Parkinson's Disease Classification Using SHAP, LIME, and Fisher Score with XGBoost and White-Box Machine Learning Models: An Explainable AI Perspective

Pratiyush Guleria
National Institute of Electronics and Information Technology (NIELIT) Shimla, Himachal Pradesh, India
E-mail: pratiyush@nielit.gov.in

Keywords: classifier, disease, explainable, learning, machine, parkinson, xgboost

Received: June 6, 2025

Parkinson's disease (PD) is a chronic neurological disorder that progressively impairs motor functions, often characterized by tremors, rigidity, and bradykinesia. Early diagnosis plays a vital role in improving patient outcomes. This study applies machine learning (ML) classifiers for the accurate detection of PD using a publicly available dataset comprising 195 voice recordings with 24 real-valued speech attributes. The preprocessing phase involved normalization and labeling, where the status attribute encoded healthy (0) and PD (1) subjects. The models are evaluated using 10-fold cross-validation to ensure robust performance estimation. The study evaluated a diverse range of machine learning classifiers, encompassing interpretable white-box models along with complex black-box models. Explainable Artificial Intelligence (XAI) techniques, including Fisher Score, SHAP, and LIME, are employed to interpret and rank feature importance, enhancing model transparency. Among all classifiers, XGBoost achieved the best results, with an accuracy of 94.87%, F1-score of 96.97%, and ROC-AUC of 0.91. The findings highlight that integrating XAI methods with ML not only yields high classification accuracy but also provides interpretable insights essential for clinical decision support in PD diagnosis.

Povzetek: Študija združuje XGBoost in razložljivo AI (SHAP, LIME, Fisher) za natančno in interpretabilno odkrivanje Parkinsonove bolezni iz govora.

1 Introduction

Parkinson's disease (PD) is a progressive neurological condition primarily associated with a reduction in dopamine levels in the brain [1]. This deficiency leads to worsening motor abilities and is commonly characterized by symptoms such as muscle stiffness and tremors. Additionally, individuals with PD often exhibit speech-related difficulties, including dysarthria (impaired articulation), hypophonia (reduced speech volume), and a monotone voice (restricted pitch variation). As the disease advances, it may also result in mood changes, cognitive decline, and an increased risk of developing dementia. Affecting approximately 2–3% of the elderly population, Parkinson's disease ranks as the second most common neurodegenerative disorder after Alzheimer's disease [1]. A key pathological feature of PD is the loss of dopaminergic neurons in the substantia nigra region of the brain [2]. The clinical diagnosis of PD primarily relies on identifying motor symptoms such as bradykinesia, rigidity, and resting tremor [3]. Physicians typically evaluate the patient's neurological history and assess their motor performance across different situations to establish a diagnosis.

However, due to the lack of a reliable laboratory test, especially in the early stages when motor symptoms may be mild, accurate diagnosis remains challenging. Regular follow-up visits are also required to monitor disease progression. Consequently, there is a critical need for an effective, remote screening solution that does not require in-person clinical visits. Encouragingly, recent studies have explored the potential of machine learning (ML) algorithms for analyzing voice recordings to detect Parkinson's disease. Since PD patients often display specific vocal abnormalities such as hypophonia, dysarthria, and monotone speech like, ML models trained on audio data from both affected individuals and healthy controls have demonstrated promising results in facilitating accurate, non-invasive diagnosis. One of the key benefits of utilizing voice recordings for Parkinson's disease screening lies in their non-invasive nature and the feasibility of conducting assessments remotely. Patients can capture voice samples using everyday tools like smartphones or specialized recording equipment, which can then be securely shared with medical professionals for further evaluation. If machine learning models demonstrate strong diagnostic performance, voice-based assessments could serve as a practical preliminary method for flagging individuals who may require more detailed clinical examinations. It is crucial, however, to emphasize that such techniques are meant to support and not substitute for comprehensive clinical diagnosis conducted by trained medical practitioners. While ML

technologies are powerful aids, they must be integrated with expert evaluations and other diagnostic procedures to ensure accurate results. A various machine learning techniques including classification, regression, clustering, and advanced deep learning approaches have been applied to analyze a wide spectrum of Parkinson's-related data. These data sources include patient assessments, medical imaging, genetic information, and signals from wearable devices. A central aim in this research is to develop models that can reliably distinguish between individuals with Parkinson's disease and those without it. The success of multiple studies in achieving high classification accuracy highlights the potential of ML as a diagnostic aid. Beyond diagnosis, machine learning has also been instrumental in evaluating treatment effectiveness and predicting disease progression. Additionally, by extracting relevant patterns from diverse datasets, ML has contributed to identifying potential biomarkers useful for early detection and continuous monitoring of Parkinson's disease. The application of cutting-edge algorithms has enabled researchers to uncover relationships and trends that are not easily detectable through conventional statistical analysis. Despite its promise, the field still faces challenges, such as the need for interpretable models, validation in clinical environments, and access to large, diverse datasets. Nonetheless, ML continues to offer transformative opportunities for enhancing the understanding, detection, and treatment of Parkinson's disease.

Research objectives

The primary objectives of this study are reformulated as the following research questions to ensure clarity and testability:

- **RQ1:** Does feature selection using SHAP improve the classification accuracy for PD detection compared to using all features?
- **RQ2:** How does the performance of different machine learning classifiers (SVM, KNN, LR, NN, NB, DT, and boosting) vary when applied on both complete and feature-selected datasets?
- **RQ3:** Which feature selection technique (Fisher Score, SHAP, or LIME) yields the most discriminative set of attributes for Parkinson's disease classification?

The organization of this paper is outlined as follows: Section 2 summarizes the relevant literature reviewed for this study. Section 3 details the methodology, covering data preprocessing, initial filtering, feature selection strategies, and classification techniques. Section 4 reports the experimental results and provides an interpretation of the outcomes. Finally, Section 5 concludes the study and lists the references.

2 Literature work

Parkinson's disease is a slowly progressing neurological brain disorder [4]. Brain cells die as a result of neurological disorders. Brain cells ordinarily create dopamine in specific regions of the brain that humans use. When 60–80% of the dopamine-producing cells are destroyed, the body is unable to produce enough dopamine, which causes the motor symptoms of disease. The study on feature selection for machine learning (ML) in brain surgery was reviewed by the authors [5]. The authors estimated the level of tremor in PD patients using an ML paradigm [6]. In [7], a variety of machine learning and feature extraction methods were used to identify Parkinson's disease. They showed that phonation is the simplest task for PD detection. Four classifiers evaluated in the study are SVM, KNN, multilayer perceptron (MLP), and optimum path forest. In [8], vocal features are reduced using artificial neural networks in order to facilitate ML-based PD diagnosis. Furthermore, using machine learning approaches, handwriting exercises are able to identify PD more accurately than MRI, motion, or voice data. ML approaches have been extensively employed in studies to facilitate the diagnosis, prognosis, and monitoring of Parkinson's disease. A genetically optimised neural network and linear discriminant analysis are used in an automated manner to identify Parkinson's disease based on several forms of prolonged phonations [9]. Based on clinical data, the scientists used a variety of machine learning methods, such as random forests and SVM, to forecast how disease would proceed [10]. To forecast the progression of PD, researchers employed a DL model, as opposed to techniques such as random forest, elastic net, and SVM, which simulate the impact of clinical and biofluid predictors [11, 12]. In a study, scientists examined the effectiveness of various ML techniques for categorising gait patterns in people with PD [13]. They also used sensor technology and ML approaches to predict the freezing of gait episodes in patients with PD [14]. The authors conduct a study outlining the challenges and opportunities in this field and giving a summary of the various machine learning techniques used to diagnose a disease [15]. AI and ML are being used extensively in the diagnosis, therapy, and discovery of new diagnostics in the course of Parkinson's disease. Additionally, authors have emphasised how changed lipidomics and the gut-brain axis can help control PD [16].The authors functioned on a task that analysed voice data using ML techniques to assess the seriousness of Parkinson's disease symptoms. The authors demonstrate the usefulness of these algorithms for illness diagnostic analysis in a clinical setting [17]. A review paper provides an overview of the latest advancements in ML and DL techniques for classifying Parkinson's disease. It lists the benefits and drawbacks of several algorithms and evaluates their efficacy [18]. The review paper gives an overview of different ML approaches for PD diagnosis, including wearable sensors, neuroimaging, and clinical data. Authors discuss the challenges, possible solutions, and future directions in this field [19]. A DL model is put forth to use neuroimaging data, notably functional connectivity, and hippocampus volume, for the early diagnosis of PD dementia. The authors show how DL methods can identify individuals who are at risk of cognitive decline [20]. To provide a clearer overview of recent studies, a summarized comparison of machine learning techniques,

datasets, and performance metrics used in Parkinson's disease classification is presented in Table 1.

Table 1: Summary of selected literature on ML approaches for parkinson's disease diagnosis

Ref.	ML Methods Used	Dataset Details	Performance Metrics
[7]	SVM, KNN, MLP, Optimum Path Forest	Voice dataset (multiple phonation tasks, real-valued features)	Accuracy ≈ 91%; phonation found most effective
[8]	Adaptive Kernel-based Weighted Extreme Learning Machine (AABC-KWELM)	Not specified	Accuracy ≈ 97.93%
[9]	Linear Discriminant Analysis (LDA), Genetically Optimized Neural Network (GONN)	Multiple types of sustained phonations	Training Accuracy ≈ 95%; Testing Accuracy ≈ 100%
[10]	SVM, ANN, Classification and Regression Trees	Voice features	Accuracy ≈ 93.84%
[14]	ANN, SVM, Decision Tree	Voice dataset	Accuracy ≈ 95.89%
[21]	Long Short-Term Memory (LSTM) Network	Resting-state fMRI data from 84 subjects (56 stage 2, 28 stage 1)	Accuracy 71.63%; 13.52% improvement over traditional ML methods
[22]	Random Forest, Gradient Boosting, Support Vector Machine	Stabilometric parameters from postural sway analysis	Accuracy ≈ 90%; AUC ≈ 0.91
[23]	Deep Learning (CNN, LSTM)	Gait signals from PhysioNet dataset	Accuracy 97.71%, Sensitivity ≈ 99%, Precision ≈ 98%, Specificity ≈ 96%
[24]	Contrastive Graph Cross-View Learning	Multimodal fusion of SPECT images and clinical features	Accuracy ≈ 0.91; AUC ≈ 0.93
[25]	Ensemble Classifiers (Random Forest, Gradient Boosting, SVM)	Biomechanical data from balance assessments	Accuracy ≈ 90%; AUC ≈ 0.91

The previous studies have effectively utilized ML and DL techniques for Parkinson's disease classification; however, most lacked interpretability and detailed feature-level insights**.** The present study bridges this gap by combining multiple classifiers with Explainable AI (XAI) techniques to improve both predictive accuracy and transparency, making the results more applicable for clinical decision support.

