Journal of Information Systems and Informatics
Vol. 7, No. 3, September 2025 e-ISSN: **2656-4882** p-ISSN: **2656-5935**
DOI: 10.51519/journalisi.v7i3.1172 Published By DRPM-UBD

A Comparative Analysis of Machine Learning Techniques and Explainable AI on Voice Biomarkers for Effective Parkinson's Disease Prediction

Otis Mabikwa[1], Belinda Ndlovu[2], Kudakwashe Maguraushe[3]

[1,2]National University of Science and Technology, Ascot AC 939, Bulawayo
[3]School of Computing, University of South Africa, Johannesburg
Email: [1]otismabikwa@gmail.com, [2]belinda.ndlovu@nust.ac.zw, [3]magark@unisa.ac.za

Abstract

Parkinson's disease (PD) is a neurological movement disorder that remains difficult to diagnose, although it affects millions globally. Early diagnosis can lead to more effective and improved patient outcomes. Diagnosis through traditional methods is subjective and often lacks transparency, raising concerns about reliability. In this study, the CRISP-DM framework was applied to compare eight ML algorithms, including Random Forest and Support Vector Machine (SVM). Recursive Feature Elimination (RFE) was used to preprocess, balance, refine the data and find the eight most predictive vocal features. With 195 recordings coming from the UCI Parkinson's Speech Dataset, which contains voice measurements from 31 individuals (23 with PD and 8 healthy controls), Random Forest (Entropy) had the best performance ($\boldsymbol{F_1}$ = 96.6%, ROC AUC = 0.98). Explainable AI tools (SHAP and LIME) were integrated, allowing both global and instance-level understanding of model predictions thereby identifying measures of pitch variability (MDVP: RAP, spread1, PPE) as key predictors of PD. This research contributes to the practical deployment of reliable, transparent PD prediction tools in real-world medical settings, supporting early diagnosis and improved patient care. This raises the issue of the urgent need to detect PD early among Africa's aging populations to help protect the cultural heritage contained in the voices of the elders. this research contributes to the practical deployment of reliable, transparent PD prediction tools in real-world medical settings, supporting early diagnosis and improved patient care. Future work should embark on validating these findings over much more varied cohorts, integrating additional data modalities (e.g., gait, imaging), and enhancing model robustness. Real-time speech analysis-based tools, in the end, will allow remote screening, early intervention, and tailored care.

Keywords: Parkinson's Disease (PD), Machine Learning (ML), Artificial Intelligence (AI), Explainable Artificial Intelligence (XAI), Prediction

1. INTRODUCTION

Parkinson's disease (PD) affects mostly people of the older generation [1], and in many African societies, older people go beyond family. The elderly are living libraries of community history and tradition, along with moral guidance [2]. As soon as the traditional elderly slowly fade out of action or die too soon because of complications related to PD, the entire community has lost its main custodians of language, proverbs, stories, and ritual knowledge kept orally and passed down for generations [3]. This culture isn't just academic; it erodes social ties as rites of passage, dispute reconciliation customs, and even forms of traditional healing often draw upon the wisdom of senior members [3]. Economically, more caregiving duties fall on families; they spend time on and resources away from education or income-generating schedules [4]. Emotionally, the absence of grandparents and great uncles in the home disrupts the cross-generational linkages and transmission of values like respect, reciprocity, and communal responsibility [5]. In summary, Parkinsonism in the elderly constitutes dimensions beyond the mere health of the individual; it threatens continuity into cultural identity and social stability in African communities. In numerous African societies, elders are more than relatives; they are custodians of oral history, moral instruction, and fostering community cohesion [3]. The accelerated PD related attrition of these individuals thus severs the intergenerational flows of cultural identity, norms of conduct, and traditional healing practices.

Parkinson's disease, a progressive movement disorder characterized by tremors, stiffness, and impairment with balance and coordination [6], [7] is a global healthcare burden of ranking next only to Alzheimer's among neurodegenerative disorders, normally affecting persons aged 50 and above, with a projected increase in cases [8]. It has caused a significant strain on healthcare systems [9]. Apart from the deterioration of quality of life for the patients due to disabling motor symptoms (tremor, rigidity, slowness of movement, postural instability and difficulties with balance) and non-motor symptoms (mental health problems, sleep disturbances, digestive issues, and sensory changes) this disease has high economic and social costs, particularly in low- and middle-income countries with limited resources [10].

The World Health Organization (WHO) has recognized PD as an increasing health burden of rising importance, requiring early detection and management approaches [11]. However, with traditional methods, early detection continues to be a major challenge due to the subtly of PD symptoms and the subjectivity of clinical diagnosis [12].

The diagnosis of PD primarily relies on clinical assessment by movement disorder specialists [13], involving subjective evaluations of motor symptoms, disease history, and clinical examinations which are hard to differentiate from other diseases such as Multiple System Atrophy (MSA), Pro-Gressive Supranuclear Palsy (PSP) and Huntington's Disease [14], [15], [16]. This subjectivity can lead to wrong diagnoses and treatments [9], [17].

