GMR
The Original

AN EXPLAINABLE MACHINE LEARNING FRAMEWORK WITH ROBUST CROSS-VALIDATION FOR PARKINSON'S DISEASE PREDICTION USING VOICE BIOMARKERS

Sonu Rana[1], Sreelekha paul[2], Shivnath Ghosh[3], Sanjay Kumar[4], Ahinshubra Nandi[5], Anusmita Bhattacharjee[6].

[1,2,3,4,5,6]Department of Computer Science and Engineering, Brainware University, Barasat, Kolkata, West Bengal, Pin 700125, India
sranasonu85@gmail.com, paulsree350@gmail.com, shivghosh.cs@gmail.com, dsk.cs@brainwareuniversity.ac.in, ahinsubhranandi@gmail.com, bhattacharjeeanusmita3@gmail.com

Abstract
A chronic neurological condition that impairs mobility is Parkinson's disease. Slower motions, tremors, and stiffness might be symptoms of Parkinson's disease. Although a precise diagnostic test does not yet exist, machine learning can be used to estimate a person's likelihood of having Parkinson's disease based on particular biomarkers. The purpose of this article is to forecast Parkinson's disease using machine learning algorithms.
Early detection may be delayed by traditional clinical diagnosis, which mostly relies on medical knowledge and symptom monitoring. Thus, machine learning (ML)-based automated prediction systems can greatly enhance early diagnosis and assist medical professionals.
This work uses clinical and biological speech datasets to develop a machine learning-based classification model for Parkinson's disease identification. XGBoost (Extreme Gradient Boosting), Random Forest, and Support Vector Machine (SVM) are three potent supervised learning techniques used by the system. Data preparation procedures, such as feature scaling, normalisation, cleaning, and feature selection, are carried out to raise the dataset's quality and boost model performance. In order to properly evaluate the dataset, it is then separated into training and testing sets. Metrics including F1-score, confusion matrix, recall, accuracy, and precision are used to assess the models' performance. In general, XGBoost and Random Forest offer more accuracy and resilience, according to comparative studies, whereas SVM handles complicated categorisation boundaries well.

Keywords: Parkinson's Disease, Machine Learning, Parkinson's Disease Prediction, Support Vector Machine (SVM), Random Forest, XGBoost, Voice Signal Analysis, Early Diagnosis.

1. INTRODUCTION

Parkinson’s Disease (PD) is a progressive condition that affects the brain's nerve cells, specifically the dopaminergic neurons in the substantia nigra region. It is the second most common type of neurodegenerative disorder after Alzheimer’s disease and mostly occurs in older adults. Medically, PD presents with movement-related symptoms like resting tremors, slowed movement, stiffness, and balance issues, as well as non-movement-related symptoms such as memory problems, sleep issues, and problems with automatic bodily functions. Diagnosing Parkinson’s Disease early and accurately is difficult because its symptoms can be similar to those of other brain disorders, and the way the disease progresses varies from person to person. In the past, doctors mainly relied on their experience and physical exams to diagnose such conditions, which can sometimes lead to inconsistencies [1,2,3].
Recently, there have been developments in using computer-based systems to help in diagnosis. These systems use data from medical tests and apply machine learning techniques to identify patterns that are hard to detect through traditional means. Machine learning techniques have shown promise in helping to classify diseases by understanding complex connections in health data. Specific types of machine learning, such as supervised learning, include methods like Support Vector Machines (SVM), Random Forest (RF), and gradient boosting. Ensemble methods like Random Forest and Extreme Gradient Boosting (XGBoost) are effective because they improve accuracy by combining multiple models and handle a lot of data effectively. Kernel-based classifiers like SVM are also useful because they can manage large data sets and find the best way to separate different types of data. Although there have been many studies on using machine learning to detect Parkinson’s Disease, there is a need for an organized comparison of different machine learning methods using standard evaluation tools. It is also important to interpret how well these models perform using tools like confusion matrices, ROC-AUC measurements, and analysis of feature relationships to ensure that the models are reliable and can be used in medical settings [5,7,8,30].

This study introduces a complete machine learning approach for determining if someone has Parkinson's Disease using a dataset with 195 samples and 24 features related to health. The data is split into training and testing groups using an 80:20 ratio. Three models—XGBoost, Random Forest, and Support Vector Machine—are tested. Their performance is measured using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC. A correlation heatmap is also used to look at how different features might influence the results of the classification [9,11,24,25].

2 LITERATURE REVIEW

The use of machine learning techniques for Parkinson's disease early detection and prediction has been the subject of several studies. To increase diagnostic accuracy, researchers have used a variety of datasets, such as voice recordings, handwriting patterns, and clinical assessments. When it comes to differentiating Parkinson's sufferers from healthy people, conventional machine learning methods like Support Vector Machine (SVM), Random Forest (RF), and XGBoost have shown encouraging results.

Year	Reference / Study	Algorithms Evaluated	Dataset & Methods	Key Findings
2019	Handwriting classifier (SVM focus) — Ujjwal et al.	SVM	Handwriting features for PD vs control	SVM used to rank age/sex-dependent features; achieved 79–84 % accuracy (SVM used for feature importance ranking)[6]
2022	Voice PD diagnosis (ensemble + XGBoost) — Ghaheri et al.	XGBoost, LightGBM, GB, Bagging	Voice signal dataset with SHAP selection	Ensemble with XGBoost achieved ~85 % accuracy with explainable features; SHAP used for interpretability[7]
2024	Early PD detection — Adiga et al.	XGBoost, Random Forest, SVM, NN	UCI & Oxford Parkinson's datasets, voice features	ML models (RF, XGBoost, SVM) achieved high diagnostic accuracy; multimodal features improve results
2024	Multi-algorithm comparison — Y. Yan	XGBoost, RF, SVM, AdaBoost, BP Neural	Multiple classifiers evaluated on PD datasets	XGBoost and ensemble models included; confirmed performance benefits of tree-based methods[21]
2025	Systematic comprehensive ML review — Shokrpour et al.	RF, SVM, others	133 PD papers reviewed (2021-2024)	Detailed survey of datasets, ML models, evaluation challenges; RF and SVM frequently used [30]
2025	Explainable AI for PD prediction — Esan et al.	RF, XGBoost, SVM, ensemble	Large multimodal dataset (demographics, clinical)	Random Forest with feature selection + XAI achieved ~93 % accuracy and AUC ~0.97 [31]
2025	ML & DL review — Rabie & Akhloufi	SVM, RF, CNN, etc.	PD detection & progression review	Discussed traditional and DL models, noted SVM & RF effectiveness with hybrid data [32]

Table 1. Literature Review of Parkinson's Disease Prediction Studies

3 DATASET- OVERVIEW AND CROSS-VALIDATION 10-FOLD

This study's dataset, known as the Oxford Parkinson's Disease Dataset, was acquired from the UCI Machine Learning Repository. It is made up of biological voice measurements taken from both healthy persons and those with Parkinson's disease. The collection includes 195 speech recordings from 31 persons, 8 of whom were healthy controls and 23 of whom had a Parkinson's disease diagnosis. There are 24 features in each instance, such as nonlinear dynamical complexity, amplitude change, and different vocal frequency measurements[1,2]. A value of 1 denotes the existence of Parkinson's disease, whereas a value of 0 denotes a healthy person. The target variable, status, is binary in nature. This dataset offers a useful standard for assessing machine learning algorithms for disease prediction because of its dependability and extensive application in Parkinson's disease research. In this work, 10-fold cross-validation was used to produce a more accurate estimate of model performance. Ten roughly equal subsets (folds) were randomly selected from the whole dataset. One fold served as the testing set for each iteration, and the model was trained using the other nine folds. To ensure that each fold was used as the testing set exactly once, this procedure was carried out ten times[20,21]. To offer a reliable assessment of the machine learning models, the performance indicators from all 10 iterations were averaged. For each of the ten cross-validation folds, the messages "Fold 1 saved" through "Fold 10 saved" show that the train and test datasets were successfully created and saved for later model training and assessment [35,36].

4 METHODOLOGY

Using clinical and voice-related characteristics, this study suggests a machine learning-based paradigm for Parkinson's disease prediction. To enhance data quality and model performance, the gathered dataset is put through preparation procedures such as data cleaning, normalisation, and feature selection. Next, an 80:20 ratio is used to separate the dataset into training and testing sets. Three supervised machine learning algorithms—Random Forest, Support Vector Machine (SVM), and XGBoost—are put into practice and assessed. To choose the best method for Parkinson's disease prediction, assessment measures like accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC are used to evaluate these models' performance.