3 Research methodological context

In recent years, the incorporation of XAI methods into medical diagnosis systems has become more prominent, providing transparency and interpretability to previously complex machine learning models. This research utilizes SHAP to provide insight into the predictions made by ML classifiers. SHAP calculates a unique importance value for each feature within a specific prediction, based on the principles of cooperative game theory. In medical settings, this approach is particularly valuable; as it helps clinicians make informed decisions by understanding how characteristics, like vocal parameters, impact disease diagnosis. The research methodology followed in this paper for predicting PD using feature selection and ML techniques typically involves several steps to achieve accurate and reliable results. This study employs two complementary research frameworks to integrate explainable AI (XAI) techniques into machine learning-based classification. The first framework shown in Figure 1 adopts a linear pipeline that begins with data preprocessing, followed by feature selection using Fisher Score, SHAP, and LIME, and then proceeds to train multiple classifiers including SVM, KNN, and Neural Networks. This approach emphasizes the quantitative contribution of each feature in prediction and follows a structured cross-validation and evaluation process. The second framework shown in Figure 2 is the XAI-based framework categorizing ML models by interpretability and feature selection approach. ML classifiers are trained on the most significant features to improve performance, and feature extraction techniques are used to extract significant features. Each feature's impact on performing prediction is determined, and significant features are chosen. The important features are used to train the machine learning classifiers used in the research. The outcomes are interpreted following the model's

examination. The dataset is divided into training and testing sets, initially following an 80:20 ratio. However, to enhance robustness and reduce bias, the study employed 10-fold stratified cross-validation instead of relying solely on a single 80:20 split. Stratification ensured class balance in each fold, thereby improving the reliability and statistical validity of the reported results.

3.1 Dataset information

The study's dataset, which has 24 attributes and 195 cases overall, can be accessed online [26]. There are several different biological voice metrics in this dataset. Each row in the table corresponds to one of the 195 voice recordings made by these people as mentioned in the 'name' attribute, with each attribute representing a specific voice measure. The dataset characteristics are of multivariate scope and the attribute characteristics are real in nature. The task associated with the dataset is to perform classification. The 'status' column, which is set to 0 for healthy and 1 for PD, serves as the primary means of differentiating between individuals without PD and those with it. The information is in comma-separated values (CSV) format. Each row in the CSV file represents a single voice recording occurrence. Each patient has about six recordings. For the purpose of experimentation, a Jupyter Notebook version 6.1.4 is utilized. Table 2 displays the sample dataset, and Table 3 displays the attribute information for the sample dataset.

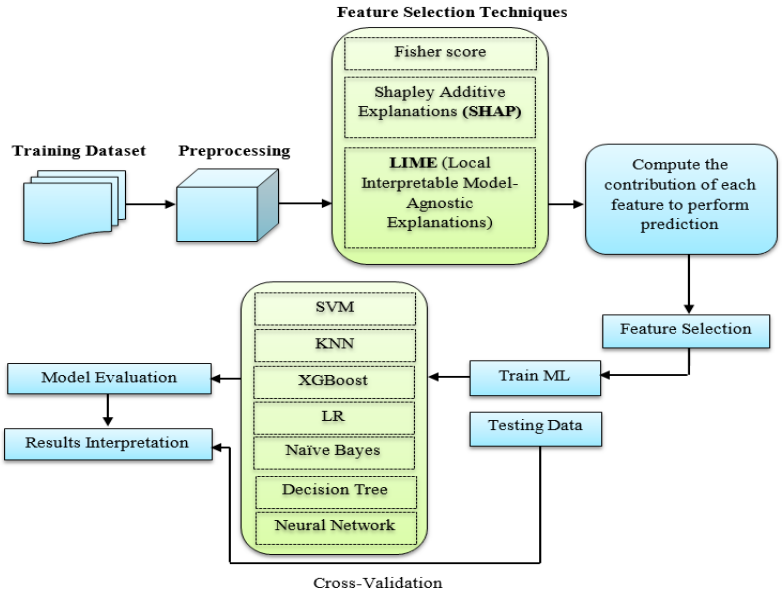


Figure 1: Workflow of ML classification with integrated feature selection techniques and evaluation

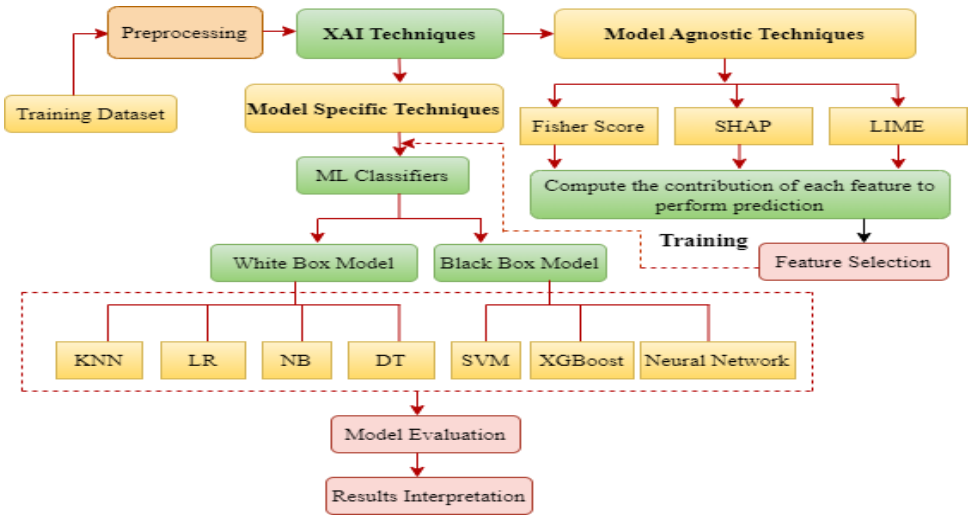


Figure 2: XAI-based framework categorizing ML models by interpretability and feature selection approach

Table 2: Sample Dataset of PD

name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(d	Shimmer:APQ3	Shimmer:APQ5	MDVP:APQ	Shimmer:DDA	NHR	HNR	status	RPDE	DFA	spread1	spread2	D2	PPE
phon_R01_S01	119.99	157.3	74.997	0.0078	7E-05	0.004	0.0055	0.011	0.044	0.426	0.0218	0.0313	0.02971	0.0655	0.022	21.03	1	0.4148	0.8153	-4.81303	0.26648	2.3014	0.2847
phon_R01_S01	122.4	148.65	113.82	0.0097	8E-05	0.005	0.007	0.014	0.061	0.626	0.0313	0.0452	0.04368	0.094	0.019	19.09	1	0.4584	0.8195	-4.07519	0.33559	2.4869	0.3687
phon_R01_S01	116.68	131.11	111.56	0.0105	9E-05	0.005	0.0078	0.016	0.052	0.482	0.0276	0.0386	0.0359	0.0827	0.013	20.65	1	0.4299	0.8253	-4.44318	0.31117	2.3423	0.3326
phon_R01_S01	116.68	137.87	111.37	0.01	9E-05	0.005	0.007	0.015	0.055	0.517	0.0292	0.0401	0.03772	0.0877	0.014	20.64	1	0.435	0.8192	-4.1175	0.33415	2.4056	0.369
phon_R01_S01	116.01	141.78	110.66	0.0128	0.0001	0.007	0.0091	0.02	0.064	0.584	0.0349	0.0483	0.04465	0.1047	0.018	19.65	1	0.4174	0.8235	-3.74779	0.23451	2.3322	0.4103
phon_R01_S01	120.55	131.16	113.79	0.0097	8E-05	0.005	0.0075	0.014	0.047	0.456	0.0233	0.0353	0.03243	0.0699	0.012	21.38	1	0.4156	0.8251	-4.24287	0.29911	2.1876	0.3578
phon_R01_S02	120.27	137.24	114.82	0.0033	3E-05	0.002	0.002	0.005	0.016	0.14	0.0078	0.0094	0.01351	0.0234	0.006	24.89	1	0.596	0.7641	-5.63432	0.25768	1.8548	0.2118
phon_R01_S02	107.33	113.84	104.32	0.0029	3E-05	0.001	0.0018	0.004	0.016	0.134	0.0083	0.0095	0.01256	0.0249	0.003	26.89	1	0.6374	0.7633	-6.1676	0.18372	2.0647	0.1638

Table 3: Attribute Information in a dataset

Attribute	**Description of measurement methods applied to acoustic signals recorded from each subject**
name	ASCII subject name and recording number
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP, MDVP:PPQ,Jitter:DDP	Multiple indicators of fundamental frequency fluctuation
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5, MDVP:APQ,Shimmer:DDA	Multiple amplitude variation measurements
NHR,HNR	Two measures of ratio of noise to tonal components in the voice
RPDE,D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1,spread2,PPE	Three nonlinear measures of fundamental frequency variation
status	Health status of the subject (one) - Parkinson's, (zero) - healthy

Note: MDVP stands for (Kay Pentax) Multi-Dimensional Voice Program; PPE stands for Pitch period entropy, RPDE stands for Recurrence Period Density Entropy, DFA stands for Detrended Fluctuation Analysis

3.1.1 Data preprocessing and feature selection techniques

An essential phase in machine learning activity is pre-processing. In this process, the dataset is cleaned by handling missing values, outlier removal, and normalization/standardization of features. This step is crucial to ensure the data is in a consistent and suitable format for ML algorithms. To eliminate outliers from the training dataset, one-hot encoding techniques are employed to convert the non-numerical data as a part of pre-processing techniques. The features are selected and relevant features are transformed from the raw data. The features and labels are separated, and fisher scores are calculated to determine the importance of each feature. The features are scaled using standard scale method to ensure all features have the same scale. The proposed study uses several strategies for improving the interpretability and understanding of model predictions in machine learning, including Fisher Score, Shapley values, and LIME. The discriminative ability of features in classification problems is measured using the fisher score, which is then utilised for feature selection. Shapley values are based on ideas from cooperative game theory and offer a means of attributing the model's predictions to specific traits. LIME approximates individual predictions of complicated models with more easily understood models close to the prediction of interest, producing local explanations for each prediction.