The lack of definitive biomarkers in PD diagnosis, unlike some other diseases (e.g., Alzheimer's disease), makes PD diagnosis harder [9], [18], [13]. The diagnosis relies on a constellation of symptoms and observations rather than a definitive test [19], [11]. Standard clinical assessments are expensive, time-consuming and may not capture subtle changes in speech or movement patterns indicative of early-stage PD [20]. Machine learning has emerged as a promising non-invasive technology for predicting the risk and progression of PD [21].

Machine Learning and AI offer promising solutions by leveraging their ability to analyze complex datasets and learn intricate patterns [8]. Machine Learning algorithms can build predictive models from PD patient datasets to classify between PD and healthy individuals [13]. These algorithms can analyze subtle changes and complex patterns in various data modalities (such as speech, gait, handwriting, or neuroimaging) that may elude human perception [12], [8], [19]. This leads to high diagnostic accuracies, often exceeding traditional methods for instance in a study by [9], Support Vector Machine (SVM) achieved accuracies ranging from 65.2% to 99.99%.

Non-Invasive Detection is made possible, ML models can process and intergrate complex, non-invasive data modalities like speech/voice recordings, gait patterns, handwriting patterns, and even health screening data [12], [11], [22]. These can reveal subtle changes indicative of early-stage PD that are often overlooked in traditional assessments [13], [11], [19]. According to [12], [11] and [13] speech impairment is a common and early non-motor symptom of PD, making speech analysis a valuable tool for early detection and monitoring.

Explainable AI methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) improve interpretability by providing insights into "why" a model makes a particular prediction [14], [23]. This allows healthcare professionals to understand the decision-making process, fostering trust and facilitating clinical adoption as many advanced ML models, particularly deep learning, can be "black boxes", lacking transparency [14]. Explainable AI can highlight the specific features whether in speech or brain regions that contribute most to the model's decision, making the diagnosis evidence-based and interpretable for clinicians [1], [14].

Machine Learning provides the possibility of automated diagnosis using various PD biomarkers such as voice abnormalities, posture analysis in gait, brain imaging like Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) [23], [24], [25]. A significant advantage of ML is its ability to identify relevant features that are not traditionally used in clinical diagnosis, paving the way for new biomarker discovery in PD prediction [19], [26], [27] supporting clinical decision making while enabling early diagnosis [6], [7], [28].

However, only a few preliminary studies in small and clinically heterogenous cases have been reported [29], [30]. This highlights the lack of large and clinically well-characterized examinations of voice impairment in the prediction of PD; therefore, a lack of comparative analysis of algorithms. While the use of ML in healthcare continues to grow, the demand for transparency and interpretability in ML models is also growing [23]. Explainable AI (XAI) techniques, such as SHAP and LIME, have shown promise in explaining the decision-making processes of complex models, thereby producing confidence among both clinicians and patients [31], [7].

This study aims to establish a detailed comparative analysis of various ML techniques, such as Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) Gini, Random Forest (RF) Entropy, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GaussianNB), Bernoulli Naive Bayes (Bernoulli NB), Voting Ensemble, alongside XAI approaches SHAP and LIME for the efficient predicting PD on voice biomarkers. These eight algorithms were used to guarantee a balanced approach in measuring predictive power, interpretability, and methodological variance. Logistic Regression and Linear SVM, among others, were linear candidates.

Then we had tree ensemble types, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, followed by instance-based and probabilistic paradigms provided by K-Nearest Neighbors (KNN) and Naïve Bayes. This variety helped the study to capture the essence of different paradigms from simply linear decision boundaries to sharply non-linear and hierarchical relationships within the voice and demographic data. Using an array of algorithms served two purposes: first, to benchmark performances to ensure that success in prediction was an advantage that could not arise from just any single modeling approach; and second, to discern which model family handled best the complex, possibly correlated, and high-dimensional nature that Parkinson's-themed features potentially exhibit. This gave us an advantage in robustness, trustworthiness, and yielded increased insight into how differently model architectures would react to one and the same medical dataset.

To ensure that model predictions were transparent and clinically interpretable, two complementary XAI methods, SHAP and LIME, were employed. Using SHAP, one could explain feature importance globally and locally, with a solid mathematical foundation from the Shapley value concept, thereby giving a fair attribution to how each feature might have contributed to the output of the models. This allowed researchers to identify the voice or demographic features that always influenced predictions across the dataset. LIME works in conjunction with SHAP by providing a local explanation for individual predictions through simple and interpretable local surrogate models, which allow physicians to understand the reasoning behind a prediction for a particular patient. Thus, the combined application of SHAP and LIME makes the system more trustworthy by offering high transparency at the population level and clarity at the individual patient level-a requirement for placing AI models in real-world medical decision-making.

