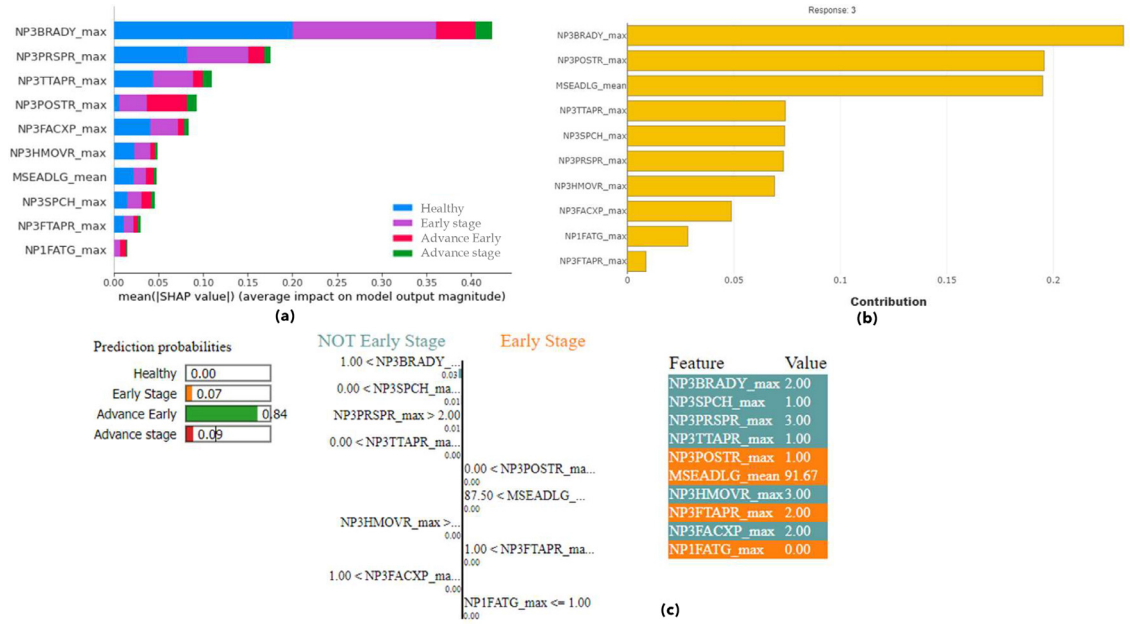

Fig. 10. Explainability of the ET ML classifier for predicting the four-class case of PD by (a) SHAP, which shows the global feature importance by the class; (b) SHAPASH, which showsthe local feature importance by a single instance; and (c) LIME, which shows the local feature importance by a single instance.

(rank = 1.2), SVC (rank = 2.3), and RF (rank = 2.70) are lower than the crucial distance. ET (rank = 3.80), ET (rank = 4.0), and SGD (rank = 5.0) are examples of models with a greater rank than the critical distance.

Explainability In the medical realm, it is inadequate to simply provide a meaningful outcome and the highest model accuracy. Doctors want to know why certain judgments were made. These lead us to the important questions: What is the rationale behind the system's decision? What considerations or parameters are used in making the decision? Are the selected attributes medically important and sufficient? This subsection discusses the implementation of XAI features for the four-class problem in the present research. XAI capabilities may be delivered to the physician in a variety of ways, including feature significance based on global and local explainers. Similar to the three-class problem, we used the three models to test various explainers.

❶ *ETC Model:* The explainer constructed the feature significance plot by stacking the effect of a feature on the classes. Figure 10 illustrates the importance of global (SHAP) and local (SHAPASH and LIME) features. SHAP provides a global understanding of the prediction of the correct class. As demonstrated in Fig. 10(a), NP3BRADY is a critically important feature for determining whether a patient is healthy or in an early stage of disease, but it is less important for determining if the patient is in an advanced early or an advanced stage. SHAP also assigns a weighting to the features based on their relevance. The other key features are NP3PRSPR, NP3TTAPR, and NP3FTAPR, and these are included in the rank list.

After reviewing the global feature relevance of the used feature set, we must analyze the significance of features for different situations (i.e., local feature importance). For the relevance of local features, we employed SHAPASH and LIME. As shown in Fig. 10(b), SHAPASH indicates the features that are most relevant in predicting class 3. (i.e., advanced early stage). The topmost features, according to SHAPASH, are NP3BRADY and NP3POSTR (with maximum statistical attributes), which match the SHAP explainer's feature ranking for predicting the advanced early class of PD patients. Further, similar to how SHAP assigns low importance scores to features in predicting the advanced early and advanced stage globally, the LIME explanation assigns low values to features in the predicting task of the advanced early and advanced stage of PD patients, as shown in Fig. 10(c). The highest score feature, NP3BRADY, is the same as in SHAP and SHAPASH. SHAP's global and local explanations of LIME and SHAPASH in predicting the correct class of Parkinson's patients are shown to be consistent.

❷ *LGBM Model:* As shown in Fig. 11(a), the global feature importance plot created by SHAP through the LGBM ML approach, in contrast to ETC, reveals a significant difference. SHAP, like ETC, presents the highest-ranking NP3BRADY feature. On the other hand, the NP3BRADY feature contribution is better for predicting healthy or early patients than advanced early or advanced stage. MSEADLG, the next-best feature, predicts the advanced early and advanced stage better than the healthy and early-stage patients. Further, the NP3POSTR and NP3SPCH features have a considerable impact on deciding the advanced stage PD patient in the SHAP rank list of feature importance.