4.1 XGBOOST

XGBoost is a strong machine learning algorithm that uses Gradient Boosting techniques. It functions by bringing together several weak decision trees to build a strong predictive model that has better accuracy. In finding Parkinson's disease, XGBoost is useful for examining complicated medical data and boosts how well we can classify it. The algorithm lowers prediction mistakes by improving the loss function through a method called Gradient Descent. It also has regularisation techniques that help stop overfitting and make the model more trustworthy [17].

$$\mathbf{F(x)} = \sum_{\mathbf{m=1}}^{\mathbf{M}} \mathbf{h_m(x)}$$

Where, F(x) = final prediction, M = number of trees in the group, $h_m(x)$ = prediction from each tree

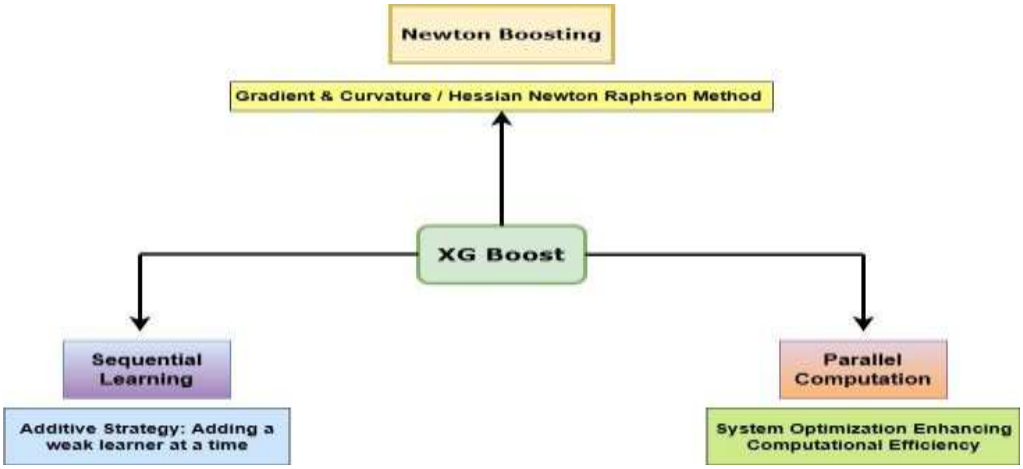

Figure 1: XGBoost

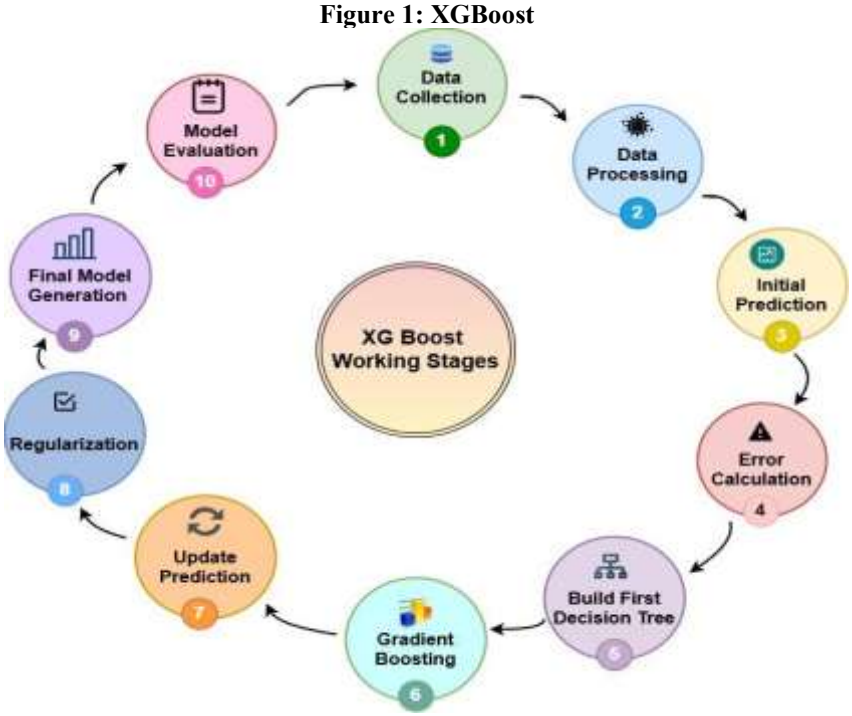

Figure 2: XG Boost Working Process

4.2 Random Forest

Random Forest is a group learning method that exercises multiple decision trees for categorising binary outcomes. Different small groups of data and features are habituated to following each tree. This assists in jack legging the trees more varied and keeps them from apt too closely to the training data. Random Forest assists in categorising patients by coming across different medical features and symptoms when diagnosing Parkinson's disease. The final result is determined by the most votes from all the decision trees. This method can offer reliable and precise results for medical diagnosis systems [18].

$\hat{Y} = \text{Majority Vote}(T_1, T_2, T_3, \ldots, T_n)$ Where, T_n = Each decision tree

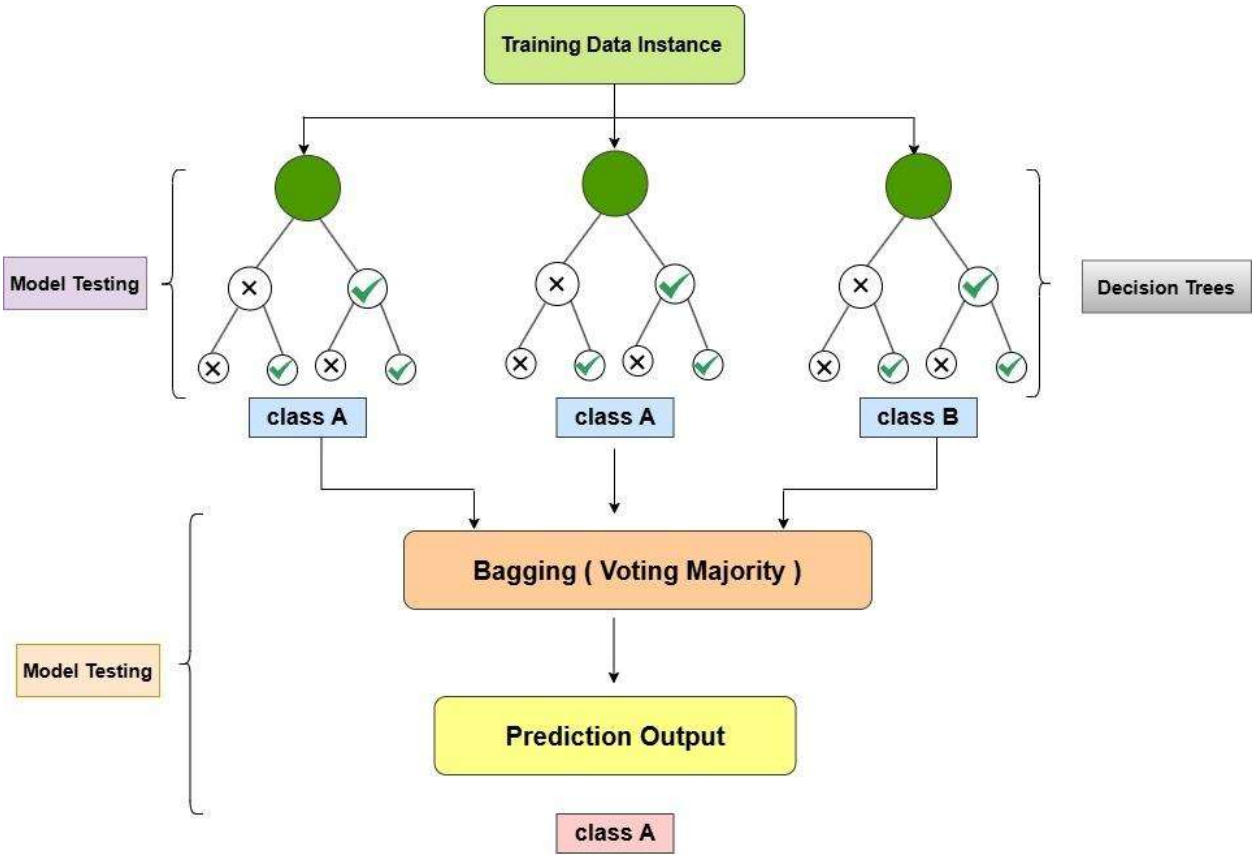

Figure 3: Random Forest Algorithm in Machine Learning

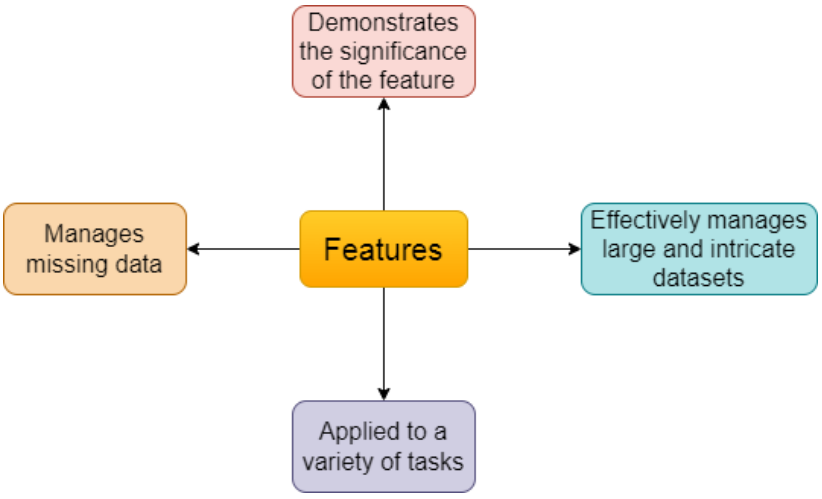

Figure 4: Features of Random Forest

4.3 SVM

Supervised learning algorithm known as Support Vector Machine (SVM) primarily used for classification. It does this by identifying an optimal hyperplane that maximizes the margin between the different classes of data. For Parkinson's disease detection, SVM is used to classify patients with Parkinson's and healthy people according to their medical traits [19]. The algorithm works well for low- and high-dimensional data. Kernel functions can also be used to better classify nonlinear data.$f(x) = w^t x + b$Where,w = weight vector,b = bias term