3.2 XAI tools

XAI tools are pivotal in closing the knowledge divide between sophisticated ML systems and human comprehension. The increasing use of data-driven algorithms in sensitive areas like healthcare, education, and finance has led to a substantial rise in the need for transparency and accountability. XAI tools are designed to make the decision-making process of models more understandable by showing the impact of input features, explaining the reasoning behind predictions, and allowing users to verify or dispute the results. These tools differ in their approach, with some designed to function independently of any specific model, whereas others are customized for particular types of algorithms. Improving model transparency through the use of XAI tools enables not only increased user trust but also facilitates model debugging, bias identification, and adherence to ethical or regulatory requirements.

3.2.1 Model-Agnostic techniques

Interpretable ML systems require model-agnostic tools, particularly when dealing with complex or black-box models whose inner workings are unclear. These tools are designed to function autonomously of the underlying algorithm, enabling them to be highly

adaptable across various types of classifiers and datasets. By focusing solely on the relationship between input features and model outputs, they enable researchers and practitioners to extract meaningful insights without needing access to the inner workings of the model. This versatility enables the use of model-agnostic methods across a wide range of tasks including feature selection, explaining individual predictions, and assessing overall model performance, ultimately increasing transparency and trust in data-driven decision-making processes.

3.2.1.1 Fisher discriminant analysis

Fisher's score function known as fisher discriminant analysis is deeply related to maximum likelihood estimation as mentioned in (1). Fisher's score to find the maximum of the likelihood function all along, just without explicitly using the term. Assume some dataset X where each observation is identically and independently distributed according to a true underlying distribution parameterized by θ. Given this probability density function $f\theta(x)$, the likelihood function as follows:

$$p(x|\theta) = \prod_{i=1}^{n} f_\theta(x_i) \quad (1)$$

The maximum likelihood estimate of the distribution's parameter is given by (2).

$$\theta_{mle} = arg_\theta max\, \mathrm{p(x|\theta)} \quad (2)$$

$$= arg_\theta max \log p\,(\mathrm{x|\theta}) \quad (3)$$

$$= arg_\theta \max \sum_{i=1}^{n} log f_\theta\ (x_i) \quad (4)$$

The gradient of log likelihood function is shown below:

$$u(\theta) = \nabla_\theta log p(x|\theta) \quad (5)$$

The feature importance i.e. Fisher Score obtained from the PD dataset is shown in Figure 3. In model selection and training phase, ML classifiers for the task are taken, such as SVM, KNN, XGBoost, LR, NN, DT and NB. The dataset is split into training and testing subsets for training a model. The training data is used to train the ML models and optimize their hyperparameters. Apart from it, the cross-validation techniques are also being used to avoid overfitting and assess model generalization. In performance evaluation, the trained models are evaluated using appropriate metrics like accuracy, precision, recall, F1 score, or area under the receiver operating characteristic curve (AUC-ROC). The performance of different models is compared to identify the best-performing one. Since Parkinson's disease diagnosis and management require interpretability, it's essential to explore methods to explain the models' decisions. Therefore, the techniques like SHAP, LIME are used in the work to help and provide insights into the features that significantly influence the model's predictions. The bar chart in Figure 3 shows the importance of features calculated using the Fisher Score. The y-axis lists the feature names, and the x-axis represents the Fisher Score values. Features like PPE, MDVP:Fo(Hz), and spread1 have the highest Fisher Scores, indicating they contribute most to class separation in the dataset.

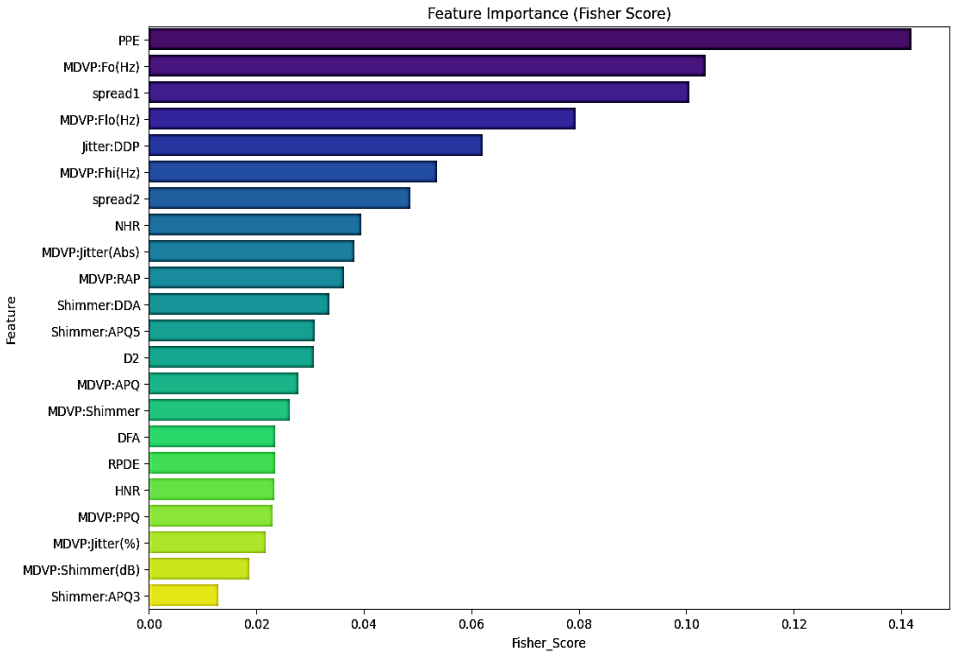


Figure 3: Feature importance using fisher score

3.2.1.2 Shapley additive explanations

Shapley values are a useful tool for explaining ML model outcomes and can be used to explain feature contributions. It's a notion from game theory that utilized to calculate each group member's participation. Assuming that features are the players or team members and the model prediction is the game's or teamwork's contribution. For every iteration, the feature values for all features are drawn in an arbitrary sequence. The difference between the prediction with and without feature i to determine the significance of feature i is computed. Calculating the mean difference across every possibility yields the Shapley value. It is the feature's

average marginal impact when all possible combinations are taken into account. It is only possible to determine the Shapley value with a subset of combinations. Assuming that a group t_m consists of n_m members, and that each member contributes significantly to the work during the group time frame, the teamwork results in the total value $tot_{val} = tot_{val}(t_m)$. The outcome for each member of the group m is represented by the shapley value, $\varphi_m\,(tot_{val})$. The definition of $\varphi_m\,(tot_{val})$ is

$$\varphi_m\,(tot_{val}) = \frac{1}{n_m}\sum_{s_{subset}} \frac{[tot_{val}(s_{subset}\cup\{m\}) - tot_{val}(s_{subset})]}{\binom{n_m-1}{K(s_{subset})}}, m = 1,2,3,\ldots..n_m \quad (6)$$

The sum of the values for a particular member,m is the total of all the subsets s_{subset} of the group $t_m = \{1,2,3,\ldots..,n_m\}$, that can be formed once m is eliminated. The size of a subteam is denoted by $K(s_{subset})$, its accomplished value is $val(s_{subset})$, and its actual value once m joins s_{subset} is $val(s_{subset} \cup \{m\})$.The appropriate collection is defined by the Shapley value as a distribution that meets requirements for linearity, symmetry, and efficiency. The plot in Figure 4 visualizes the impact of each feature on the XGBoost model output using SHAP values. The x-axis shows SHAP values (impact on prediction), and the y-axis lists the features. Colors indicate feature value (red = high, blue = low. Features like PPE and spread1 have the largest range of SHAP values, meaning changes in these features have the most influence on model predictions.

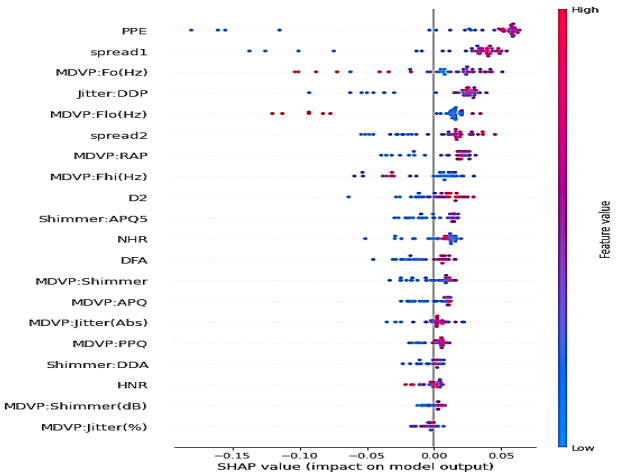


Figure 4: SHAP value impact on model output

The SHAP value and its impact on model output from low to high is shown in Figure 4.The mean SHAP value i.e. average impact on model output magnitude is shown in Figure 5. It shows the average absolute SHAP values for each feature, separated by class. The x-axis represents mean SHAP values, and the colors indicate Class 0 (blue) and Class 1 (red). PPE, spread1, and MDVP:Fo(Hz) are the most influential features for distinguishing between the two classes. Some features have higher impact on one class than the other. Figure 6 displays the beeswarm plot for feature relevance in Parkinson's disease prediction. It is inferred from Figure 4, Figure 5, and Figure 6, where the most significant features are 'spread1', 'MDVP:Fo (Hz)', and 'PPE' among other features.