Regarding the local explanation, SHAPASH and LIME share similar consistency properties. Figure 11(b) shows how SHAPASH determines the most essential elements for predicting the class 3 (i.e., advanced early PD). The topmost features in time-series data, according to SHAPASH, are MSEADLG and NP3BRADY, both of which have the maximum statistical attributes. When using SHAPASH, ETC and LGBM list the same top features but in reverse order, as the comparison is made between the models. When comparing explainers within the same classifier (LGBM), SHAP and SHPASH have NP3BRADY in common. Furthermore, similar to how SHAP provides a high weight to features in predicting the advanced early PD patient globally (see Fig. 11(c)), the LIME approach assigns a slightly greater value to features in predicting the advanced stage. However, the correct class prediction probability is 71% and the NP3BRADY feature is second topmost in the LIME feature importance list. SHAP, LIME, and SHAPASH have been found to be consistent across models in predicting the correct target class.

❸ *RF Model:* The RF model is used to compare the SHAP, SHPASH, and LIME explanations to the previous two models (ETC and LGBM). As shown in Fig. 12(a), NP3BRADY, like ETC and LGBM, is at the top of the SHAP explainer's global feature relevance list in predicting the four-class case of PD patients. The NP3BRADY feature

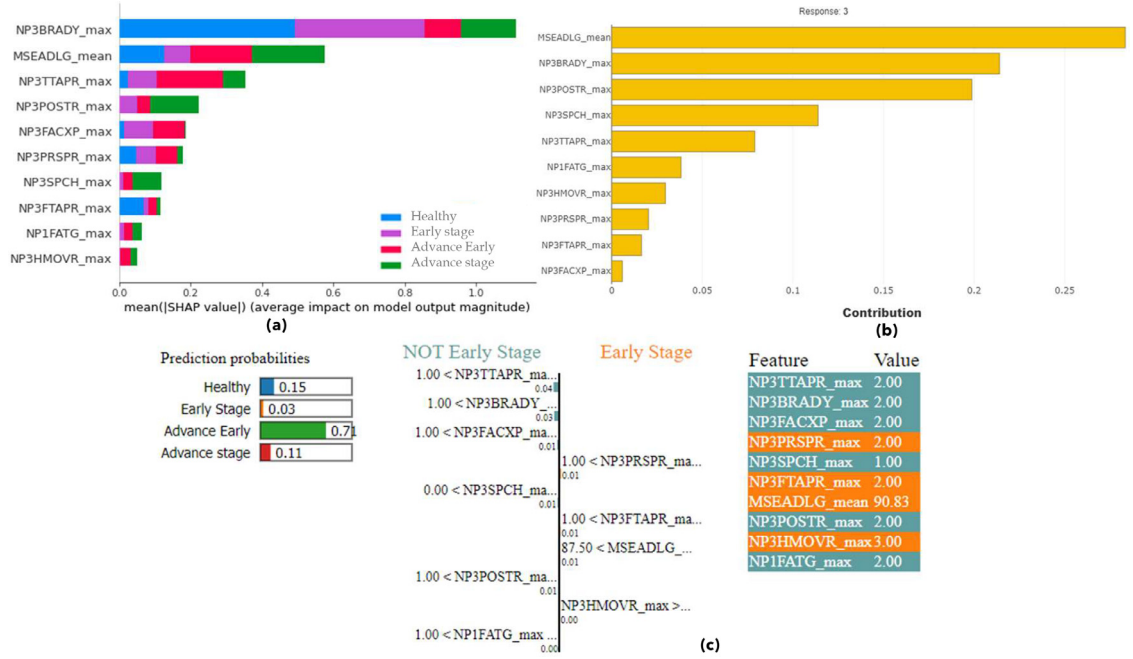

Fig. 11. Explainability of the LGBM ML classifier for predicting the four-class of PD by (a) SHAP, which shows the global feature importance by the class, (b) SHAPASH, which shows the local feature importance by a single instance, and (c) LIME, which shows the local feature importance by a single instance.

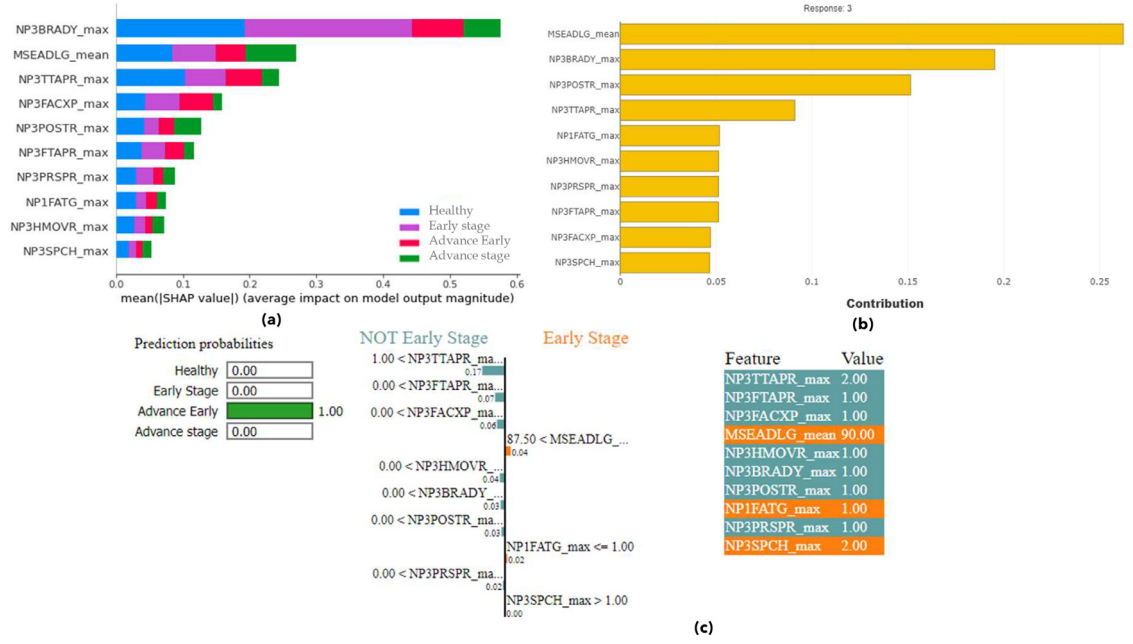

Fig. 12. Explainability of the RF ML classifier for predicting the four-class case of PD by (a) SHAP, which shows the global feature importance by the class, (b) SHAPASH, which shows the local feature importance by a single instance, and (c) LIME, which shows the local feature importance by a single instance.

aids in distinguishing the healthy and early-stage patients. The feature also contributes significantly to advanced early and advanced stage prediction when compared to other features. The NP3FACXP feature is listed as the second most important feature in SHAP for RF models, similar to LGBM but not ETC. Further, for LGBM and ETC, SHAPASH shows MSEADLG as the most essential feature.