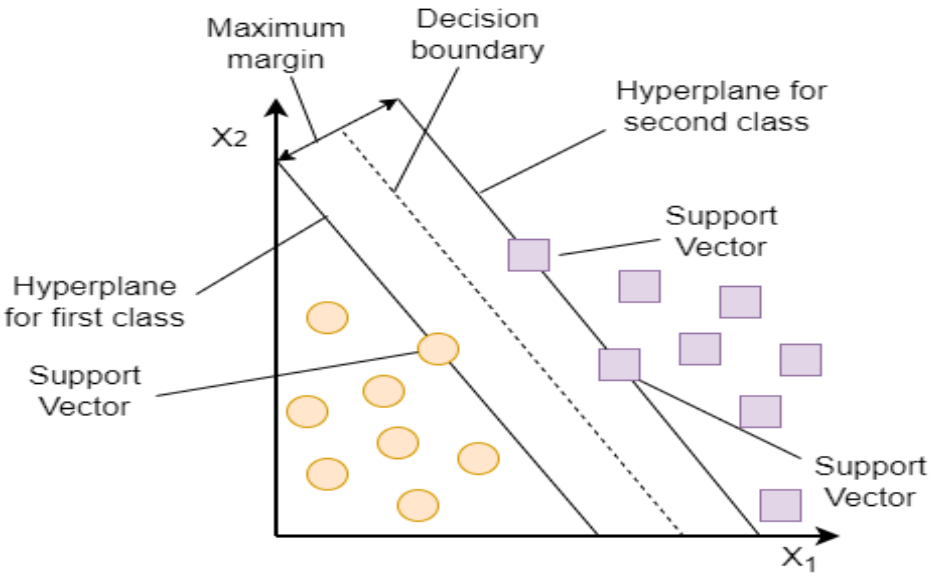

Figure 5: Support Vectors & Hyperplanes

4.4 Confusion Matrix

A Confusion Matrix is a table of performance metrics in machine learning that helps evaluate the performance of classification models. It is used to compare actual values to predicted values and to aid in the determination of how well the model is working. In Parkinson's disease prediction, it is employed to determine the accuracy of the system in detecting whether patients suffer from Parkinson's or not. The confusion matrix displays key evaluation metrics like accuracy, precision, recall, and F1 score. It allows researchers to find out the reliability and mistakes of the classification model.

Metric	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Measures the overall proportion of correctly classified instances.
Precision	TP / (TP + FP)	Indicates the proportion of positive predictions that are actually positive.
Recall (Sensitivity)	TP / (TP + FN)	Measures the ability of the model to correctly identify positive instances.
F1-Score	(2 × Precision × Recall) / (Precision + Recall)	Represents the harmonic mean of precision and recall.

Table 2. Evaluation Metrics Used for Model Performance Assessment

Table 2 explains TP (True Positive): The model predicts positive, and the prediction is correct, TN (True Negative): The model predicts negative, and the prediction is correct. FP (False Positive): The model predicts positive, but the prediction is incorrect. FN (False Negative): The model predicts negative, but the prediction is incorrect[20,33].

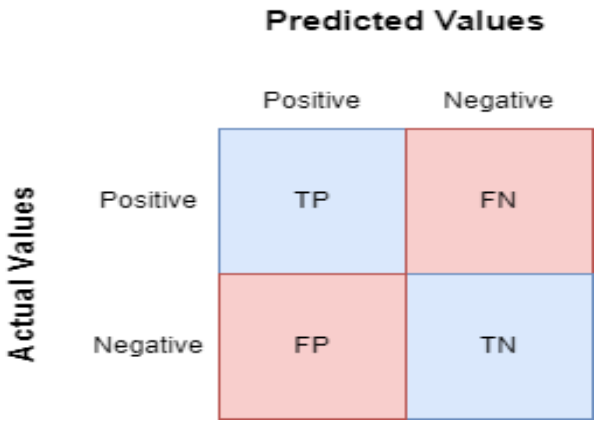

Figure 6: Confusion Matrix

4.5 ROC-AUC

ROC stands for Receiver Operating Characteristic Curve. It is a graphical representation that is commonly used to assess the performance of a classification model in distinguishing between two classes. In your project: Class 1 → Parkinson's Disease Class 0 → Healthy Person. The ROC curve shows the relationship between the true positive rate (TPR) and the false positive rate (FPR). The threshold values for at different values. AUC stands for Area Under the Curve. It is the area covered by the ROC curve. It tells that the model discriminates well between positive and negative classes[31].

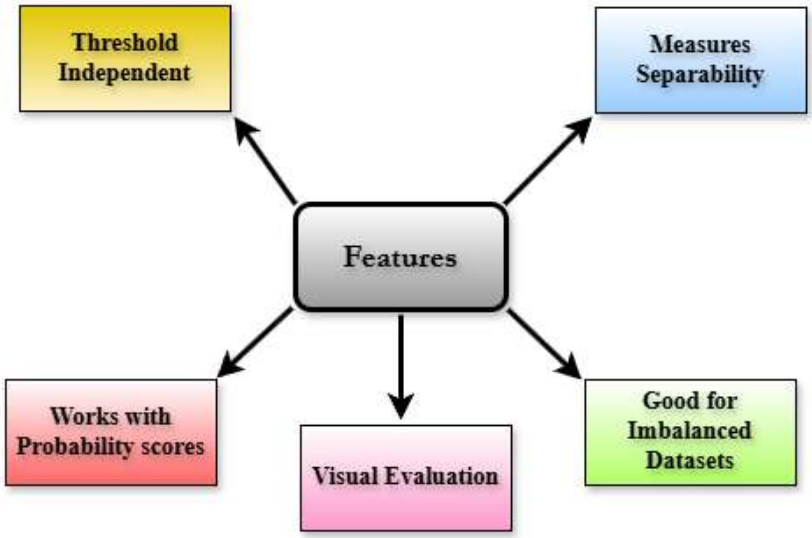

Figure 7: Features of ROC-AUC

Metric	Formula	Description
True Positive Rate (TPR)	TP / (TP + FN)	Measures the proportion of actual positive cases correctly identified by the model.
False Positive Rate (FPR)	FP / (FP + TN)	Measures the proportion of actual negative cases incorrectly classified as positive.

Table 3. ROC-AUC Evaluation Metrics

TPR = True Positive Rate, FPR = False Positive Rate, TP = True Positive, FN = False Negative, FP = False Positive, TN = True Negative. AUC-ROC curve allows us to understand the power of the classification model to separate the two classes[21].

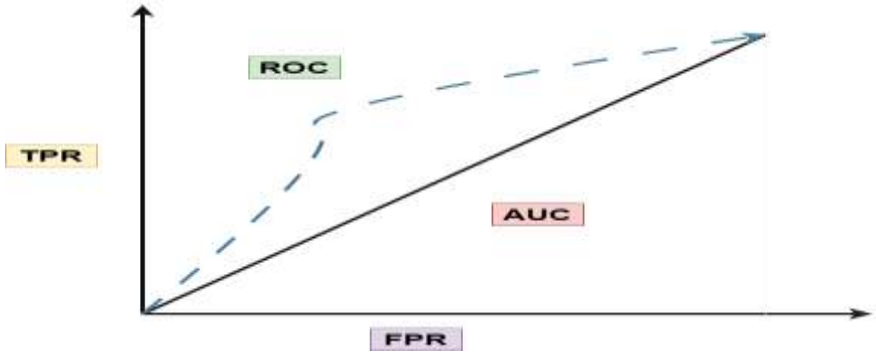

Figure 8 : Graph of ROC-AUC

It calculates the sum of the area under ROC curve and above the X axis. Mathematical formula for AUC curve:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Where,AUC= Area Under CurveTPR= True Positive RateFPR= False Positive Rate ∫ = Integration (area calculation)d(FPR) = Small change in FPR [31,32]

4.6 Binary Features Classification Using SVC

In Binary Classification, the output can take only two values to be considered two classes: 1 represents Parkinson Present and 0 represents Parkinson Not Present. The model is trained to identify patterns in the data and to make a prediction of one of these two categories [19].

SVC stands for Support Vector Classifier. It is a supervised machine learning algorithm that is based on Support Vector Machine (SVM), and it is used for Classification, Pattern recognition, and disease prediction. SVC tries to find the best decision boundary that takes the place of the dividing line between two classes. Parkinson's disease-related voice anomalies are taught to the SVC model. Patients are divided into Diseased or Healthy and it utilizes measurements of voice signals. The main concept of the SVM algorithm is to construct the hyperplane separating two classes with maximum margin. This margin is the distance from the hyperplane to the nearest data points (support vectors) on each side[24].

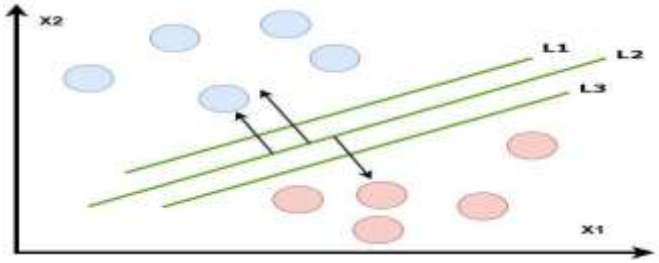

Figure 9: Several hyperplanes that divide the data into two classes.

Component	Mathematical Expression	Description
SVC Decision Function	$f(x) = w^T x + b$	Computes the decision score for classifying an input sample.
Classification Rule	$f(x) \geq 0 \rightarrow$ Parkinson (Class 1) $f(x) < 0 \rightarrow$ Healthy (Class 0)	Assigns samples to Parkinson's or Healthy classes.
Optimization Objective	Minimize: $\frac{1}{2}\|w\|^2 + C \Sigma Ei$	Finds the maximum-margin hyperplane while minimizing classification errors.
Constraint	$yi(w^T xi + b) \geq 1 - Ei$	Ensures correct classification with allowance for errors.