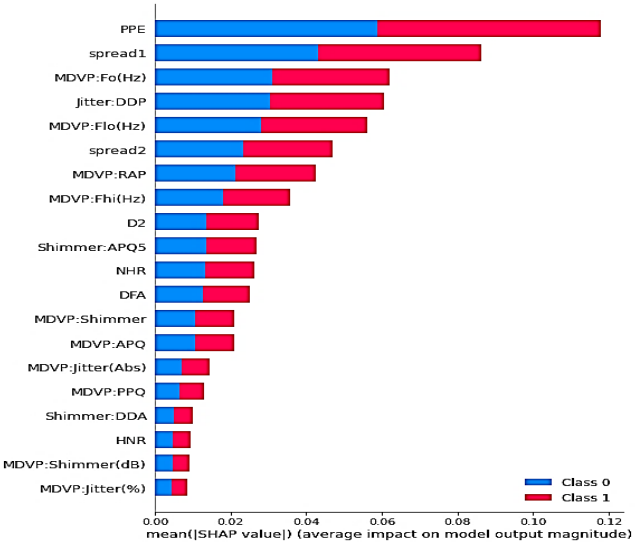


Figure 5: Mean SHAP value

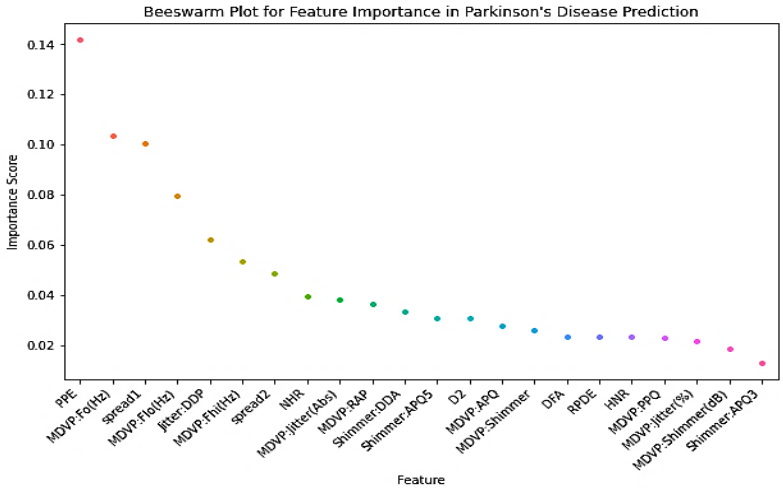


Figure 6: Beeswarm plot for feature importance in PD prediction

A force-directed visualisation that breaks down the model's prediction into components attributed to each feature is displayed in Figure 7 is shape force plot. This visualisation shows how each feature contributes to pushing the model's output from a base value, the average model output to the predicted value for the particular instance. The base value of the model's output, known as the mean prediction, is shown by the horizontal line in the center of the plot. It serves as a starting point before taking into account the influence of any features. A corresponding bar on the plot corresponds to each feature in the instance that is being explained. Each bar's length indicates how that feature affects the model's forecast. Positive values are represented by the bars pointing right, indicating that the value of the feature raises the output of the model, while negative values are represented by the bars pointing left, indicating a drop in the output of the model. In addition to the global feature ranking shown in the beeswarm plot in Figure 6, local interpretability is explored using the SHAP force plot in Figure 7. The beeswarm visualization highlights that PPE**,** MDVP:Fo(Hz)**,** Spread1**,** and MDVP:Fhi(Hz) are the most influential predictors of Parkinson's disease. To further interpret individual cases, the force plot demonstrates how specific feature values drive predictions: for instance, high PPE**,** low spread1, and elevated MDVP:Fhi(Hz) values collectively push the model output toward the Parkinson's (PD) class, whereas higher MDVP:Fo(Hz) values contribute toward the healthy prediction. These local explanations confirm that the classifier's decisions align with clinically recognized voice impairments in PD, thereby enhancing the model's interpretability and practical relevance.

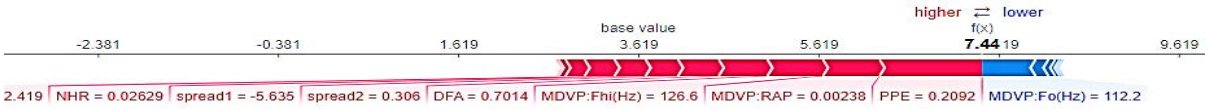


Figure 7: Shapley force-directed plot of models prediction attributed to each feature

3.2.1.3 LIME

Local Interpretable Model-Agnostic Explanations are referred to as LIME. It is a ML technique that approximates the behavior of black-box models locally to explain the predictions made by the models. It highlights the most significant characteristics and offers insights into why a ML model produces a particular prediction for a particular instance. Rather than explaining the model's overall behavior, LIME focuses on producing local explanations for specific predictions. It produces understandable approximations of the complex model behavior based on a particular occurrence.

$$rbf(y^{(j)}) = exp(-\frac{\|y^j - y^{(ref)}\|^2}{kw}) \quad (7)$$

LIME uses a Gaussian kernel, or radial basis function, to assign a weight to each generated point. Since it's a local method, it doesn't care about distant methods and ignores them. The Gaussian Kernel gives a value in the interval [0, 1] to an equation; the greater the value, the closer the equation is to the reference point. The value of the kernel width 'kw' parameter determines the diameter of the meaningful weights' circle surrounding the reference point $'ref'$. Figure 8 displays a LIME plot as a bar chart that illustrates how much the value of a feature changes and how much the anticipated value changes as well. It highlights the most important features and how they affect the forecast, visualising the local interpretability of an ML model's prediction for a particular instance.

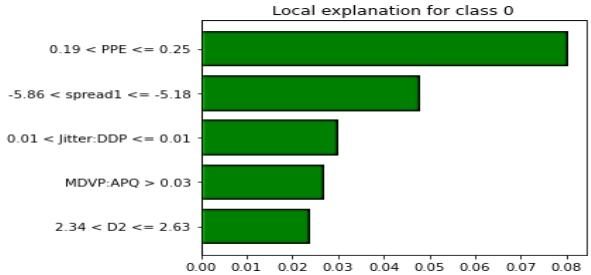


Figure 8: Local explanation of a model prediction using LIME

3.2.2 Model specific techniques

Model-specific techniques are explainability methods tailored to the internal structure and behavior of particular machine learning algorithms. Unlike model-agnostic approaches, these techniques leverage the mathematical and logical characteristics unique to a specific model type to provide deeper and more accurate insights into how predictions are made. For instance, decision trees naturally offer interpretability through their hierarchical structure of feature-based splits, while linear models provide direct access to feature weights that indicate their influence on the output. In neural networks, techniques such as saliency maps or layer-wise relevance propagation are used to trace the contribution of input features through the network's layers. By focusing on the internal mechanisms of each model, these techniques often yield more precise and informative explanations, although they are limited in scope to the models they are designed to interpret.

classifiers can help determine whether Parkinson's illness will develop or not. By differentiating PD from other neurological disorders that present with similar symptoms, they can help in the process of diagnosis, resulting in faster and more accurate diagnoses. To improve the effectiveness of classifiers for tasks involving the categorization of Parkinson's disease, feature engineering and selection are essential. Using the Shapley, LIME and Fisher score algorithms on the dataset, the salient features are retrieved.

Logistic regression

The probability of a certain class or occurrence can be modeled using the supervised machine learning approach known as logistic regression. When the outcome is binary or dichotomous in nature and the data is linearly separable, it is employed. Thus, situations involving binary classification typically employ logistic regression. Predicting an output variable that alternates between two discrete classes is known as binary classification.

Naïve Bayes

The Naive Bayes classification technique relies on the assumption that all features that predict the target value are independent of one another, and it is based on the bayes theorem. It determines each class's probability before selecting the one with the highest likelihood. The Bayes theorem as shown in (8) asserts that, given a class variable y and a features vector X= $(x_1, x_2, \ldots, x_n)$.

$$p(y|X) = \frac{P(X|y)*P(y)}{P(X)} \quad (8)$$

The probability $P(X \mid y)$ can be broken down as shown in (9):

$$P(X|y) = P(x_1|y) * P(x_2|y) \ldots\ldots\ldots P(x_n|y) \quad (9)$$

3.2.2.1 White box models

White-box machine learning models are characterized by their high degree of interpretability and transparency, making them particularly suitable for applications where understanding the reasoning behind a prediction is essential. These models allow researchers and practitioners to examine the internal structure, such as decision boundaries, feature weights, or probability distributions, to gain insights into how inputs are transformed into outputs. Unlike more complex black-box models, white-box approaches do not obscure their logic, which fosters trust and supports accountability in sensitive domains like healthcare. In the context of Parkinson's disease diagnosis, white-box models offer the added benefit of enabling clinicians to verify and understand the diagnostic criteria identified by the algorithm, thereby bridging the gap between computational predictions and medical decision-making. Their simplicity, coupled with ease of validation, makes them a valuable choice when explainability is a priority. Using diagnostics or early symptoms, ML

Decision Tree

A supervised ML approach based on the nested-if classifier is the decision tree. It creates models for regression or classification using a tree topology. The quantity of data required to precisely characterize a sample is known as its entropy, and it is computed as (10).

$$entropy = -\sum_{x=1}^{n} p_x * \log(p_x) \quad (10)$$

In this case, the fraction of examples in class x is denoted by p_x. Entropy equals 1 when classes are evenly distributed, and it equals 0 when one class is being most prevalent. Gini impurity is computed as follows:

$$gi = 1 - \sum_{x=1}^{n} (p_i)^2 \quad (11)$$

Finding a characteristic that yields the maximum information gain is the primary step in building a decision tree, as indicated by (12).

$$I_f G_n(d_{pnode}, attrib) = I_f(d_{pnode}) - \frac{N_{sleftchildnode}}{N} I_f(d_{leftchildnode}) - \frac{N_{srhtchildnode}}{N} I_f(d_{rgtchildnode}) \quad (12)$$

In equation above,$I_f G_n$ is the information gain, d_{pnode} is the dataset of the parent node. $d_{leftchildnode}$ is the dataset of left child node and $d_{rgtchildnode}$ is the dataset of right child node. $N_{sleftchildnode}$ represents the number of samples at left child node and $N_{srhtchildnode}$ is the number of samples at right child node.

KNN

K-Nearest Neighbor (KNN) is a straightforward method that identifies new data points according to the majority class of their k closest neighbors. Both classification and regression tasks can be accomplished with it. KNN presupposes that related instances frequently have related labels and targets. For distance metrics, the euclidean metric as shown in (13) is used:

$$d(x, \acute{x}) = \sqrt{(x_1 - \acute{x}_1)^2 + \cdots + (x_n - \acute{x}_n)^2} \quad (13)$$

Finally, the input x gets assigned to the class with the largest probability in (14).

$$P\ (y = j\ |X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j) \quad (14)$$

KNN measures the similarity between instances using distance metrics. Manhattan distance and Minkowski distance are two additional distance metrics; however euclidean distance is the one that is most frequently employed.