In the case of the local explanation, SHPASH and LIME explainers are used for the RF model, which are equivalent to ETC and LGBM for feature comparison. An example of a label advanced early class is presented to the RF model (response 3). As shown in Fig. 12(b), the SHAPASH explanation generates a list of essential features, which includes MSEADLG, that has the mean statistical attribute. Other features, such as NP3BRADY and NP3POSTR respectively contribute 19% and 15% of the total. On the other hand, the LIME explanation correlates the feature significance score with the likelihood of prediction. LIME has a 100% probability of accurately classifying the input instance, as illustrated in Fig. 12(c). The NP3TTAPR feature makes the greatest significant contribution to the prediction. The analysis found that NP3BRADY is the most essential feature in predicting the correct target class globally, whereas MSEADLG found the most important feature locally. The results are evaluated on three models (ETC, LGBM, and RF) and

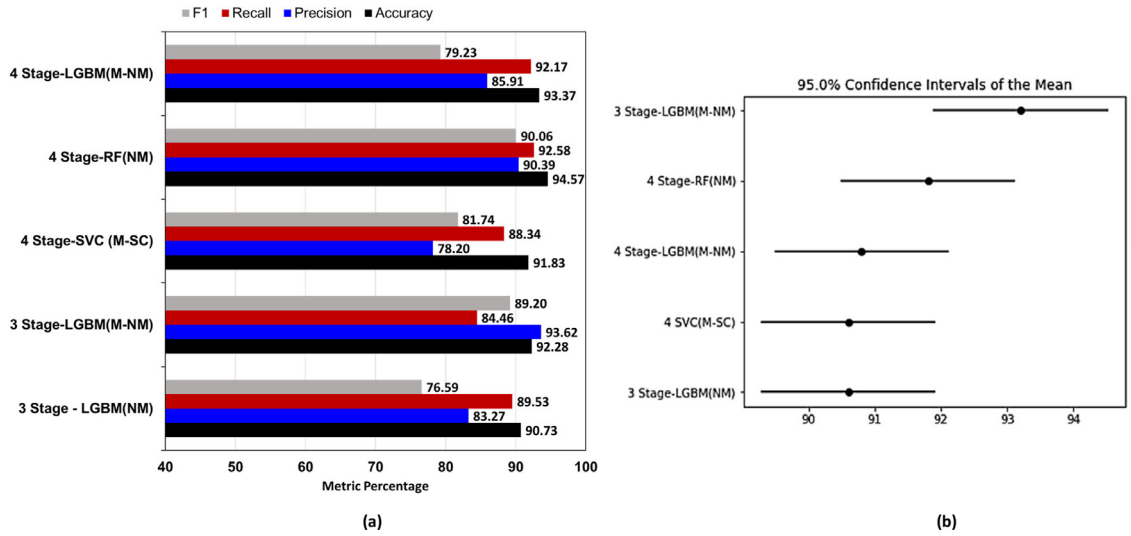

Fig. 13. Best performance of modalities in overall experiments.

evaluated using three separate explainers (SHAP, SHAPASH, and LIME), including global and local explainers.

5. Discussion

Three-class: In this section, we will delve into the results of the three-class and four-class predictions, as well as the findings from the explainability analysis. Our findings reveal that the LGBM model demonstrated excellent performance in the three-class prediction task, both for single-modality and multi-modality cases, as illustrated in Fig. 13(a). Our results indicate that all models performed exceptionally well when using the NM modality, which was found to be superior to the other modalities. In the single-modality scenario, the LGBM model achieved a maximum cross-validation accuracy of 90.73% using the NM modality. Furthermore, when using the NM-M modality in the multi-modality case, the LGBM model achieved a maximum cross-validation accuracy of 92.28%. The addition of the M modality to the NM modality improved performance by approximately 2%. Conversely, we observed a decline in performance as more modalities were combined.

The LGBM model was evaluated for its explainability compared to the RF and ETC machine learning-based approaches. To thoroughly assess the performance of these three machine learning models, three different explanation methods, as well as global and local approaches, were employed to ensure the robustness and generalization of the model. The results of the SHAP explanation showed that the NP3BRADY feature was the most significant for all three categories of PD patients, with the LGBM classifier having double the importance magnitude compared to the other models. The NP3BRADY feature was ranked at the top of the SHAPASH feature list in the local explanation. The contributions of the ETC, LGBM, and RF classifiers to predicting class 2 (early stage) were 55%, 50%, and 40%, respectively. The LIME explanation also revealed the NP3BRADY feature as the top-ranked feature for correct target class prediction. However, the prediction probability of LGBM (81%) was much lower compared to the ETC (98%) and RF (100%).

Four-class The results of the four-class prediction test showed mixed performance for multi-modality, with the NM modality providing the best results in single-modality scenarios. Although the LGBM model had the highest accuracy in the three-class prediction, the RF model had the best CV accuracy in the four-class prediction, reaching 94.57% as depicted in Fig. 13(a). The combination of the NM and M modalities provided the highest accuracy with the LGBM model, while the combination of the NM and SC modalities had the best results for the other classifiers. In particular, the SVC model had a maximum accuracy of 91.83% using the NM-SC combination, whereas the LGBM model achieved an even higher accuracy of 93.37% with the NM-M combination.