Table 4. Mathematical Formulation of Support Vector Classifier (SVC)

Where x : Input feature vector, w: Weight vector learned by the model, b: Bias term,f(x): Decision score,$\|w\|^2$: Margin control term, C: Penalty parameter, Ei: Misclassification error (slack variable),yi: True class label

4.7 Correlation Heatmap

The correlation heat map is a graphical representation that displays how strongly two features are related to each other. It's employed in machine learning to learn: Feature relationships, Redundant features, and important features. A correlation heatmap is a 2-dimensional visualization of a correlation matrix of several variables. It displays the values of the correlation in various shades of cells, making the pattern and relationships in data easier to read. The intensity of the colour of each cell indicates the level of correlation:1 (or close to 1) shows high positive relationship (dark colours),0 represents no relationship (neutral colors) and -1 (or close to -1) represents Very negative correlation (light colors).

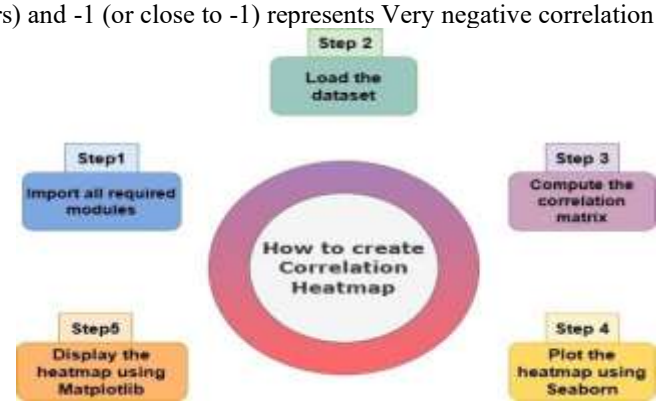

Figure 10: How to create a Correlation Heatmap

Pearson Correlation Coefficient:

$r = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \sqrt{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]}$ where r is the description$_i$ is each value of feature X, y_i is each value of feature Y, $\bar{x}$ is the mean of feature X, $\bar{y}$ is the mean of feature Y, and r_{ij} is the correlation value between variables i and j in the correlation matrix [25].

5 METHOD ARCHITECTURE

Figure 11 illustrates the overall architecture of the proposed explainable machine learning framework for Parkinson's disease prediction using voice biomarkers. Initially, the UCI Parkinson's Disease dataset, comprising 195 voice recordings represented by 22 acoustic features, was acquired and categorized into healthy individuals and Parkinson's disease

patients. During the preprocessing stage, missing value analysis, feature scaling, label encoding, and an 80:20 train-test split were performed to prepare the data for classification. Subsequently, four machine learning algorithms, namely Support Vector Machine (SVM), Random Forest (RF), XGBoost, and Stochastic Gradient Descent (SGD) classifier, were trained and evaluated. The performance of these models was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. To ensure the robustness and generalizability of the proposed framework, 10-fold cross-validation was conducted, and the mean performance of each classifier was compared. Furthermore, SHAP (SHapley Additive exPlanations) analysis was employed to interpret the contribution of individual voice biomarkers and provide both global and local explanations of model predictions. Finally, the comparative analysis demonstrated the effectiveness of the proposed framework, with XGBoost achieving the best predictive performance, thereby highlighting the potential of explainable machine learning techniques for reliable and clinically interpretable Parkinson's disease prediction.

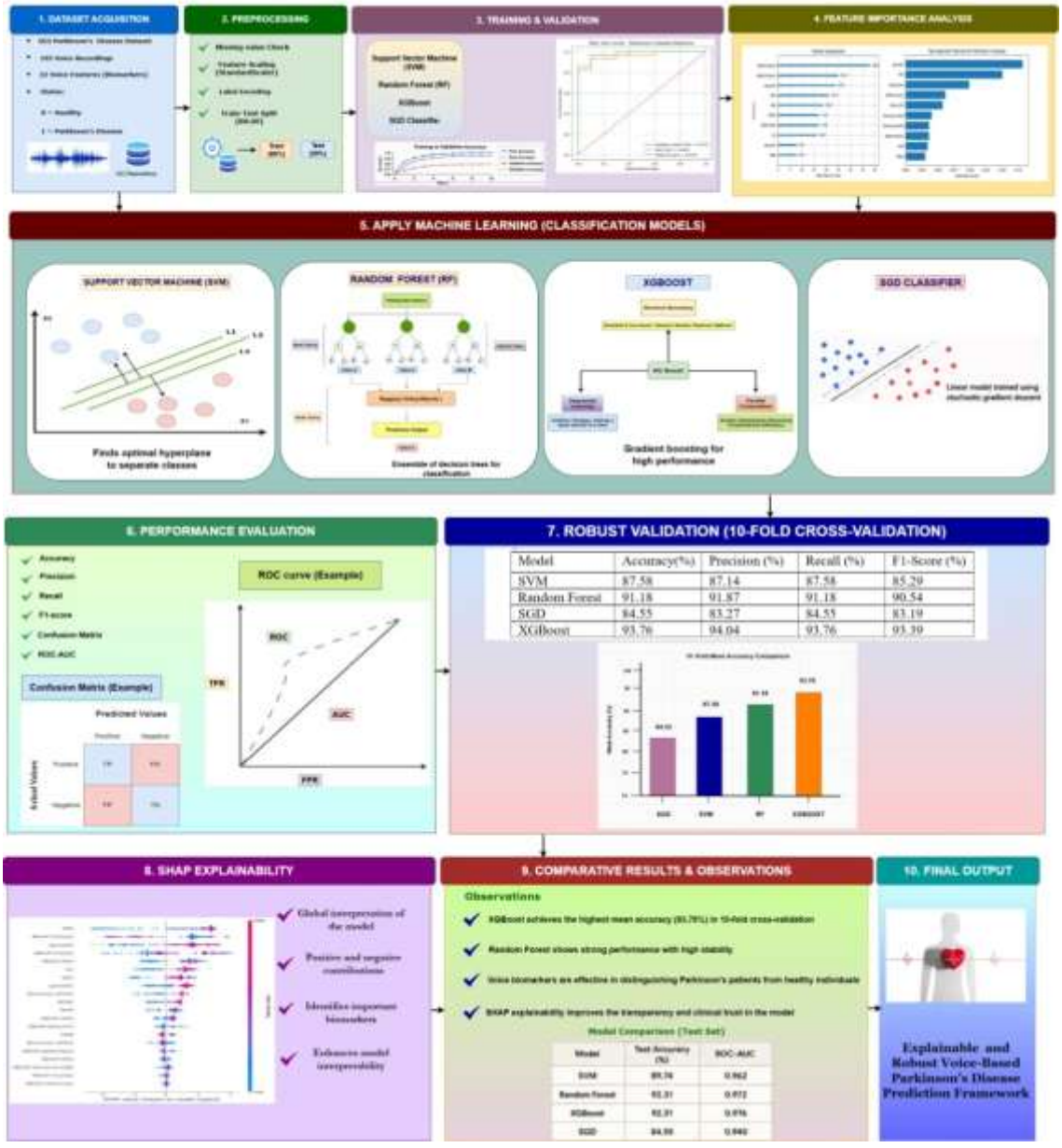

Figure 11: Overall Method Architecture

6 DATA AND RESULT ANALYSIS

The Table-5 explaining that the predictive accuracy of three machine learning models—Random Forest, Support Vector Machine (SVM), and XGBoost—was evaluated to predict Parkinson's disease from vocal characteristics. Performance was excellent, with an accuracy of almost 92% obtained by both the Random Forest and XGBoost classifiers. The test set accuracy was highest for the Random Forest and XGBoost classifiers,

Model	Data	Accuracy	Precision	Recall	F1-score
Random Forest	Train	1.0000	1.0000	1.0000	1.0000
Random Forest	Test	0.9231	0.9219	0.9231	0.9217
SVM	Train	0.9808	0.9807	0.9808	0.9807
SVM	Test	0.8974	0.8974	0.8974	0.8974
XGBoost	Train	1.0000	1.0000	1.0000	1.0000
XGBoost	Test	0.9231	0.9219	0.9231	0.9217

Table 5: All Model Comparison Table

with values close to 92.31%, and the precision, recall, and F1-score were also close to 92%. SVM was competitive, albeit slightly less accurate, with 89.74% on the test set. Though they've brought 100% accuracy in training, the test results of Random Forest and XGBoost shows good prediction capability with slight overfitting. Overall, the results of the study indicate that ensemble methods (Random Forest, XGBoost) are proving to be more effective than SVM and are the most accurate methods for correctly diagnosing PD [30].

6.1 ROC-AUC Curve

The ROC-AUC Curve was created to evaluate the performance and compare the performance of three machine learning models like Random Forest, Support Vector Machine (SVM) and XGBoost to detect Parkinson's Disease. In this graph the X-axis shows the False Positive Rate and Y-axis shows the True Positive Rate. The AUC values of Random Forest Classifier is 0.972,SVM is 0.962 and XGBoost is 0.976, those values indicate the differentiation of each model between Parkinson's patients and healthy individuals. Among all three models XGBoost has the highest AUC value (0.976) and it indicates the best performance to predict a positive case. If the AUC value score is greater than 0.95, can conclude that chosen voice features are highly predictive and reliable for predicting Parkinson's disease.