3.2.2.2 Black box models

Black-box machine learning models are known for their ability to capture complex, non-linear patterns within data, often achieving high predictive accuracy. However, what sets them apart is their limited interpretability; the internal processes and decision rules they employ are typically not transparent or easily understood by humans. These models, such as neural networks, support vector machines, and ensemble methods like XGBoost, consist of intricate layers, mathematical transformations, or combinations of multiple weak learners, making it challenging to trace how a specific output is derived from the input features. In medical applications like Parkinson's disease classification, black-box models can uncover subtle patterns in high-dimensional data that may not be detected by simpler algorithms. While their predictive power is valuable, the opaque nature of these models necessitates the use of explainable AI techniques to enhance trust, support clinical validation, and ensure that decisions align with medical understanding.

SVM

SVM is a potent supervised machine learning technique used for both regression and classification applications. It is extensively used because it can manage complicated datasets and deliver precise results. SVM are very useful when the data cannot be separated linearly and requires intricate decision boundaries. SVM algorithm looks for the best hyperplane to divide the data points into distinct groups. A hyperplane is a linear decision boundary that divides the data into two classes in binary classification. In multi-class classification, various hyperplanes are employed to divide various classes. With n training points, each observation i (x_i) has p dimensions and p characteristics, meaning it belongs to one of two classes ($y_i = 1 \; or \; y_i = -1$). Assume that two linearly separable classes of observations exist. This means that a feature space can be used to construct a hyperplane with all instances of one class on one side and all instances of the other class on the other. A hyperplane is defined in (15) as

$$x.\tilde{w} + \tilde{b} = 0 \tag{15}$$

Here, $\tilde{w}$ is a p vector and $\tilde{b}$ is a real number. When $\tilde{w}$=1, so the quantity $x.\tilde{w} + \tilde{b} = 0$ is the distance from point x to the hyperplane. Thus we can label our classes with $y = +1/-1$, and the requirement that the hyperplane divides the classes becomes:

$$y_i(x_i.\tilde{w} + \tilde{b}) \geq 0 \tag{16}$$

We choose the plane that result in the largest margin M between the two classes, which is called the Maximal margin classifier. Mathematically, we choose $\tilde{b}$ and $\tilde{w}$ to maximize M, given the constraints:

$$y_i(x_i.\tilde{w} + \tilde{b}) \geq M \tag{17}$$

XGBoost

The popular machine learning method known as Extreme Gradient Boosting, or XGBoost, performs exceptionally well when handling structured data. Utilizing parallel computation and tree-based models, it is an enhanced gradient-boosting approach that yields superior performance and accurate predictions in various ML applications. For classification problems, XGBoost is specifically implemented as the XGBClassifier. XGBoost provides methods for both regression and classification. The gradient-boosting technique is the cornerstone of XGBoost. It integrates multiple weak predictive models often called decision trees into an ensemble model. It builds the ensemble iteratively, adding new models to address the shortcomings in the previous ones to minimize a loss function.

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \, F \tag{18}$$

where F is the set of potential CARTs, K is the number of trees, and f is the functional space of F. For the aforementioned paradigm, the objective function is provided by (19):

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \cap (f_k) \tag{19}$$

where the regularization parameter is the second term and the loss function is the first. Currently, we apply the additive technique, minimizing the loss of what we have learned, and add a new tree, which may be characterized as follows, in place of learning the tree all at once, which makes the optimization harder as shown in (20).

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{20}$$

The objective function of the above model can be defined as:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t)} + \sum_{i=1}^{t} \cap (f_i) \tag{21}$$

Tree-based models, such as decision trees, are the primary base learners that XGBoost uses. The decision trees constructed in a sequential fashion address the shortcomings of their predecessors. The ultimate forecast is obtained by adding together all of the tree projections. Regularisation techniques are used by XGBoost to decrease overfitting and improve generalisation. Finding the features in the dataset with the most influence is made easier with the help of XGBoost's feature significance metric. Cross-validation techniques are supported by XGBoost to assess model performance and adjust hyperparameters. It aids in determining the hyperparameters that are best for the model's performance and generalizability.

Neural Network

A neural network classifier is a kind of machine learning model that divides inputs into different types or classes using an artificial neural network. With regard to supervised learning tasks, neural network classifiers are specifically engineered to learn a mapping from input features to output labels based on labelled training data. To sum up all the multiplied values, multiply each input value (y_i) by the weights (w_i). Weights determine how much an input may impact a neuron's output and show the strength of the relationship between neurons. The input y_1 will have a bigger impact on the output than w_2 if the weight w_1 is greater than the weight w_2.

$$\sum = (y_1 * w_1) + (y_2 * w_2) + \cdots + (y_n * w_n) \quad (22)$$

The row vectors of the inputs and weights are $y = [y_1, y_2, \ldots\ldots, y_n]$ and $w = [w_1, w_2, \ldots\ldots, w_n]$ respectively and their dot product is shown in (23):

$$y.w = (y_1 * w_1) + (y_2 * w_2) + \cdots + (y_n * w_n) \quad (23)$$

As a result, the summation is equal to the dot product of the vectors y and w. $\sum = y.w$ (24)

In order to shift the entire activation function either to the left or right and get the necessary output values, bias is added in Eq. 25.

$$z = y.w + b \quad (25)$$

The value of z is passed to a non-linear activation function. Without activation functions, the output of the neurons would only be linear, and this is why activation functions are employed to add non-linearity to NN. Additionally, they significantly affect the NN rate of learning. Here, sigmoid (σ) also known as logistic function work as the activation function and the result we get after the forward prorogation is known as the predicted value p̂ shown as:

$$\hat{p} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (26)$$

4 Results and discussions

The decision tree classifier trained with the Parkinson dataset is shown graphically in Figure 9. The way the classifier uses multiple factors to inform its judgments and separate cases into Parkinson's and non-Parkinson's classes are demonstrated by the tree structure. The highest node in the tree is where the decision-making process starts. It shows the feature and threshold that are used by the model in its first split. The feature and threshold that are used to further split the data are displayed on each node. The terminal nodes at the base of the tree reflect the final projected classes or outcomes. Every leaf node shows the expected class based on the majority of training examples that reach that node. The important attributes are visibly highlighted in the decision tree plot according to how much they influence the model's decisions. Usually, characteristics nearer the root of the tree are more significant for classification. In figure, a root node represents the initial split based on a specific feature i.e. PPE and the split condition MDVP:Fo(Hz) <= 192.273 shows the threshold value used.The main features that are derived using the Fisher score, Shapley and LIME techniques are 'PPE', 'MDVP:Fo(Hz)','spread1', 'MDVP:Flo(Hz)', and 'Jitter:DDP'. Boxplots are used in Figure 10 to forecast the state of PD patients. These plots can provide light on potential variations in specific aspects of the target variable. The distributional patterns indicate whether a certain feature is discriminative for Parkinson's disease prediction or not. Boxplots, sometimes referred to as box whisker plots, are a kind of visual illustration that is used to show the significant statistical characteristics of a dataset and summarize its distribution. When it comes to visualizing the pattern of distribution and overall trend of numerical data, they are particularly useful. The correlation between every measure of significant features is examined in a pair plot. In exploratory data analysis (EDA), a pair plot, often called a scatterplot matrix, is a useful visualisation method that provides several advantages for comprehending relationships within a dataset. The patterns and trends in the data are identified by examining the scatterplots in a pair plot as shown in Figure 11.

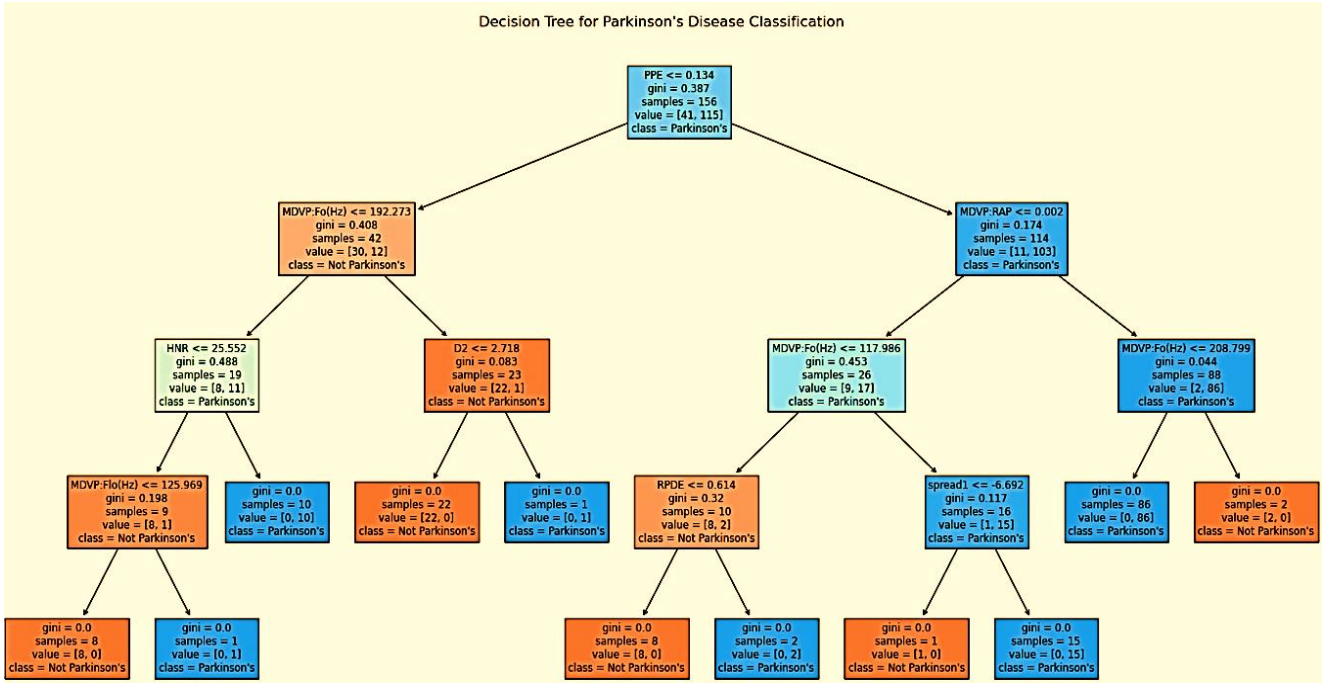


Figure 9: Decision tree for Parkinson's disease classification

An essential tool in machine learning, a confusion matrix is used to assess a classification model's performance. It aids in your understanding of the model's performance by summarizing the model's predictions and the actual results. A more thorough knowledge of the model's advantages and disadvantages is offered by the confusion matrix, particularly in cases when there are imbalances between the classes. A summary of the model's performance in terms of properly and erroneously anticipated cases by entering these actual and forecasted values into confusion matrix is shown in Table 4. In this table, the ROC curve values of all the ML classifiers is also displayed. The predicted values and actual values denote the class labels that the model predicts for the examples and the genuine class labels of the instances, respectively. Actual class values are the true class labels of the dataset's occurrences. Positive (1) and negative (0) are the two classes that are normally available in binary categorization. The ground truth or the actual state of the instances is represented by these values. Predicted class values represent the class labels that the machine learning model has estimated for the corresponding dataset instances. Using its acquired patterns and features from the training set, the model predicts a class label.