The performance and explainability of the LGBM model were evaluated in comparison to the RF and ETC machine learning-based approaches. The robustness and generalizability of the three machine learning models were assessed through the use of three different explanations, as well as both global and local methods. The results of the SHAP explanation showed that the NP3BRADY feature was the most significant for all four categories of Parkinson's disease patients, however, the LGBM classifier had a 2.5% higher importance magnitude compared to the ETC and RF models. In terms of the SHAPASH feature ranking list, the NP3BRADY feature was the most important for the ETC model, whereas the MESEADLG feature was the most important for the LGBM and RF models. The contributions of the ETC, LGBM, and RF classifiers towards predicting class 3 (advanced early stage) were 28%, 27%, and 26% respectively. Additionally, the NP3BRADY feature was the first in the LIME feature ranking list for the ETC model, while the NP3TTAPR feature was the first for the LGBM and RF models. In terms of prediction accuracy, the RF model was the most accurate at 100%, while the LGBM was the least accurate at 71% and the ETC was 84% accurate.

We also conducted an analysis of variance (ANOVA) to assess whether the mean values of the modalities differed significantly. If the confidence intervals of the two populations did not overlap, the populations were considered distinct. The results are depicted in Fig. 13(b), which displays the mean value for each population. There is a significant difference between the two modalities: In the three-class test, LGBM with the NM-M modality had a mean of 93.20, while LGBM with NM had a mean of 90.60. For the four-class test, the RF model with NM modality had a mean of 91.80, the LGBM model with NM-M modality had a mean of 90.80, and the SVC model with NM-SC modality had a mean of 90.60.

Our study was compared to existing state-of-the-art literature on predicting Parkinson's patients as presented in Table 5. A review of the literature shows that numerous studies have been conducted in this field since 2014, each using a unique dataset and modality. Many of these studies utilized single modalities from various

Table 5
Comparison with existing works for PD prediction.

Reference	Year	Dataset	Modalities	Total patients	Tasks	Feature engineering	Base classifier	CV	Performance (Accuracy %)	Method
[86]	2015	Non-wearable sensors	Gait data By Microsoft Kinect	7	Detection	Yes	NB,SVM,DT,RF	No	94%	SVM
[52]	2015	PPMI and VV	DaTSCAN image	497	Progression	Yes	SVM	No	91%, 94%	SVM
[57]	2015	Vowel speech data	Voice recordings.	80	Progression	Yes	SVM	No	85%	SVM
[54]	2015	CuPiD multimodal	Sensory data	18	Detection	Yes	Supervised ML	Yes	90%	ML
[22]	2016	PaWaH database	PD hand writing	37	Detection	No	SVM	No	84%	SVM
[67]	2016	3D MRI	T1-weighted MRI	60	Progression	Yes	SVM	No	94%	SVM
[42]	2017	PD telemonitoring	Dysphonia measurements voice	390	Progression	No	LDA, k-NN, NB, RT, RBFNN, SVM	Yes	92%	SVM
[88]	2017	PC-GITA	pronounce several speech	208	Progression	Yes	CNN,CWT	No	89%	CNN
[87]	2018	Phonetically Speech	Voice signal setence segmentation	268	Detection	No	CNN	No	86%	CNN
[26]	2018	UCI's PD Voice	Telemonitoring Voice	42	Progression	Yes	DNN	No	81%	DNN
[53]	2019	PPMI	Motor and Nonmotor	40	Detection	Yes	KNN + MLP	Yes	91.82%	SVM
[95]	2019	University of California	Pd patients	45	Progression	Yes	SVM	No	94.93%	ANN
[74]	2020	National Centre for Voice and Speech, Denver	Speech	31	Detection	Yes	SVM, ANN,CART	Yes	93.84%	ANN,SVM
[85]	2020	Sage Bionetwork	Motor and Non motor	2,289	Progression	Yes	SVM,NB,RF	Yes	85%	RF
[27]	2019	UCI PD Speech	Time Frequency characteristics	252	Progression	Yes	CNN,SVM	Yes	80%	CNN
[36]	2020	UCI PD Data	Voice feature	31	Progression	Yes	CART, SVM, ANN	No	93%	SVM
[84]	2021	GYENNO SCIENCE PD	Dynamic speech features	45	Detection	Yes	LSTM	Yes	84%	LSTM
[82]	2022	BioVid Heat Pain Database (BVDB)	Pd patients	52	Progression	Yes	RF	No	93.26%	RF,MLP,CNN
[25]	2023	Parkinson dataset	35	Detection	Yes	15 ML classifiers	Yes	91.45%	LGBM	
Our Work	–	PPMI	SC,BS,MH,M,NM	1059	Progression	Yes	RF,LGBM,ET,SVC,SGD	Yes	94%,94%	LGBM,RF

datasets, with some examples including speech data, time-series analysis, and imaginary data.

However, many of these studies had limitations, such as insufficient sample size, lack of cross-validation, utilization of non-time series data, and lack of validation of testing results. In comparison, our study used time-series data from 1059 patients, employed 10 well-known machine learning-based classifiers, and investigated both single and multi-modality approaches through advanced feature engineering techniques. Furthermore, we ensured the stability and robustness of our model by employing 10-holdout and 10-fold cross-validation. Our proposed framework achieved a cross-validation accuracy of 94% (LGBM) and a testing accuracy of 94% (RF).