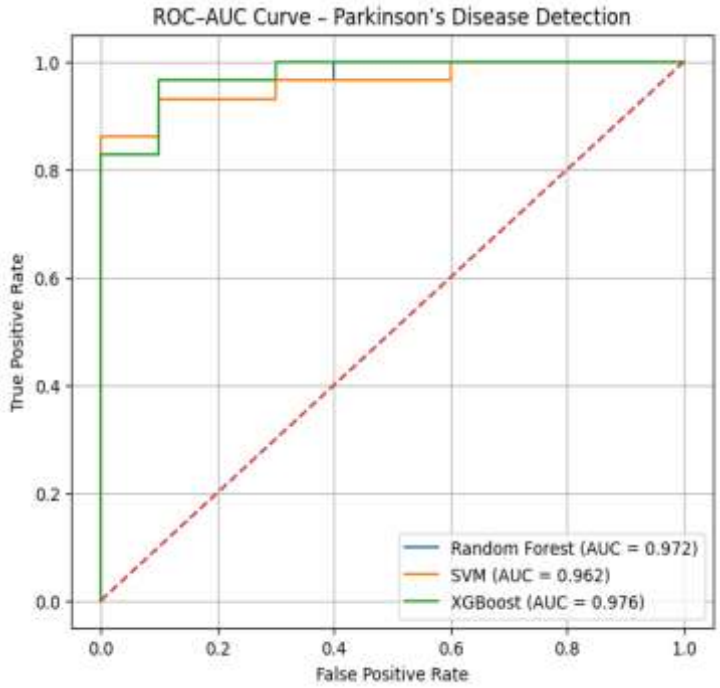

Figure 12: ROC-AUC Curve Comparison

6.2 Binary Classification Using SVC

Based on the output, get the performance of the Support Vector Classifier (SVC) model for Binary Classification of Parkinson's Disease. There are 195 samples and 24 columns, the input features are 22 and the last columns contain the target information. The status indicates the target variable and it has two categories like 1 and 0. 1 indicates the patient has Parkinson's Disease and 0 indicates healthy.147 samples shows 1, means there is presence of parkinson disease and 48 samples shows 0, means there are no presence of parkinson disease. Using the selected voice related features, the SVC

model correctly found 89.74% (nearly 90%) of the unseen data.The classification report gives an in-depth analysis of each class. In the healthy subjects (class 0), the model had a precision 0.80, a recall of 0.80, and an F1-score of 0.80, with 80% accuracy.The model's performance on the Parkinson's class (1) resulted in a precision of 0.93, a recall of 0.93, and an F1-score of 0.93, indicates its high accuracy to identify Parkinson's patients.High score of class 1 show the model is better for detecting disease-related voice pattern. To show all over performance, the model achieved a macro average of 0.87 across both classes and the size-adjusted weighted average is 0.90. The final output [0 1 1 1 0 1 1 1 0 1] is the first 10 predictions made by the SVC model, with 0 indicating a healthy person and 1 indicating a Parkinson's patient. Overall these results show that the selected voice features provided excellent data for binary classification. The SVC model effectively separates healthy and Parkinson's subjects with high accuracy and makes itself a trustworthy framework for predicting disease.

```
Dataset Shape: (195, 24)
Number of features: 22
Class Distribution:
status
1    147
0     48
Name: count, dtype: int64
Test Accuracy: 0.8974358974358975
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.80      0.80        10
           1       0.93      0.93      0.93        29

    accuracy                           0.90        39
   macro avg       0.87      0.87      0.87        39
weighted avg       0.90      0.90      0.90        39

[0 1 1 1 0 1 1 1 0 1]
```

Figure 13: Binary classification of the dataset

6.3 Correlation Heatmap Analysis

The correlation heatmap illustrates the relationship among the extracted voice features to predict Parkinson's disease. In the first heatmap, the correlation matrix of all voice features in the data set is shown, while the second heatmap contains only the most important features selected in the feature selection process, namely PPE, RPDE, MDVP:Jitter(%), MDVP:Shimmer, HNR, spread1, and spread2. The degree and direction of feature correlations are visually represented by color,where red,blue and white denote strong positive,strong negative and negligible association respectively. Large red blocks indicate strong positive correlations between measures tracking comparable Parkinson's disease-related voice problems. The second, detailed heatmap shows an almost perfect correlation (0.96) between spread1 and PPE, indicating that these two important characteristics provide nearly the same predictive information.MDVP:Jitter(%) and MDVP:Shimmer are highly correlated (0.77), suggesting that there is a strong relationship between the two parameters in Parkinson's patients and the other features, such as MDVP:Shimmer, MDVP:Jitter(%) and PPE, show strong negative correlations with HNR, with correlation coefficients of -0.84, -0.73 and -0.69 respectively. This indicates that speech quality decreases as vocal abnormalities become more noticeable, a trend that is known to happen in Parkinson's disease. The total heat map confirms disease-related patterns in the data set, shows significant correlations between the vocal biomarkers, and offers a way to choose the most useful features for machine learning categorization. The findings proved that the voice is a trustworthy predictor of Parkinson's disease.

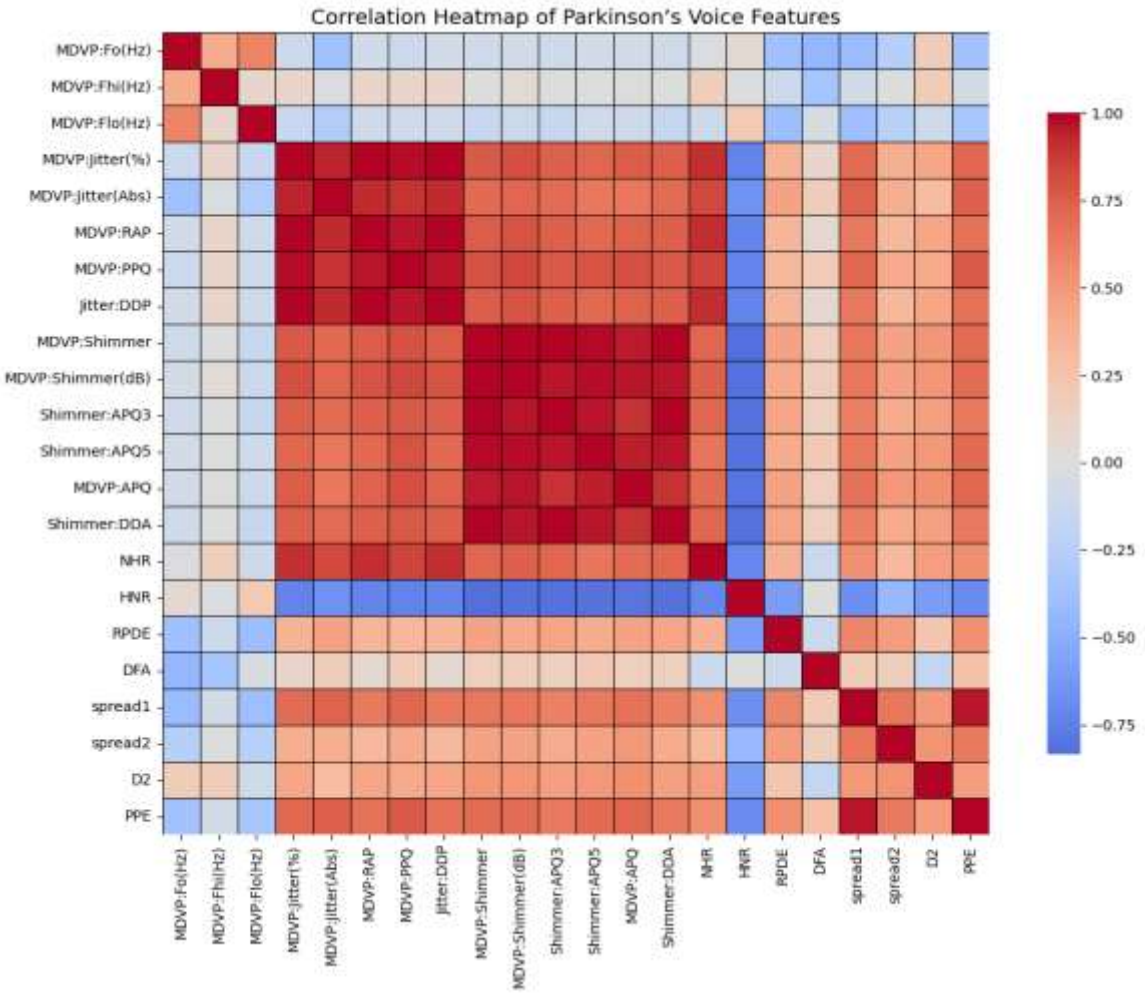

Figure 14: Correlation Heatmap of Disease Features

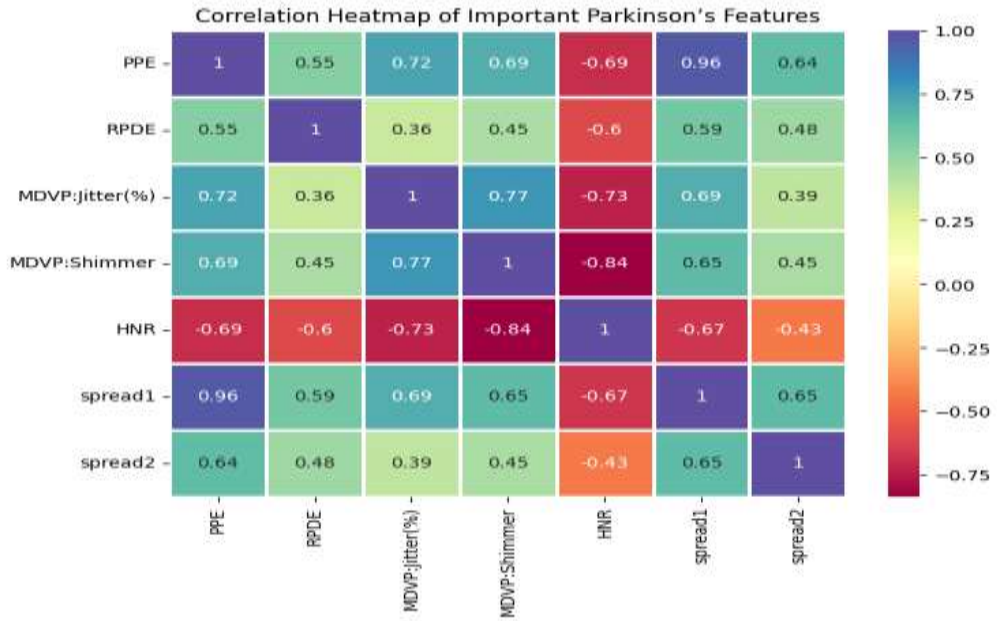

Figure 15: Top Important Features Numeric Value

6.4 Top Important Features for Parkinson's Disease Prediction

The top 10 characteristics are presented in this table 6 that the machine learning model has identified as being important in predicting the presence of the disease Parkinson's. The significance score shows the importance of the features with respect to the classification capacity of the model. The features are sorted by their relevance score and within the features the relevance score for Spread1 is highest (0.1819) and is the most relevant in differentiating PD patients from normal