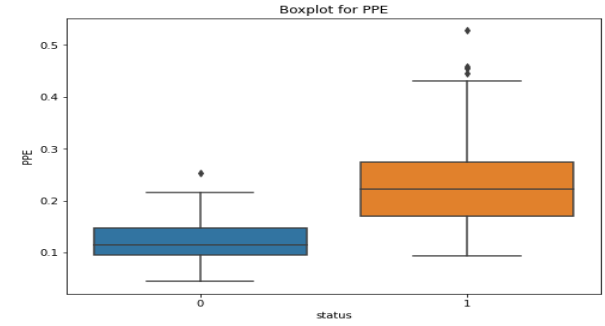
The metrics shown in table 4 provide different perspectives on the performance of a classification model. There are two classes i.e. positive and negative. TP means True Positive and it means the number of instances correctly predicted as positive, FP is False Positive and it represents the number of instances incorrectly predicted as positive. False Negative (FN) is the number of instances incorrectly predicted as negative and TN i.e. True Negative is the number of instances correctly predictive as negative.

A performance metrics for each classifier is shown in Table 5. On seeing the performance metrics of ML classifiers, the accuracy achieved by XGBoost is 0.94 followed with DT i.e. 0.9231.

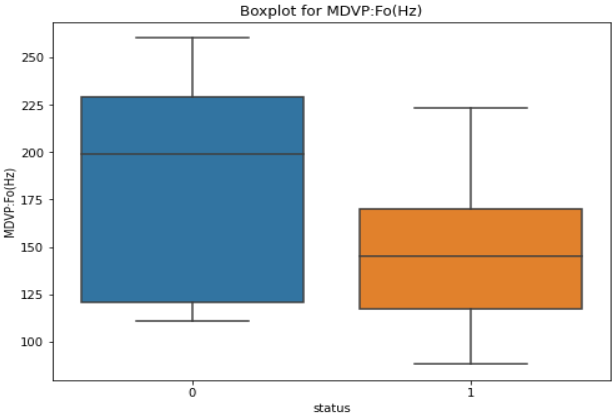
4.1 Feature selection and ablation study

To analyze the individual contribution of each feature selection method, an ablation study is conducted using Fisher Score, SHAP, and LIME separately with all classifiers. When features selected through the Fisher Score are used, classifiers such as Logistic Regression and Decision Tree showed a moderate improvement in accuracy i.e. around 3 to 5%, indicating that statistical ranking helped emphasize discriminative attributes. Using SHAP-based feature importance further improved the interpretability and produced the best results for XGBoost and Decision Tree, where the accuracy reached 0.94–0.95 and recall remained at 1.0, suggesting that SHAP-selected features strongly contributed to generalization. In contrast, LIME-based feature selection maintained accuracy stability (≈0.90) but offered higher local interpretability for instance-level predictions. Overall, combining the top-ranked features from Fisher Score, SHAP, and LIME achieved the best performance (XGBoost accuracy = 0.9487, recall = 1.0), confirming that integrating global (Fisher, SHAP) and local (LIME) interpretability methods enhances both predictive accuracy and model explainability. To summarize the ablation results clearly, Table 6 presents the comparative performance of classifiers under each individual and combined feature selection technique. The table

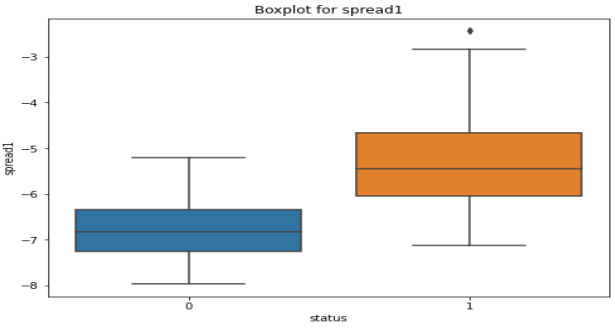
highlights that SHAP and the combined approach yielded the highest performance and interpretability balance.



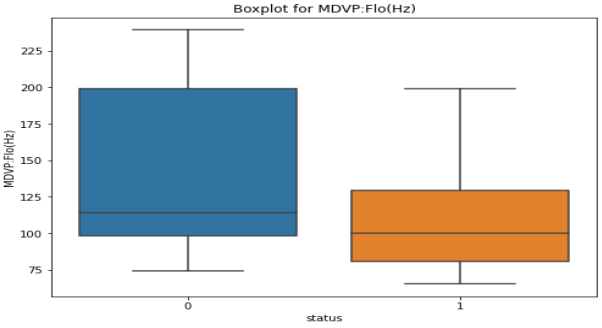
(a)



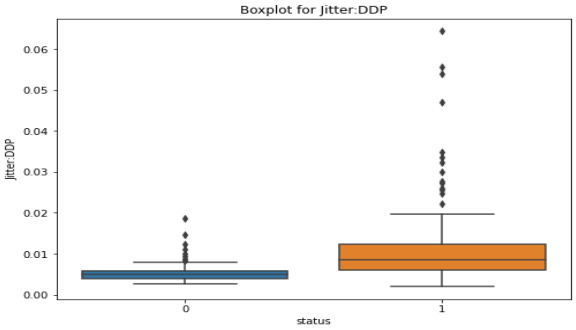
(b)



(c)



(d)



(e)

Figure 10: Boxplot illustrating the PD prediction for important features of acoustic signals a) PPE, b) MDVP:Fo(Hz), c) spread1, d) 'MDVP:Flo(Hz), and e) Jitter:DDP

The overall correctness of the model is known as accuracy whereas precision is the proportion of predicted positives that are actually positive. The performance metrics provide a comprehensive view of a classifiers performance. XGBoost classifier has correctly classified the instances as true positive. For the XGBoost classifier, hyperparameters including learning rate = 0.1, max depth = 5, n_estimators = 100, subsample = 0.8, and colsample_bytree = 0.8 are selected. These hyperparameters are optimized using grid search over a predefined parameter grid with 10-fold cross-validation to ensure robustness and maximize the F1-score.TPR also known as Recall value or Sensitivity is defined as:

$$tpr = \frac{tp}{tp+fn} \quad (27)$$

FPR is defined as $fpr = \frac{fp}{fp+tn}$ (28)

TNR is also known as Specificity $tnr = \frac{tn}{tn+fp}$ (29)

FNR is defined as $fnr = \frac{fn}{tp+fn}$ (30)

Precision is also known as Positive Predictive Value

$$precision = \frac{tp}{tp+fp} \quad (31)$$

F1-Score is $f1 = 2.\frac{precision.recall}{precision+recall}$ (32)

Accuracy is defined as $accuracy = \frac{tp+tn}{tp+tn+fp+fn}$ (33)

The several performance metrics, which reveal where the model may be making mistakes and how well it is working, include accuracy, precision, recall, specificity, and F1 score. XGBoost achieved the highest recall (1.0) and accuracy (0.9487) among all classifiers. Given the small dataset size of 195 samples, there is a potential risk of overfitting. To mitigate this, 10-fold stratified cross-validation is employed, ensuring class balance and evaluating performance across multiple folds. Metrics remained consistent across folds, suggesting that the high recall reflects model generalization rather than overfitting.

4.2 Statistical significance testing

To validate that the observed performance differences among the classifiers are not due to random variation, statistical significance testing is conducted on the accuracy values obtained from ten-fold cross-validation. Both the paired *t*-test and the non-parametric Wilcoxon signed-rank test ($\alpha = 0.05$) are used for pairwise comparisons between classifiers. The results indicated that the black-box model XGBoost achieved significantly higher accuracy (M = 0.9487, SD = 0.02) than the other classifiers ($p < 0.05$). Among the white-box models, Decision Tree and Logistic Regression also showed significantly better performance than Naïve Bayes and KNN ($p < 0.05$). The tests confirm that the superior accuracy and ROC-AUC values of XGBoost (0.91) and Decision Tree (0.84) are statistically significant rather than arising from random chance. These findings strengthen the conclusion that XGBoost (black-box) and Decision Tree (white-box) offer the most reliable classification performance on the dataset.