Limitations and future work Although our suggested framework has the best accuracy and has been tested across ten various architectural classifiers, it does have some drawbacks: ❶ From a single PPMI dataset, we employed five modalities. Most notably, all the modalities are numerical features with ML models that have been validated for classification tasks. Beyond numerical features, it would be advantageous to incorporate imaginary modalities such as MRI, PET, and CT scans since imaginary modalities play a vital role in the prediction of PD. ❷ We implemented the H&Y scale to categorize the patients. Other scales, such as the Schwab and England scale, the Brain Bank of the United Kingdom PD Society, and Gelb's criteria and fatigue scale, may be used instead. As a multi-classification problem, it is also conceivable to integrate the progression and diagnosis of the Parkinson's stage prediction. ❸ When dealing with medical problems, deep learning models including LSTM and RNN may be trained and evaluated across diverse datasets to increase model generalizability. For example the model might be trained on PPMI and tested on the PC-GITA dataset. ❹ Further, ensemble learning and dynamic ensemble models are two sophisticated strategies that might be used to exploit the data. End-to-end learning for multi-tasking problems might also be thought of as a method that combine heterogeneous models and data. ❺ By extracting the statistical feature across the patients' visits, we turned the time-series modality into a non-time series; it is highly possible that some information was lost through that procedure. For our future study, we will exploit time-series data in conjunction with the most advanced ML/DL model and feature approach.

6. Conclusion

In this work, we comprehensively evaluated different proposed ML frameworks to predict PD progression. We explored different combinations of time series data fusion, different feature selection approaches, and different data balancing methods. We extend the resulting framework by implementing three different XAI techniques including SHAP, LIME, and SHAPASH l. Our experimental analysis demonstrated that the **NM** modality is the most informative feature set. In the 3-class case, **LGBM** achieved the maximum accuracy of 90.73%, whereas RF achieved the maximum accuracy of 94.57% in the four-class case. Both LGBM and RF are ensemble learning techniques based on tree base classifiers. However, RF classifier separately trains every tree with random sets of examples and features. This randomization makes the RF less likely to overfit the training data and more robust than every single decision tree. This is the possible reason for the superior performance of RF in the 4-class problem than LGBM.. The optimial ML models have been used to provide XAI features. We provide both global and local XAI features for different ML models using the three XAI techniques. The consistency of these techniques has been explored. XAI techniques highlighted **NP3BRADY** feature as the most important feature. Local explainers highlighted **MESEADLG** feature as the best feature with both LGBM and RF models. Resulting models are

expected to be trustworthy and medically acceptable in real environments because they provide both accurate and explainable decisions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience Program (IITP-2021-2020-0-01821) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1011198).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107495.