subjects. Pitch changes (PPE (Pitch Period Entropy)) is the second most important feature with significance value of 0.1505, which shows that voice pitch changes are significant features for illness identification. MDVP:APQ is the third most ranked measure, with a score of 0.0989, indicating that changes in the amplitude of the speech signal are important for PD .

Feature	Importance Score
spread1	0.1819
PPE	0.1505
MDVP:APQ	0.0989
MDVP:Fo(Hz)	0.0618
Jitter:DDP	0.0577
Shimmer:APQ5	0.0405
Shimmer:APQ3	0.0362
MDVP:Flo(Hz)	0.0361
HNR	0.0342
RPDE	0.0305

Table 6: Top Important Features for Parkinson's Disease Prediction

The other features, such as MDVP: Fo(Hz), Jitter: DDP, Shimmer: APQ5, Shimmer: APQ3, MDVP: Flo(Hz), HNR and RPDE, help to make the prediction by measuring various components of voice abnormalities – fluctuations in frequency, instability of the voice, vibrations, noise level and nonlinear properties of speech. A higher relevance value for the spread1, PPE and MDVP: APQ traits suggests that the most relevant information in these traits is related to Parkinson's disease. Overall, the feature importance analysis results also show that the voice indicators are quite effective in predicting illnesses and are able to determine the most important ones that influence the prediction of the machine learning model.
The most significant feature, spread1 (0.1819), indicates a robust correlation between Parkinson's disease and changes in nonlinear speech dynamics. PPE (0.1505): Pitch Period Entropy is a key sign of Parkinsonian vocal impairment and reflects abnormalities in speech patterns. MDVP:APQ (0.0989): Captures voice instability frequently seen in Parkinson's patients and measures amplitude perturbation in speech. MDVP:Fo(Hz) (0.0618): Shows the average fundamental frequency of the voice, suggesting that variations in vocal pitch aid in the identification of illness. Jitter:DDP (0.0577): Indicates poor vocal control and quantifies cycle-to-cycle fluctuations in voice frequency. Shimmer:APQ5 (0.0405) and Shimmer:APQ3 (0.0362): Measure short-term amplitude changes in speech that reveal instability and tremor in the voice. The minimum vocal frequency, or MDVP:Flo(Hz) (0.0361), offers more details regarding pitch irregularities. HNR (0.0342): Parkinson's disease is frequently linked to decreased voice quality, which is measured by the Harmonics-to-Noise Ratio. RPDE (0.0305): Recurrence Period Density Entropy is a nonlinear biomarker that captures the intricacy and irregularity of speech signals.

6.5 Confusion matrix Comparison of All model

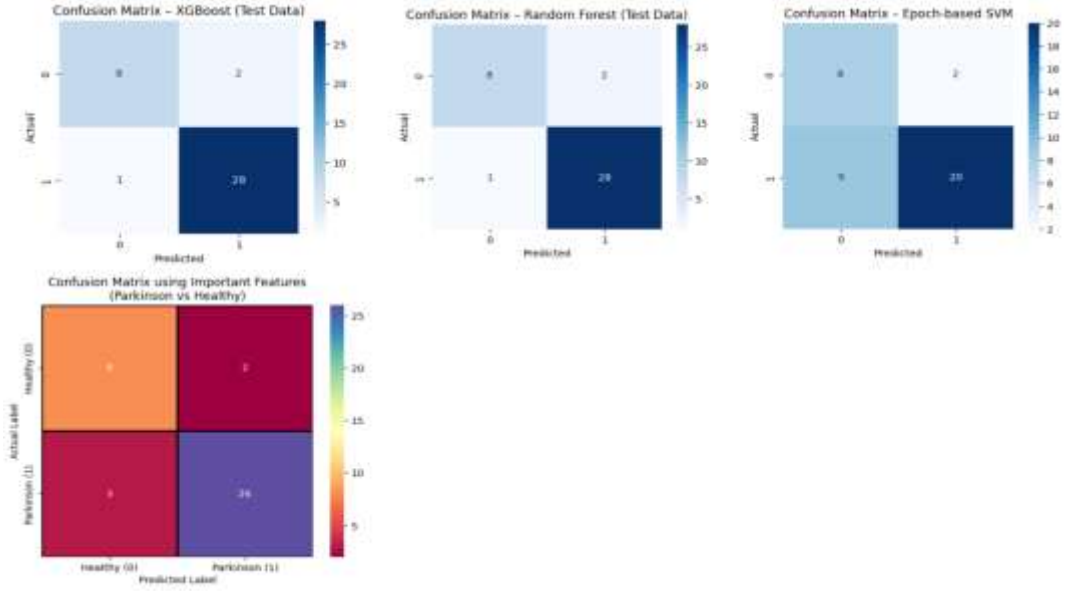

Figure 16: All Confusion Matrix Comparison

The study on the confusion matrix showed that the accuracy of the classification of the samples using the Random Forest algorithm and XGBoost algorithm was 36 of 39 samples with only 3 samples being misclassified. High sensitivity and reliability for illness prediction were demonstrated by both models' effective detection of 28 out of 29 Parkinson's disease patients. The SVM model was not very well suited, with a poor detection rate of nine patients with Parkinson's disease and misclassification of eleven patients. With 34 accurate predictions out of 39 samples, XGBoost trained on the most significant features maintained good performance, indicating that a smaller feature set may successfully retain predictive accuracy. In conclusion, this study determined the best models to detect Parkinson's disease, which were Random Forest and XGBoost, as explained in Figure 16.

6.6 Accuracy Comparison

Fig. 17 below shows the training and testing accuracies of the three machine learning models used to identify Parkinson's disease. The SVM model outperformed the other models in terms of training and testing accuracy as it performed the least well with a training accuracy of nearly 81% and a testing accuracy of 72%. Random Forest on the other hand, along with XGBoost, appeared to have higher classification accuracy, with 100% training accuracy and 92.31% testing accuracy.

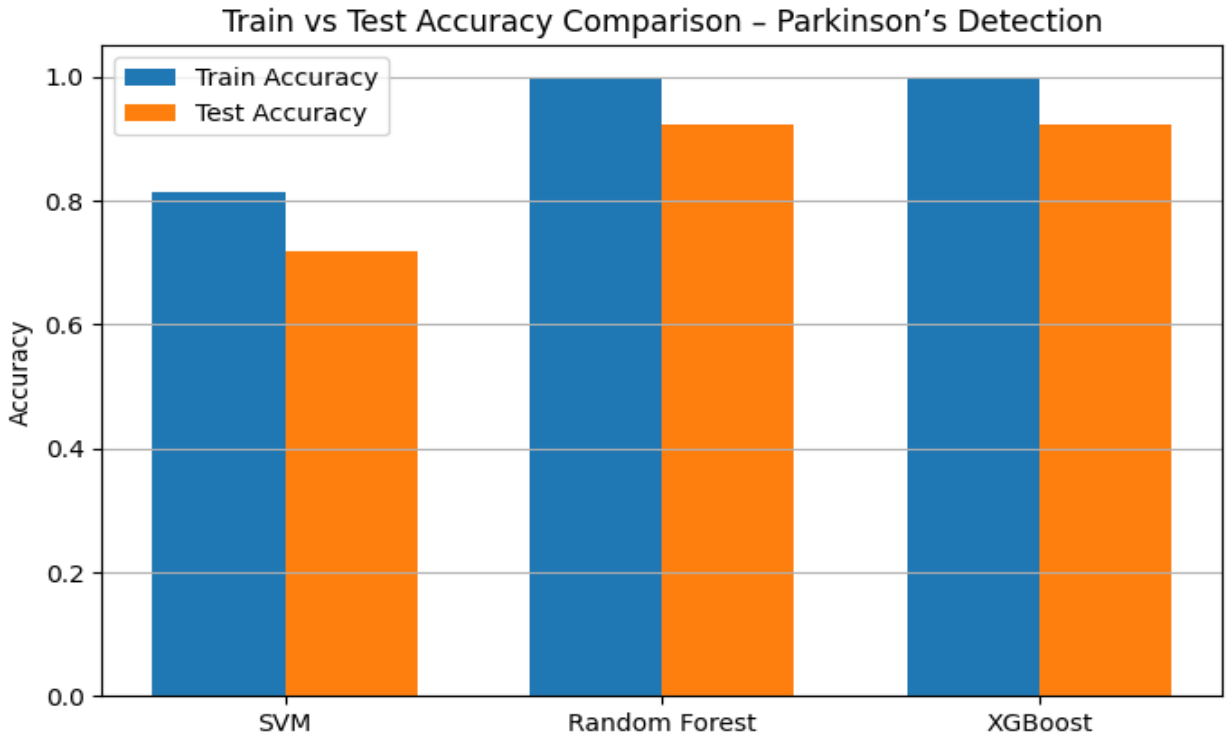

Figure 17: Train vs Test Accuracy Comparison Graph

The high test accuracy in Random Forest and XGBoost indicates good generalisation to unknown data, but a difference between the training and test accuracy means there is a bit of overfitting. The performance of ensemble-based models is significantly better than that of SVM, and the most successful models in predicting Parkinson's disease using voice-based characteristics are Random Forest and XGBoost.