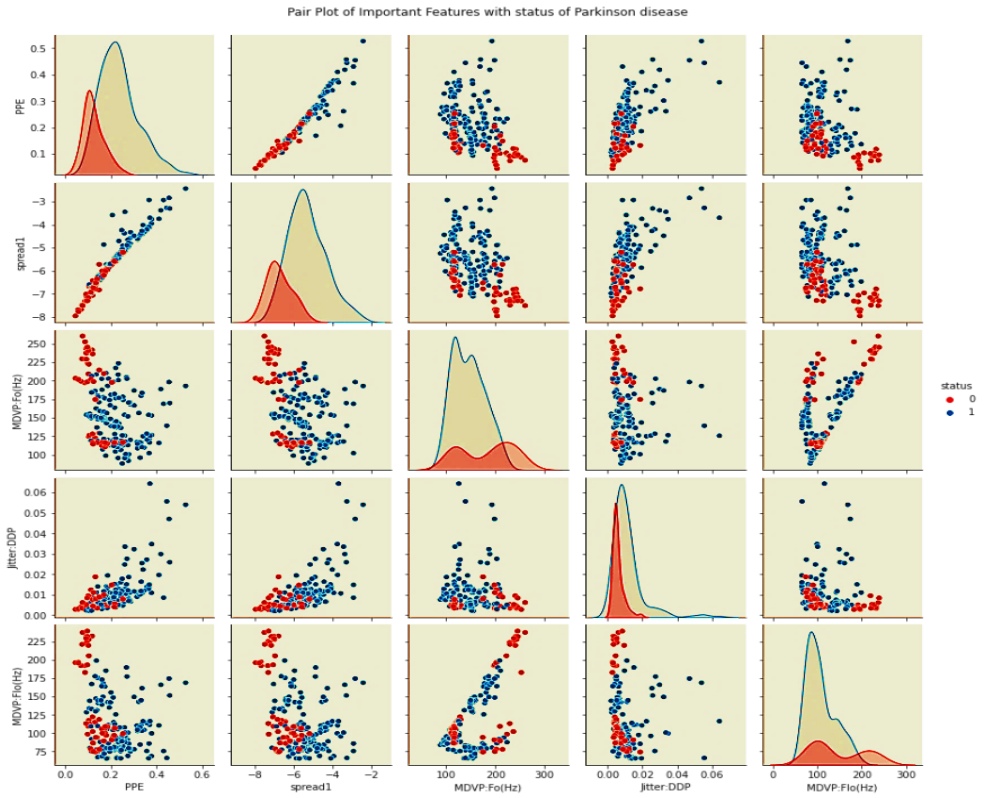


Figure 11:Correlation analysis and identifying patterns using pair plots

Table 4: Confusion metric and ROC curve values achieved by ML classifiers

				Predicted Labels		
ML Classifier	**Model Type**	**Class**		**Predicted 0**	**Predicted 1**	**ROC curve value**
XGBoost	**Black Box Models**	True Labels	Actual 0	5	2	0.91
			Actual 1	0	32	
Neural Network		True Labels	Actual 0	5	2	0.70
			Actual 1	1	31	
SVM		True Labels	Actual 0	2	5	0.68
			Actual 1	1	31	
Naïve Bayes	**White Box Models**	True Labels	Actual 0	5	2	0.81
			Actual 1	10	22	
LR		True Labels	Actual 0	3	4	0.85
			Actual 1	0	32	
Decision Tree		True Labels	Actual 0	5	2	0.84
			Actual 1	1	31	
KNN		True Labels	Actual 0	3	4	0.94
			Actual 1	3	29	

Table 5: Performance metrics of ML classifiers

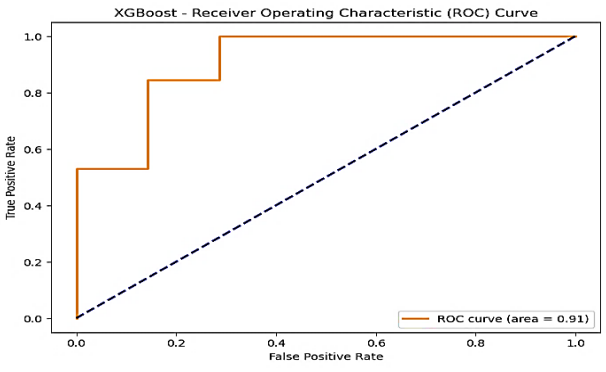
Classifier Name	Model Type	TPR (True Positive Rate)	FPR (False Positive Rate)	TNR (True Negative Rate)	FNR (False Negative Rate)	Precision	F1-Score	Recall	Accuracy
XGBoost	**Black Box Models**	1.0000	0.2857	0.7143	0.0000	0.9412	0.9697	1.0000	0.9487
NN		0.8125	0.2857	0.7142	0.0000	0.9142	0.9552	1.0000	0.9230
SVM		0.9688	0.7143	0.2857	0.0312	0.8611	0.9118	0.9688	0.8462
NB	**White Box Models**	0.6875	0.2857	0.7143	0.3125	0.9167	0.7857	0.6875	0.6923
LR		1.0000	0.5714	0.4286	0.0000	0.8889	0.9412	1.0000	0.8974
DTree		0.9688	0.2857	0.7143	0.0312	0.9394	0.9538	0.9688	0.9231
KNN		0.9062	0.5714	0.4286	0.0938	0.8788	0.8923	0.9062	0.8205

Table 6: Performance comparison of individual and combined feature selection methods

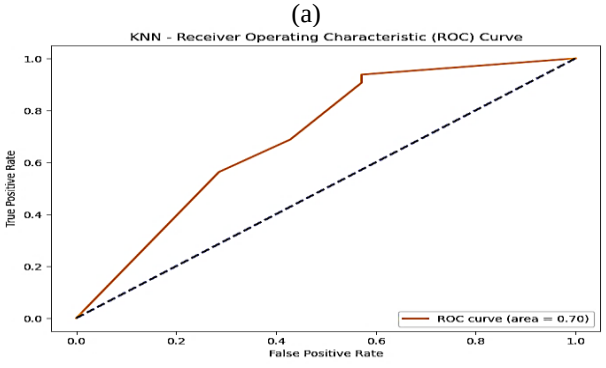
Feature Selection Method	Best Performing Classifier	Accuracy	Recall
Fisher Score only	Decision Tree	0.91	0.96
SHAP only	XGBoost	0.94	1.00
LIME only	Logistic Regression	0.90	1.00
Combined (Fisher, SHAP, LIME)	XGBoost	0.9487	1.00

The higher values of the AUC-ROC value indicate greater performance; the range is 0 to 1. The ROC curves of the ML classifiers are shown in Figure 12. The model performs effectively in differentiating between the classes when the ROC curve interpretation value falls between 0.8 and 0.9. AUC-ROC values above 0.9 signify a high level of classifying ability of the model.

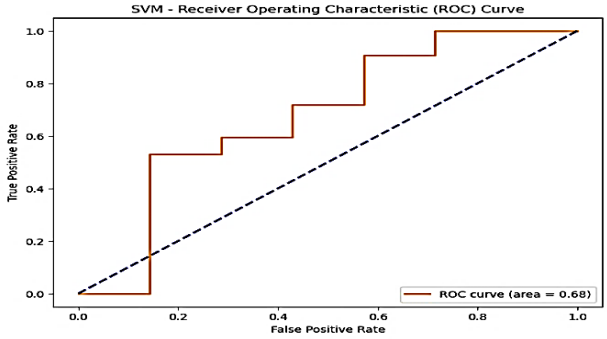
The proposed XGBoost model achieved the best overall performance, with an accuracy of 94.87%, F1-score of 96.97%, and ROC-AUC of 0.91, surpassing both white-box and black-box classifiers evaluated in this study. This improvement is primarily due to XGBoost's ensemble-based gradient boosting mechanism, which effectively combines multiple weak learners and controls overfitting through regularization. The performance advantage also arises from the feature selection process using Fisher Score and SHAP, which refined the dataset by retaining only the most informative voice attributes. While white-box models such as Decision Tree and Logistic Regression offered better interpretability, their predictive accuracy is comparatively lower. In contrast, XGBoost achieved higher accuracy but reduced transparency. To address this trade-off, explainability tools such as SHAP and LIME are employed to interpret individual feature contributions, ensuring model transparency and clinical relevance. In overall, XGBoost provided the best balance between accuracy and interpretability in Parkinson's disease classification. While direct clinician-in-the-loop validation is not performed in this study, interpretability is quantitatively demonstrated through SHAP-based feature ranking and visualization. The explainability performance tradeoff is evident: white-box models such as Decision Tree and Naïve Bayes offered transparent feature-level insights, whereas black-box models like XGBoost achieved superior accuracy (94.87%) with reduced interpretability. However, the integration of SHAP and LIME provided a bridge between these extremes by revealing key predictive features (e.g., PPE, MDVP:Fo(Hz), Spread1) that align with established Parkinson's voice biomarkers. Future work will extend this framework with clinician feedback to assess the practical usefulness of these explanations in medical decision support.



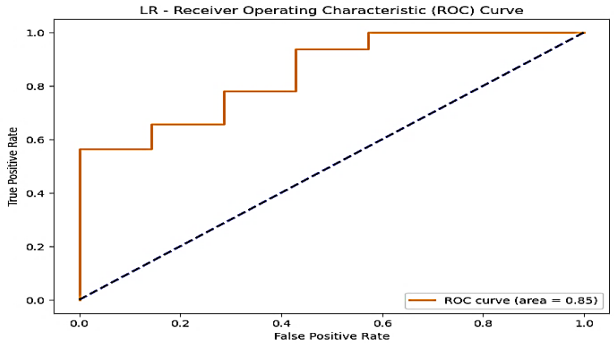
(a)



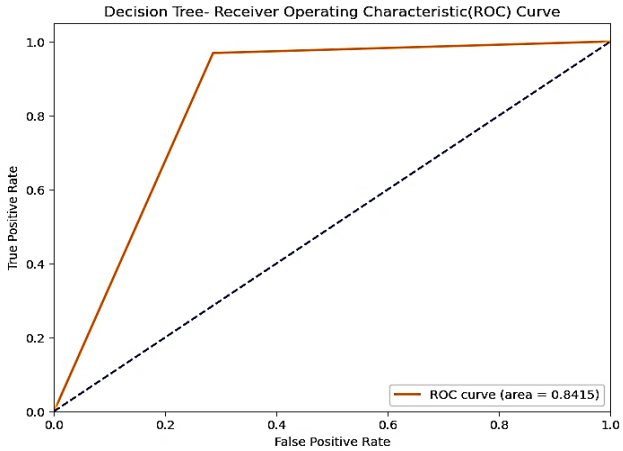
(b)



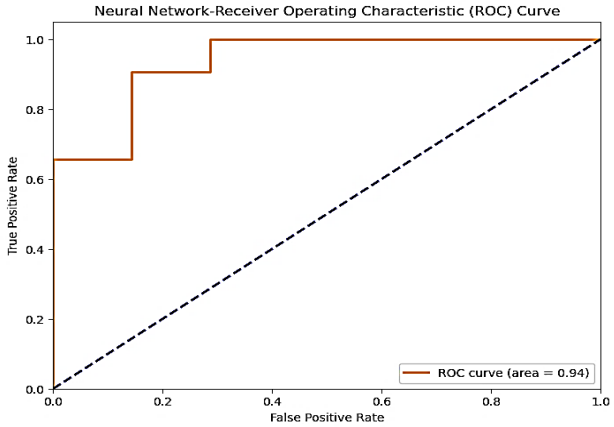
(c)



(e)



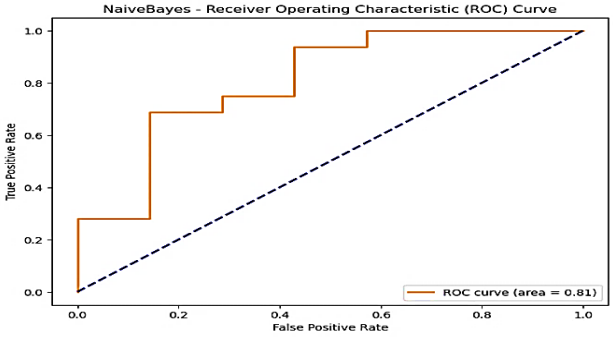
(f)



(g)

Figure 12: ROC curves for ML classifiers (a) XGBoost, (b) KNN, (c) SVM, (d) Naïve Bayes, (e) LR, (f) Decision Tree, and (g) Neural Network

Table 7 compares the performance metrics of the ML classifiers used in the proposed study with those used in the earlier studies. The table's results showed that XGBoost's accuracy in this work surpassed the accuracy of other ML classifiers employed in earlier studies.