References

[1] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, V. Venkatraman, Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease, Future Gener. Comput. Syst. 83 (2018) 366–373.
[2] T. Abuhmed, S. El-Sappagh, J.M. Alonso, Robust hybrid deep learning models for Alzheimer's progression detection, Knowledge-Based Syst. 213 (2021) 106688, doi:10.1016/j.knosys.2020.106688.
[3] M. Ahmadlou, H. Adeli, Enhanced probabilistic neural network with local decision circles: a robust classifier, Integr. Computer-Aided Eng. 17 (3) (2010) 197–210, doi:10.3233/ICA-2010-0345.
[4] A.H. Al-Fatlawi, M.H. Jabardi, S.H. Ling, Efficient diagnosis system for Parkinson's disease using deep belief network, in: 2016 IEEE Congress on Evolutionary Computation (CEC), 2016, pp. 1324–1330, doi:10.1109/CEC.2016.7743941.
[5] S. Alabdulwahab, B. Moon, Feature selection methods simultaneously improve the detection accuracy and model building time of machine learning classifiers, Symmetry 12 (9) (2020) 1424, doi:10.3390/sym12091424.
[6] J.S. Almeida, P.P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, V.H.C. de Albuquerque, Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques, Pattern Recognit. Lett. 125 (2019) 55–62.
[7] P. Bahad, P. Saxena, R. Kamal, Fake news detection using bi-directional LSTM-recurrent neural network, Procedia Comput. Sci. 165 (2019) 74–82, doi:10.1016/j.procs.2020.01.072.
[8] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowl. Inf. Syst. 34 (3) (2013) 483–519.
[9] B.M. Boshkoska, D. Miljković, A. Valmarska, D. Gatsios, G. Rigas, S. Konitsiotis, K.M. Tsiouris, D. Fotiadis, M. Bohanec, Decision support for medication change of Parkinson's disease patients, Comput. Methods Programs Biomed. 196 (2020) 105552.
[10] C. Camara, P. Peris-Lopez, J.E. Tapiador, Human identification using compressed ecg signals, J. Med. Syst. 39 (11) (2015) 1–10, doi:10.1007/s10916-015-0323-2.
[11] F. Cavallo, A. Moschetti, D. Esposito, C. Maremmani, E. Rovini, Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning, Parkinsonism Relat. Disord. 63 (2019) 111–116.
[12] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[13] A.T. Connolly, W.F. Kaemmerer, S. Dani, S.R. Stanslaski, E. Panken, M.D. Johnson, T. Denison, Guiding deep brain stimulation contact selection using local field potentials sensed by a chronically implanted device in Parkinson's disease patients, in: 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2015, pp. 840–843, doi:10.1109/NER.2015.7146754.
[14] Y. Dai, W. Kuang, B.W. Ling, Z. Yang, K.-F. Tsang, H. Chi, C.-K. Wu, H.S.-H. Chung, G.P. Hancke, Detecting Parkinson's diseases via the characteristics of the intrinsic mode functions of filtered electromyograms, in: 2015 IEEE 13th International Conference on Industrial Informatics (INDIN), IEEE, 2015, pp. 1484–1487, doi:10.1109/INDIN.2015.7281952.
[15] R. Das, A comparison of multiple classification methods for diagnosis of Parkinson disease, Expert Syst. Appl. 37 (2) (2010) 1568–1572.
[16] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease, Artif. Intell. Med. 67 (2016) 39–46, doi:10.1016/j.artmed.2016.01.004.
[17] N. El-Rashidy, S. Abdelrazik, T. Abuhmed, E. Amer, F. Ali, J.-W. Hu, S. El-Sappagh, Comprehensive survey of using machine learning in the COVID-19 pandemic, Diagnostics 11 (7) (2021) 1155.
[18] N. El-Rashidy, T. Abuhmed, L. Alarabi, H.M. El-Bakry, S. Abdelrazek, F. Ali, S. El-Sappagh, Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning, Neural Comput. Appl. 34 (5) (2022) 3603–3632.
[19] S. El-Sappagh, T. Abuhmed, B. Alouffi, R. Sahal, N. Abdelhade, H. Saleh, The role of medication data to enhance the prediction of Alzheimer's progression using machine learning, Comput. Intell. Neurosci. 2021 (2021) 1–8.
[20] S. El-Sappagh, F. Ali, T. Abuhmed, J. Singh, J.M. Alonso, Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers, Neurocomputing 512 (2022) 203–224.
[21] S. El-Sappagh, M. Elmogy, F. Ali, T. Abuhmed, S. Islam, K.-S. Kwak, A comprehensive medical decision–support framework based on a heterogeneous ensemble classifier for diabetes prediction, Electronics 8 (6) (2019) 635.
[22] M. Faundez-Zanuy, J. Fierrez, M.A. Ferrer, M. Diaz, R. Tolosana, R. Plamondon, Handwriting biometrics: applications and future trends in e-security and e-health, Cogn. Comput. 12 (5) (2020) 940–953, doi:10.1007/s12559-020-09755-z.
[23] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results, Mov. Disord. 23 (15) (2008) 2129–2170.
[24] G. Goetz Christopher, P. Werner, R. Olivier, S. Cristina, T. Stebbins Glenn, C. Carl, G. Nir, G. Holloway Robert, G. Moore Charity, K. Wenning Gregor, Movement disorder society task force on rating scales for Parkinson's disease movement disorder society task force report on the Hoehn and Yahr staging scale: status and recommendations, Mov Disord 19 (9) (2004), doi:10.1002/mds.20213. 1020–8
[25] P. Gouverneur, F. Li, K. Shirahama, L. Luebke, W.M. Adamczyk, T.M. Szikszay, K. Luedtke, M. Grzegorzek, Explainable artificial intelligence (XAI) in pain research: Understanding the role of electrodermal activity for automated pain recognition, Sensors 23 (4) (2023) 1959.
[26] S. Grover, S. Bhartia, A. Yadav, K. Seeja, et al., Predicting severity of Parkinson's disease using deep learning, Procedia Comput. Sci. 132 (2018) 1788–1794, doi:10.1016/j.procs.2018.05.154.
[27] H. Gunduz, Deep learning-based Parkinson's disease classification using vocal feature sets, IEEE Access 7 (2019) 115540–115551, doi:10.1109/ACCESS.2019.2936564.
[28] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[29] T.J. Hirschauer, H. Adeli, J.A. Buford, Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network, J. Med. Syst. 39 (11) (2015) 1–12, doi:10.1007/s10916-015-0353-9.
[30] I. Illán, J. Górriz, J. Ramírez, F. Segovia, J. Jiménez-Hoyuela, S. Ortega Lozano, Automatic assistance to Parkinson's disease diagnosis in DaTSCAN spect imaging, Med. Phys. 39 (10) (2012) 5971–5980, doi:10.1118/1.4742055.
[31] D. Jain, V. Singh, Feature selection and classification systems for chronic disease prediction: a review, Egyptian Inform. J. 19 (3) (2018) 179–189.
[32] M. Jalal, P. Arabali, Z. Grasley, J.W. Bullard, H. Jalal, Behavior assessment, regression analysis and support vector machine (SVM) modeling of waste tire rubberized concrete, J. Clean. Prod. 273 (2020) 122960, doi:10.1016/j.jclepro.2020.122960.
[33] J. Jankovic, Parkinson's disease: clinical features and diagnosis, J. Neurol., Neurosurg. Psychiatry 79 (4) (2008) 368–376.
[34] G. Je, S. Arora, S. Raithatha, R. Barrette, N. Valizadeh, U. Shah, D. Desai, A. Deb, S. Desai, Epidemiology of Parkinson's disease in rural Gujarat, India, Neuroepidemiology 55 (3) (2021) 188–195, doi:10.1159/000515030.
[35] D. Joshi, A. Khajuria, P. Joshi, An automatic non-invasive method for Parkinson's disease classification, Comput. Methods Programs Biomed. 145 (2017) 135–145.
[36] R.M. JOSI, R. MINU, Review of computational approaches to Parkinson's disease gene prediction, in: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 780–788, doi:10.1109/ICISS49785.2020.9315926.
[37] F. Juraev, S. El-Sappagh, E. Abdukhamidov, F. Ali, T. Abuhmed, Multilayer dynamic ensemble model for intensive care unit mortality prediction of neonate patients, J. Biomed. Inform. 135 (2022) 104216.
[38] S.K. Khare, V. Bajaj, U.R. Acharya, Detection of Parkinson's disease using automated tunable q wavelet transform technique with EEG signals, Biocybern. Biomed. Eng. 41 (2) (2021) 679–689, doi:10.1016/j.bbe.2021.04.008.
[39] S.K. Khare, V. Bajaj, U.R. Acharya, Pdcnnet: an automatic framework for the detection of Parkinson's disease using eeg signals, IEEE Sens. J. 21 (15) (2021) 17017–17024, doi:10.1109/JSEN.2021.3080135.
[40] T. Ko, A.M. Brenner, N.P. Monteiro, M.S. Debastiani, A.C. Nesello, A. Hilbig, Abnormal eye movements in Parkinsonism: a historical view, Arq. Neuro-Psiquiatria 79 (2021) 457–459, doi:10.1590/0004-282X-ANP-2020-0406.
[41] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, Machine learning-based classification of simple drawing movements in Parkinson's disease, Biomed. Signal Process. Control 31 (2017) 174–180.
[42] S. Lahmiri, D.A. Dawson, A. Shmuel, Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures, Biomed. Eng. Lett. 8 (1) (2018) 29–39, doi:10.1007/s13534-017-0051-2.
[43] A.E. Lang, A.M. Lozano, Parkinson's disease, N. Engl. J. Med. 339 (16) (1998) 1130–1143.