6.7 10 Fold Cross-Validation Mean Accuracy

Table 7 explains that 10-Fold Cross-Validation Mean Accuracy is important in Parkinson's disease detection because it provides a more reliable estimate of how well the proposed model will perform on unseen patient data. In medical diagnosis applications, datasets are often limited in size, and the use of a single train-test split may lead to biased or unstable results. The reported accuracy may vary depending on how the data are divided. In 10-fold cross-validation, the dataset is divided into ten equal subsets. During each iteration, nine subsets are used to train the model, while the remaining subset is used for testing. The accuracy of the average values or mean values obtained from all ten iterations. are averaged to calculate the mean accuracy [32].

Model	Fold Accuracies (%)	Mean Accuracy (%)	Std. Dev (±)	Interpretation
SVM	90.00, 90.00, 95.00, 90.00, 95.00, 89.47, 78.95, 84.21, 84.21, 78.95	87.58	5.50	Moderate performance with stable results.

Random Forest	95.00, 100.00, 90.00, 95.00, 95.00, 89.47, 94.74, 94.74, 78.95, 78.95	91.18	6.71	High predictive accuracy with slight variation.
SGD	90.00, 90.00, 90.00, 75.00, 90.00, 84.21, 84.21, 89.47, 78.95, 73.68	84.55	6.20	Lowest performance among all models.
XGBoost	95.00, 100.00, 90.00, 100.00, 100.00, 89.47, 94.74, 100.00, 84.21, 84.21	93.76	6.11	Best overall performance for PD prediction.

Table 7: Performance comparison of SVM, Random Forest, SGD, and XGBoost models using 10-fold cross-validation on the Oxford Parkinson's Disease dataset.

6.8 SHAP Feature

Random Forest (91.18%), SVM (87.58%), SGD (84.55%), and XGBoost (93.76%) had the best mean accuracy of all the classifiers that were assessed. These results suggest that ensemble-based methods, specifically XGBoost, are superior for voice-based feature-based Parkinson's disease prediction [20].

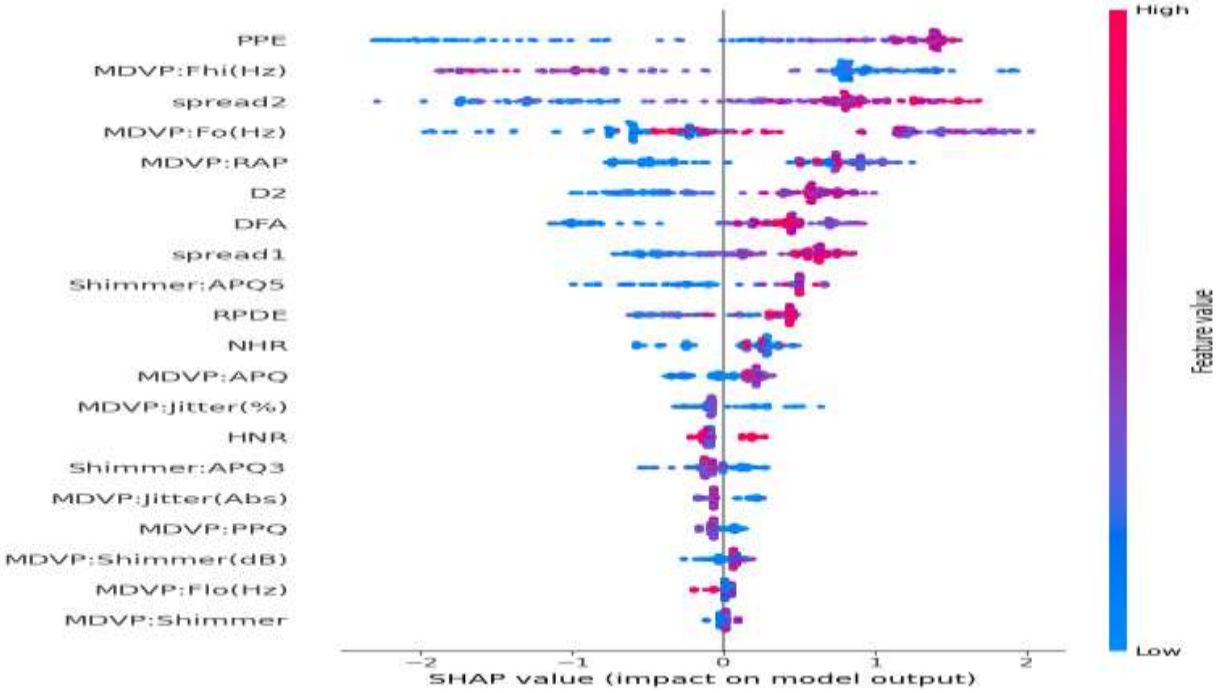

Figure 18. SHAP Feature Importance

Blue dots show low feature values in the SHAP summary graphic, whereas pink dots show high feature values. Each dot's location on the horizontal axis indicates how it affects the model's forecast. The prediction is pushed toward Parkinson's disease by positive SHAP values (right side of the plot) and toward the healthy class by negative SHAP values (left side of the plot). Additionally, the elements toward the top of the plot have the most influence, suggesting that they play a bigger role in the model's decision-making. The findings illustrate the significance of voice-based biomarkers in the diagnosis of Parkinson's disease by demonstrating that features like PPE, MDVP(Hz), spread2, and MDVP(Hz) have the biggest influence on the disease's prediction. Because it enhances machine learning models' interpretability and transparency, SHAP analysis is crucial. Despite having excellent prediction accuracy, models like XGBoost are frequently referred to as "black-box" models due to the complexity of their decision-making process. By indicating whether a trait raises or lowers the chance of Parkinson's disease, SHAP demonstrates how each feature contributes to the final prognosis. This increases clinicians' confidence in the model, aids in identifying the most significant voice biomarkers linked to the illness, and promotes the creation of more dependable and comprehensible diagnostic tools. Consequently, SHAP analysis raises the clinical relevance and uniqueness of the suggested method while also improving model interpretability.

Model	Accuracy(%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	87.58	87.14	87.58	85.29
Random Forest	91.18	91.87	91.18	90.54
SGD	84.55	83.27	84.55	83.19
XGBoost	93.76	94.04	93.76	93.39

Table 8. Performance comparison of machine learning models for Parkinson's disease prediction using 10-fold cross-validation.

Table 8 shows that XGBoost outperformed all other models, achieving the highest accuracy (93.76%), precision (94.04%), recall (93.76%), and F1-score (93.39%). Random Forest also produced strong results, whereas SVM and SGD exhibited comparatively lower performance. These findings indicate that XGBoost is the most effective model for Parkinson's disease prediction using the selected voice features

6.9 Comparison of Feature Importance Before and After 10-Fold Cross-Validation

Table 9 presents the comparison of important features identified before and after 10-fold cross-validation. Although the importance values are derived from different scoring scales, several key features remain consistent, demonstrating the robustness of the proposed Parkinson's disease detection model.

Feature	Before 10-Fold	After 10-Fold	Observation
MDVP:Fo(Hz)	0.062	38	Important in both analyses
PPE	0.150	19	Consistently influential feature
RPDE	0.031	17	Stable predictor across folds
spread1	0.182	8	Reduced importance after cross-validation
spread2	-	24	Emerged as important after 10-fold validation
HNR/NHR	0.034	8	Moderate importance

Table 9: Comparison of importance features identified before and after 10-fold cross – validation

6.10 Interpretation

The comparison indicates that MDVP: Fo(Hz), PPE, and RPDE consistently appeared among the top-ranked features before and after 10-fold cross-validation, where the left figure shows before and the right figure shows after 10 fold cross validation features. The slight variations in ranking reflect the effect of different training subsets, while the overall consistency confirms the stability and reliability of the selected voice biomarkers.