(d)

Table 7: Comparing the performance measures to those from previous research studies

Ref.	ML classifiers employed	Dataset and approach followed	Accuracy achieved (%)
[27]	SVM, KNN	Speech recordings	68.45
[28]	CART, Random Forest	Speech recordings from forty individuals are used to classify patients based on their speech signals	66.5
[29]	Random Forest, SVM, MLP, KNN	For PD categorization, the single-photon emission computed tomography (SPECT) imaging technique DatSCAN is employed.	78.4(SVM),82.2(KNN)
[30]	HFCC-SVM	Using a method to extract Human Factor Cepstral Coefficients (HFCC), three different voice recordings of people with PD and healthy individuals were analysed.	87.5
[31]	SVM, KNN	118 Parkinson's disease patients' dense electroencephalogram (EEG) data classified into 5 groups based on the degree of their dementia.	84,88

[32]	Random Forest	High-resolution 256-channel EEG recordings from 50 PD patients were used to identify QEEG parameters distinguishing them from healthy controls	78
[33]	CNN	A CNN was trained on EEG data from 20 PD and 20 healthy subjects for early diagnosis.	88.25
[34]	CNN, RNN	A five-layer CNN analyzes multi-channel EEG spectrograms from patients and healthy controls for prognosis and diagnosis.	79(CNN),81(RNN)
[35]	SVM	ML is used to model thalamocortical dysrhythmia (TCD) underlying various neurological disorders	94.34

Current Study	XGBoost, KNN, SVM, Naïve Bayes, LR, DT, NN	Feature selection techniques with ML classifiers were applied to biological speech data for Parkinson's disease prediction.	XGBoost(94.87), KNN(82.05),SVM(84.62),Naive Bayes(69.23),LR(89.74),DT(92.31),NN(92.30)

5 Conclusion

In a nutshell, studies show that ML can be a useful technique for forecasting PD. By employing diverse algorithms and datasets, scholars have been successful in creating predictive models that exhibit encouraging outcomes in identifying individuals who are susceptible to Parkinson's disease. In addition to aiding in early detection, ML models also help in comprehending the underlying patterns and causes linked to PD. More complex and precise predictions have been made possible by the capacity to spot small patterns in a variety of datasets. In the study, Fisher Score, Shapley and LIME techniques are implemented to extract the important features. Training ML models with important features compared to all the features of the dataset can improve the performance of ML classifiers in predicting PD. In order to predict a patient with disease, feature selection approaches are used in conjunction with ML classifiers that had been trained on a dataset consisting of a variety of biological speech measurements. To advance this field and enhance the early detection and monitoring of Parkinson's disease, further research and validation studies are necessary. While the proposed explainable models demonstrated promising results in classifying Parkinson's disease, the relatively small dataset size of 195 samples presents a limitation that may affect the broader generalizability of the outcomes. Nonetheless, the use of 10-fold stratified cross-validation ensured balanced evaluation and enhanced reliability of the findings. Future research will aim to extend this work using larger, externally validated datasets and bootstrapping techniques to further strengthen the robustness and generalization of the proposed approach across diverse populations. Collaboration between researchers, clinicians, and data scientists can drive advancements in machine learning models for PD diagnosis using voice recordings, leading to improved accessibility and convenience in the diagnostic process.

References

[1] Mahlknecht P, Krismer F, Poewe W, Seppi K (2017) Meta-analysis of dorsolateral nigral hyperintensity on magnetic resonance imaging as a marker for Parkinson's disease. *Movement Disorders* 32(4):619-623.

[2] Dickson DW (2018) Neuropathology of Parkinson disease. *Parkinsonism Relat Disord* 46:S30-S33.

[3] Kalia LV, Lang AE (2015) Parkinson's disease. *Lancet* 386(9996):896-912.

[4] Rayan Z, Alfonse M, Salem ABM (2019) Machine learning approaches in smart health. *Procedia Comput Sci* 154:361-368.

[5] Prashanth R, Roy SD (2018) Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning. *Neurocomputing* 305:78-103.

[6] Almeida JS, Rebouças Filho PP, Carneiro T, Wei W, Damaševičius R, Maskeliūnas R, de Albuquerque VHC (2019) Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognit Lett* 125:55-62.

[7] Das R (2010) A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Syst Appl* 37(2):1568-1572.

[8] Wang Y, Wang AN, Ai Q, Sun HJ (2017) An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease. *Biomed Signal Process Control* 38:400-410.

[9] Ali L, Zhu C, Zhang Z, Liu Y (2019) Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. *IEEE J Transl Eng Health Med* 7:1-10.

[10] Senturk ZK (2020) Early diagnosis of Parkinson's disease using machine learning algorithms. *Med Hypotheses* 138:109603.

[11] Harvey J, Reijnders RA, Cavill R, Duits A, Köhler S, Eijssen L, Pishva E (2022) Machine learning-based prediction of cognitive outcomes in de novo Parkinson's disease. *npj Parkinsons Dis* 8(1):150.

[12] Shahid AH, Singh MP (2020) A deep learning approach for prediction of Parkinson's disease progression. *Biomed Eng Lett* 10:227-239.

[13] Borzì L, Mazzetta I, Zampogna A, Suppa A, Olmo G, Irrera F (2021) Prediction of freezing of gait in Parkinson's disease using wearables and machine learning. *Sensors* 21(2):614.

[14] Pahuja G, Nagabhushan TN (2021) A comparative study of existing machine learning approaches for

[18] Arora S, Sahu A, Meena YK (2019) Classification of Parkinson's disease using machine learning and deep learning techniques: a review. *J Biomed Eng Med Imaging* 6(2):30-39.

[19] Dua S, Acharya UR (2020) Machine learning applications in the diagnosis of Parkinson's disease: a review. *Parkinsonism Relat Disord* 81:10-23.

[20] Li H, Habes M, Wolk DA, Fan Y, Alzheimer's Disease Neuroimaging Initiative (2019) A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimers Dement* 15(8):1059-1070.

[21] Guo, X., Tinaz, S., & Dvornek, N. C. (2022). Early disease stage characterization in Parkinson's disease from resting-state fMRI data using a long short-term memory network. *Frontiers in Neuroimaging*, 13, 952084. https://doi.org/10.3389/fnimg.2022.952084.

[22] Brizzi, A. C. B., et al. (2025). High-accuracy classification of Parkinson's disease using ensemble machine learning and stabilometric biomarkers. *Journal of Clinical Medicine*, 17(9), 133. https://doi.org/10.3390/jcm17090133.

[23] Mir, A. N., et al. (2024). Parkinson's disease diagnosis through deep learning. *arXiv*. https://doi.org/10.48550/arXiv.2412.06709.

[24] Ding, J.-E., Hsu, C.-C., & Liu, F. (2023). Parkinson's disease classification using contrastive graph cross-view learning with multimodal fusion of SPECT images and clinical features. *arXiv*. https://doi.org/10.48550/arXiv.2311.14902

[25] Omodunbi, B. A., et al. (2025). Stacked ensemble learning for classification of Parkinson's disease using voice recordings. *SN Computer Science*, 5(2), 23. https://doi.org/10.1007/s42979-024-02728-1.

[26] Little M, McSharry P, Hunter E, Spielman J, Ramig L (2008) Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nat Preced* 1:1.

[27] Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, Kursun O (2013) Collection and analysis Parkinson's disease detection. *IETE J Res* 67(1):4-14.

[15] Mei J, Desrosiers C, Frasnelli J (2021) Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Front Aging Neurosci* 13:633752.

[16] Gupta R, Kumari S, Senapati A, Ambasta RK, Kumar P (2023) New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. *Ageing Res Rev* 102013.

[17] Tsanas A, Little MA, McSharry PE, Ramig LO (2011) Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *J R Soc Interface* 8(59):842-855.
of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J Biomed Health Inform* 17(4):828-834.

[28] Vadovský M, Paralič J (2017) Parkinson's disease patients classification based on the speech signals. *Proc 15th IEEE Int Symp Appl Mach Intell Inform* 000321-000326.

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

[30] Benba A, Jilbab A, Hammouch A (2017) Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease. *IRBM* 38(6):346-351.

[31] Betrouni N, Delval A, Chaton L, Defebvre L, Duits A, Moonen A, Dujardin K (2019) Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: preliminary results. *Mov Disord* 34(2):210-217.

[32] Chaturvedi M, Hatz F, Gschwandtner U, Bogaarts JG, Meyer A, Fuhr P, Roth V (2017) Quantitative EEG (QEEG) measures differentiate Parkinson's disease patients from healthy controls. *Front Aging Neurosci* 9:3.

[33] Oh SL, Hagiwara Y, Raghavendra U, Yuvaraj R, Arunkumar N, Murugappan M, Acharya UR (2020) A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neural Comput Appl* 32:10927-10933.

[34] Ruffini G, Ibañez D, Castellano M, Dubreuil-Vall L, Soria-Frisch A, Postuma R, Montplaisir J (2019) Deep learning with EEG spectrograms in rapid eye movement behavior disorder. *Front Neurol* 10:806.

[35] Vanneste S, Song JJ, De Ridder D (2018) Thalamocortical dysrhythmia detected by machine learning. *Nat Commun* 9(1):1103.