[44] J.H. Lanskey, P. McColgan, A.E. Schrag, J. Acosta-Cabronero, G. Rees, H.R. Morris, R.S. Weil, Can neuroimaging predict dementia in Parkinson's disease? Brain 141 (9) (2018) 2545–2560.
[45] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563, doi:10.5555/3122009.3122026.
[46] H.W. Loh, W. Hong, C.P. Ooi, S. Chakraborty, P.D. Barua, R.C. Deo, J. Soar, E.E. Palmer, U.R. Acharya, Application of deep learning models for automated identification of Parkinson's disease: a review (2011–2021), Sensors 21 (21) (2021), doi:10.3390/s21217034.
[47] H.W. Loh, C.P. Ooi, E. Palmer, P.D. Barua, S. Dogan, T. Tuncer, M. Baygin, U.R. Acharya, Gaborpdnet: gabor transformation and deep neural network for Parkinson's disease detection using eeg signals, Electronics 10 (14) (2021), doi:10.3390/electronics10141740.
[48] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.
[49] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749, doi:10.1038/s41551-018-0304-0.
[50] Y.-W. Ma, J.-L. Chen, Y.-J. Chen, Y.-H. Lai, Explainable deep learning architecture for early diagnosis of Parkinson's disease, Soft Comput 27 (2023) 2729–2738.
[51] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al., The Parkinson progression marker initiative (PPMI), Prog. Neurobiol. 95 (4) (2011) 629–635, doi:10.1016/j.pneurobio.2011.09.005.
[52] F. Martínez-Murcia, J. Górriz, J. Ramírez, I. Illán, A. Ortiz, Automatic detection of Parkinsonism using significance measures and component analysis in DaTSCAN imaging, Neurocomputing 126 (2014) 58–70, doi:10.1016/j.neucom.2013.01.054.
[53] R. Mathur, V. Pathak, D. Bandil, Parkinson disease prediction using machine learning algorithm, in: Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, Springer, 2019, pp. 357–363.
[54] S. Mazilu, U. Blanke, G. Tröster, Gait, wrist, and sensors: detecting freezing of gait in Parkinson's disease from wrist movement, in: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), 2015, pp. 579–584, doi:10.1109/PERCOMW.2015.7134102.
[55] S. Mishra, P.K. Mallick, H.K. Tripathy, A.K. Bhoi, A. González-Briones, Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier, Appl. Sci. 10 (22) (2020) 8137, doi:10.3390/app10228137.
[56] A. Mohammed, M. Zamani, R. Bayford, A. Demosthenous, Patient specific Parkinson's disease detection for adaptive deep brain stimulation, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 1528–1531, doi:10.1109/EMBC.2015.7318662.
[57] L. Naranjo, C.J. Pérez, Y. Campos-Roca, J. Martín, Addressing voice recording replications for Parkinson's disease detection, Expert Syst. Appl. 46 (2016) 286–292, doi:10.1016/j.eswa.2015.10.034.
[58] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, M. Farahmand, A hybrid intelligent system for the prediction of Parkinson's disease progression using machine learning techniques, Biocybern. Biomed. Eng. 38 (1) (2018) 1–15.
[59] M.B.T. Noor, N.Z. Zenia, M.S. Kaiser, S. Al Mamun, M. Mahmud, Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia, Brain Inform. 7 (1) (2020) 1–21.
[60] G. Pahuja, T. Nagabhushan, A comparative study of existing machine learning approaches for Parkinson's disease detection, IETE J. Res. 67 (1) (2021) 4–14, doi:10.1080/03772063.2018.1531730.
[61] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, The unified Parkinson's disease rating scale (UPDRS): status and recommendations, Mov. Disord. 18 (7) (2003) 738–750.
[62] T.Í. Pedrosa, F.F. Vasconcelos, L. Medeiros, L.D. Silva, Machine learning application to quantify the tremor level for Parkinson's disease patients, Procedia Comput. Sci. 138 (2018) 215–220.
[63] K. Polat, Classification of Parkinson's disease using feature weighting method on the basis of fuzzy *c*-means clustering, Int. J. Syst. Sci. 43 (4) (2012) 597–609.
[64] R. Prashanth, S.D. Roy, Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning, Neurocomputing 305 (2018) 78–103, doi:10.1016/j.neucom.2018.04.049.
[65] R. Prashanth, S.D. Roy, P.K. Mandal, S. Ghosh, High-accuracy detection of early Parkinson's disease through multimodal features and machine learning, Int. J. Med. Inform. 90 (2016) 13–21, doi:10.1016/j.ijmedinf.2016.03.001.
[66] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. Del Ser, T. Abuhmed, Prediction of Alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data, Inf. Fusion 92 (2023) 363–388.
[67] B. Rana, A. Juneja, M. Saxena, S. Gudwani, S. Senthil Kumaran, R. Agrawal, M. Behari, Regions-of-interest based automated diagnosis of Parkinson's disease using T1-weighted MRI, Expert Syst. Appl. 42 (9) (2015) 4506–4516, doi:10.1016/j.eswa.2015.01.062.
[68] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: users, values, concerns and challenges, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 19–36.
[69] M.T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[70] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F.J. Martínez-Murcia, A. Ortiz, M.G. Río, M. Moreno-Caballero, Application of empirical mode decomposition (EMD) on DaTSCAN spect images to explore Parkinson disease, Expert Syst. Appl. 40 (7) (2013) 2756–2766, doi:10.1016/j.eswa.2012.11.017.