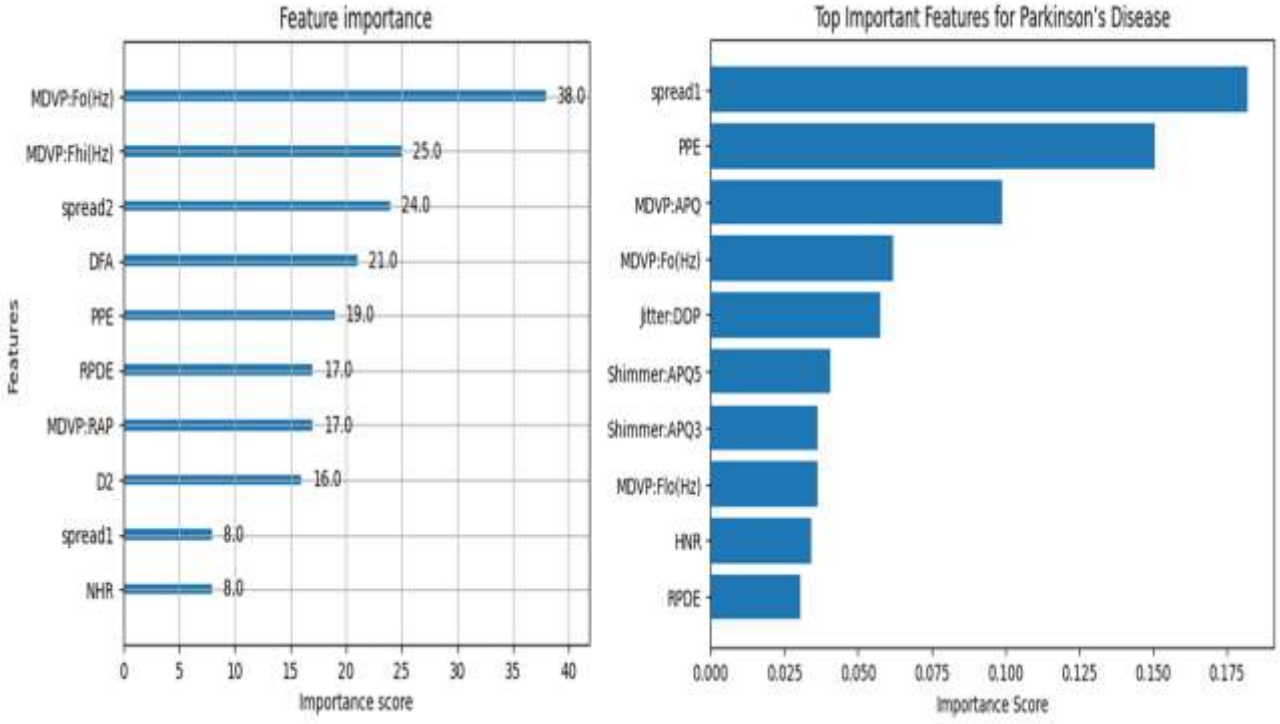

Figure 19: Feature importance comparison before and after 10-fold cross-validation.

6.11 Comparative Study :

Study	Algorithms	Validation Method	Explainability	Reported Accuracy (%)	Best Model	Remarks
Little et al. (2009)	Dysphonia Measures	Hold-out	No	91.40	Statistical Model	Early voice-based PD study

Naranjo et al. (2017)	Feature Selection + Classifier	Cross-validation	No	91.50	Variable Selection	No XAI
Gündüz (2019)	Deep Learning	Cross-validation	No	88.25	DNN	Black-box approach
Yuvaraj et al. (2020)	SVM, KNN, RF	Hold-out	No	90.64	RF	Single validation
Ghaheri et al. (2024)	Ensemble + SHAP	Cross-validation	SHAP	85.42	Ensemble	Explainable approach
Proposed (80:20 Test)	SVM, RF, XGBoost	Hold-out	SHAP + Heatmap	92.31	XGBoost/RF	Test accuracy
Proposed (10-Fold)	SVM, RF, XGBoost	10-Fold Cross-Validation	SHAP + Heatmap	93.76	XGBoost	Mean CV accuracy

Table 10: Comparison of Existing and Proposed Methods for Parkinson's Disease Prediction

The suggested framework for Parkinson's disease prediction is contrasted with previous research in Table 10. The suggested method combines hold-out validation and 10-fold cross-validation with SHAP-based explainable AI approaches, in contrast to earlier research that either depended on a single validation strategy or lacked explainability. The findings showed that XGBoost had the highest cross-validated accuracy of 93.76%, suggesting that the suggested framework offers a more dependable and comprehensible way to predict Parkinson's disease using voice biomarkers.

CONCLUSION

The main objective of this research is to create a machine learning framework for the early detection of Parkinson's disease utilizing voice-based biomarkers that is precise, dependable, and understandable. To find the best prediction model, the study evaluates the performance of SVM, Random Forest, and XGBoost classifiers using both hold-out validation and 10-fold cross-validation. Additionally, model decisions are interpreted and the most important speech qualities linked to Parkinson's disease are identified using SHAP-based explainable artificial intelligence algorithms, correlation heatmaps, and feature importance analysis. The experimental findings show that XGBoost has the best predictive performance and can be a reliable and comprehensible method for detecting Parkinson's disease.

REFERENCES:

[1] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015-1022, 2009.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," Journal of the Royal Society Interface, vol. 8, no. 59, pp. 842-855, 2011.

[3] H. Gündüz, "Deep learning-based Parkinson's disease classification using vocal feature sets," IEEE Access, vol. 7, pp. 115540-115551, 2019.

[4] N. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1925-1936, 2021.

[5] M. Vadovský and J. Paralič, "Parkinson's disease patients classification based on the speech signals," in Proc. 15th IEEE Int. Symp. Applied Machine Intelligence and Informatics, 2017, pp. 321-326.

[6] A. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," IEEE Journal of Translational Engineering in Health and Medicine, vol. 7, pp. 1-10, 2019.

[7] Q. C. Ngo, M. A. Motin, N. D. Pah, P. Drotár, P. Kempster, and D. Kumar, "Computerized analysis of speech and voice for Parkinson's disease: A systematic review," Computer Methods and Programs in Biomedicine, vol. 226, Art. no. 107133, 2022.

[8] K. Pahuja and T. N. Nagabhushan, "A comparative study of existing machine learning approaches for Parkinson's disease detection," IETE Journal of Research, vol. 67, no. 1, pp. 4-14, 2021.

[9] K. Polat and M. Nour, "Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals," Medical Hypotheses, vol. 140, Art. no. 109678, 2020.

[10] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," Computer Methods and Programs in Biomedicine, vol. 142, pp. 147-156, 2017.

[11] L. Benba, A. Jilbab, and A. Hammouch, "Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease," IRBM, vol. 38, no. 6, pp. 346-351, 2017.

[12] S. Singh and W. Xu, "Robust detection of Parkinson's disease using harvested smartphone voice data: A telemedicine approach," Telemedicine and e-Health, vol. 26, no. 3, pp. 327-334, 2020.

[13] M. A. Motin, N. D. Pah, S. Raghav, and D. K. Kumar, "Parkinson's disease detection using smartphone recorded phonemes in real world conditions," IEEE Access, vol. 10, pp. 97600-97609, 2022.
[14] R. Yuvaraj, Y. Hagiwara, U. Raghavendra, N. Arunkumar, and U. R. Acharya, "A machine learning framework for Parkinson's disease diagnosis using voice features," Neural Computing and Applications, vol. 32, pp. 10927-10933, 2020.
[15] A. Iyer et al., "A machine learning method to process voice samples for identification of Parkinson's disease," Scientific Reports, vol. 13, Art. no. 20615, 2023.
[16] G. Hinton and T. Hastie, "Statistical learning and voice-based disease prediction approaches," Journal of Machine Learning Research, vol. 22, pp. 1-28, 2021.
[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.
[18] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[19] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[20] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[21] P. Ghaheri, H. Nasiri, A. Shateri, and A. Homafar, "Diagnosis of Parkinson's disease based on voice signals using SHAP and hard voting ensemble method," Computer Methods in Biomechanics and Biomedical Engineering, vol. 27, no. 15, pp. 1858-1874, 2024.
[22] H. Gündüz, "An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification," Biomedical Signal Processing and Control, vol. 66, Art. no. 102452, 2021.
[23] R. Prashanth and S. D. Roy, "Early detection of Parkinson's disease through patient questionnaire and predictive modelling," International Journal of Medical Informatics, vol. 119, pp. 75-87, 2018.
[24] M. Mei, C. Desrosiers, and J. Frasnelli, "Machine learning for the diagnosis of Parkinson's disease: A review of literature," Frontiers in Aging Neuroscience, vol. 13, Art. no. 633752, 2021.
[25] R. Gupta, S. Kumari, A. Senapati, R. K. Ambasta, and P. Kumar, "New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease," Ageing Research Reviews, vol. 89, Art. no. 102013, 2023.
[26] S. Scimeca et al., "Robust and language-independent acoustic features in Parkinson's disease," Frontiers in Neurology, vol. 14, Art. no. 1198058, 2023.
[27] J. R. Orozco-Arroyave et al., "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in Proc. LREC, 2014, pp. 342-347.
[28] Q. Wang et al., "Early detection of Parkinson's disease from multiple signal speech: Based on Mandarin language dataset," Frontiers in Aging Neuroscience, vol. 14, Art. no. 1036588, 2022.
[29] W. Rahman et al., "Detecting Parkinson's disease from an online speech-task," in Proc. Interspeech, 2020, pp. 4457-4461.
[30] H. Rabie and M. A. Akhloufi, "A review of machine learning and deep learning for Parkinson's disease detection," Discover Artificial Intelligence, vol. 5, Art. no. 24, 2025.
[31] M. Shen, P. Mortezaagha, and A. Rahgozar, "Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis," Scientific Reports, vol. 15, Art. no. 11687, 2025, doi: 10.1038/s41598-025-96575-6.
[32] H. S. Malekroodi, M. H. Mirzaei, and M. Ghasemi, "Voice-Based Detection of Parkinson's Disease Using Machine Learning and Deep Learning: A Systematic Review," Diagnostics, vol. 15, no. 1, Art. no. 34, 2025.
[33] N. M. Nayan, M. Hossain, and S. Islam, "An interpretable and balanced machine learning framework for Parkinson's disease prediction integrating feature selection and explainable AI techniques," PLOS ONE, vol. 20, no. 10, e0333418, 2025.
[34] H. Rabie and M. Akhloufi, "A review of machine learning and deep learning techniques for Parkinson's disease detection and progression monitoring," Discover Artificial Intelligence, vol. 5, Art. no. 41, 2025.
[35] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, 1995, pp. 1137-1143.
[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.