[71] C.O. Sakar, O. Kursun, Telediagnosis of Parkinson's disease using measurements of dysphonia, J. Med. Syst. 34 (4) (2010) 591–599.
[72] M.R. Salmanpour, M. Shamsaei, A. Saberi, S. Setayeshi, I.S. Klyuzhin, V. Sossi, A. Rahmim, Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease, Comput. Biol. Med. 111 (2019) 103347.
[73] F. Segovia, J. Górriz, J. Ramírez, I. Alvarez, J. Jiménez-Hoyuela, S. Ortega, Improved Parkinsonism diagnosis using a partial least squares based approach, Med. Phys. 39 (7Part1) (2012) 4395–4403, doi:10.1118/1.4730289.
[74] Z.K. Senturk, Early diagnosis of Parkinson's disease using machine learning algorithms, Med. Hypotheses 138 (2020) 109603.
[75] K.A. Severson, L.M. Chahine, L.A. Smolensky, M. Dhuliawala, M. Frasier, K. Ng, S. Ghosh, J. Hu, Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning, Lancet Digit. Health 3 (9) (2021) e555–e564.
[76] W.M. Shaban, A.H. Rabie, A.I. Saleh, M. Abo-Elsoud, A new COVID-19 patients detection strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier, Knowledge-Based Syst. 205 (2020) 106270, doi:10.1016/j.knosys.2020.106270.
[77] A.H. Shahid, M.P. Singh, A deep learning approach for prediction of Parkinson's disease progression, Biomed. Eng. Lett. 10 (2) (2020) 227–239.
[78] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. De Freitas, Taking the human out of the loop: a review of Bayesian optimization, Proc. IEEE 104 (1) (2015) 148–175.
[79] R.R. Shamir, T. Dolber, A.M. Noecker, B.L. Walter, C.C. McIntyre, Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease, Brain Stimul. 8 (6) (2015) 1025–1032, doi:10.1016/j.brs.2015.06.003.
[80] M. Shapash, GitHub - MAIF/shapash: Shapash makes Machine Learning models transparent and understandable by everyone, 2020, https://github.com/MAIF/shapash.
[81] G. Singh, M. Vadera, L. Samavedham, E.C.-H. Lim, Machine learning-based framework for multi-class diagnosis of neurodegenerative diseases: a study on Parkinson's disease, IFAC-PapersOnLine 49 (7) (2016) 990–995.
[82] S.K. Singh, A. Khamparia, A. Sinha, Explainable machine learning model for diagnosis of Parkinson disorder, in: Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI), Springer, 2022, pp. 33–41.
[83] S.L. Smith, M.A. Lones, M. Bedder, J.E. Alty, J. Cosgrove, R.J. Maguire, M.E. Pownall, D. Ivanoiu, C. Lyle, A. Cording, et al., Computational approaches for understanding the diagnosis and treatment of Parkinson's disease, IET Syst. Biol. 9 (6) (2015) 226–233, doi:10.1049/iet-syb.2015.0030.
[84] M. Su, K.-S. Chuang, Dynamic feature selection for detecting Parkinson's disease through voice signal, in: 2015 IEEE MTT-S 2015 International Microwave Workshop Series on RF and Wireless Technologies for Biomedical and Healthcare Applications (IMWS-BIO), 2015, pp. 148–149, doi:10.1109/IMWS-BIO.2015.7303822.
[85] J.M. Tracy, Y. Özkanca, D.C. Atkins, R.H. Ghomi, Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease, J. Biomed. Inform. 104 (2020) 103362.
[86] C.S. Tucker, I. Behoora, H.B. Nembhard, M. Lewis, N.W. Sterling, X. Huang, Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors, Comput. Biol. Med. 66 (2015) 120–134, doi:10.1016/j.compbiomed.2015.08.012.
[87] E. Vaiciukynas, A. Gelzinis, A. Verikas, M. Bacauskiene, Parkinson's disease detection from speech using convolutional neural networks, in: International Conference on Smart Objects and Technologies for Social Good, Springer, 2017, pp. 206–215, doi:10.1007/978-3-319-76111-4_21.
[88] J.C. Vásquez-Correa, J.R. Orozco-Arroyave, E. Nöth, Convolutional neural network to model articulation impairments in patients with Parkinson's disease, in: INTERSPEECH, Stockholm, 2017, pp. 314–318, doi:10.21437/Interspeech.2017-1078.
[89] J. Wainer, G. Cawley, Nested cross-validation when selecting classifiers is overzealous for most practical applications, Expert Syst. Appl. 182 (2021) 115222.
[90] K.R. Wan, T. Maszczyk, A.A.Q. See, J. Dauwels, N.K.K. King, A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease, Clin. Neurophysiol. 130 (1) (2019) 145–154.
[91] Y. Wang, A.-N. Wang, Q. Ai, H.-J. Sun, An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease, Biomed. Signal Process. Control 38 (2017) 400–410.
[92] R.S. Weil, J.K. Hsu, R.R. Darby, L. Soussand, M.D. Fox, Neuroimaging in Parkinson's disease dementia: connecting the dots, Brain Commun. 1 (1) (2019) fcz006.
[93] Y. Wu, X. Luo, P. Chen, L. Liao, S. Yang, R.M. Rangayyan, Forward autoregressive modeling for stride process analysis in patients with idiopathic Parkinson's disease, in: 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings, IEEE, 2015, pp. 349–352, doi:10.1109/MeMeA.2015.7145226.

[94] Z. Xue, J. Wei, W. Guo, A real-time naive Bayes classifier accelerator on FPGA, IEEE Access 8 (2020) 40755–40766, doi:10.1109/access.2020.2976879.
[95] A. Yasar, I. Saritas, M. Sahman, A. Cinar, Classification of Parkinson disease data with artificial neural networks, in: IOP Conference Series: Materials Science and Engineering, vol. 675, IOP Publishing, 2019, p. 012031.
[96] X. Zeng, G. Luo, Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection, Health Inf. Sci. Syst. 5 (1) (2017) 1–21.