By evaluating the performance and interpretability of each model, this paper seeks to fill the existing voids in the literature. The results will help researchers and medical practitioners in choosing the best predictive models, leading to earlier interventions and better patient care.

2. METHODS

This study used the Cross-Industry Standard Process for Data Mining (CRISP-DM) [32] as a systematic framework for developing, evaluating, and deploying the prototype. Its six phases- Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment form a stringent and repeatable process.

The objective was to develop a comparison of ML models, create a reliable classification model capable of distinguishing between subjects with PD and healthy subjects based on vocal biomarkers, and to make explanations of each prediction transparent. This would foster rapid, non-invasive screening done in clinical sites to augment neurologists. While in the data understanding and cleaning phase, the dataset used was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado. The dataset was obtained from Kaggle, a widely recognized platform within the data science community.

In data understanding and cleaning, the dataset went through screening to find missing values, mismatched formats, and duplicate records- none of them were found, and no nulls were recorded. While imputation & treatment of outliers was not applicable as all features were numerical and preserved extreme values for model sensitivity. In the Data Preparation stage, Recursive Feature Elimination (RFE) with a random forest base estimator reduced 22 original features into the most predictive eight (MDVP: Fo(Hz), MDVP: Fhi(Hz), MDVP: RAP, Shimmer: APQ5, MDVP: APQ, spread1, spread2, PPE). Recursive Feature Elimination was used as it selects the most important original features by recursively removing the least useful features based on a model that allows tools like SHAP and LIME to directly explain how those features influenced predictions.

For class balancing, the Synthetic Minority Over-sampling Technique (SMOTE) was used to match numbers for each class between Parkinson's vs healthy. While for Feature scaling, the MinMaxScaler transformed all features into the [–1, 1] range to ensure stable convergence, especially for distance-based algorithms. In the Data Partitioning phase, the Train/Test split, stratified into an 80/20 split, preserved class proportions, which produce representative training and evaluation sets. While 5-Fold Cross Validation was employed during hyperparameter tuning on the training set.

2.1. ML Algorithms

Eight machine learning algorithms were chosen for this research so as to cover a wide spectrum of operative methods, each serving distinct roles in handling complex heterogeneous patterns for voice and demographic data in Parkinson's disease prediction. Logistic Regression was included as a baseline linear model because it is interpretable, efficient, and can produce probability estimates, which find utility in medical settings [33]. Support Vector Machines (SVM) with linear kernels were chosen because finding the optimal separating hyperplanes in high-dimensional feature space is vital, given that voice features may possess only subtle discriminative signals [9]. Random Forest was incorporated as a strong ensemble method able to deal with correlated features and non-linear relationships while inferring feature importance inherently [9].

Gradient Boosting, XGBoost, and LightGBM were validated as advanced boosting algorithms championed for their unprecedented predictive power, ability to detect complex feature interactions, and good regularization that circumvents overfitting [15], [9], [34]. K-Nearest Neighbors was included as a non-parametric technique [21], [33], [9]. Lastly, Naïve Bayes, with its generic probabilistic reasoning framework and computation efficiency, can be used in an exploratory evaluation despite its simplistic assumption of independence among attributes [13], [6]. Such a methodological range allowed for complete benchmarking in the realms of linear, probabilistic, distance-based, and ensemble paradigms, thereby giving even more strength to the evaluation and allowing insight into which model families are best suited to Parkinson's disease voice-based classification tasks.

2.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps understand the structure, distribution, and characteristics of data before ML techniques are used [12]. In the case of detecting PD, EDA helps bring out potential unique voice patterns that may suggest neurodegenerative symptoms [12]. This primary step is fundamentally concerned with finding correlations, detecting outliers, ensuring the quality of data, and choosing significant features for model training.

In this study, EDA techniques were used in the analysis of the PD Voice Dataset, available publicly from the UCI Machine Learning Repository, and originally prepared by Max Little and co-researchers. The dataset enables the study of voice features that may be affected somehow by PD and thus may serve as non-invasive biomarkers.

2.3. Parkinson's Disease Voice Dataset

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders. There are a total of 195 recordings in the dataset from 31 subjects, among whom 23 were diagnosed with PD and 8 were healthy controls. Multiple recordings (about six per subject) ensure a strong sample for each subject while capturing intra-subject variability. Each observation in the dataset corresponds to one voice recording, while subjects are kept anonymous by assigning unique identifiers in the name column. The target variable, status, marks the health condition: "0" for healthy and "1" for PD-affected. The data is formatted as ASCII CSV. Each row of the CSV file represents a distinct voice recording instance. Each patient has approximately six recordings, with the patient's name listed in the first column. The dataset includes 22 voice-related features, all of which are continuous and derived from signal processing techniques designed to quantify various aspects of vocal function.

2.4. Dataset Limitations

While the dataset provides a rich set of 22 voice-related features extracted from established signal processing methods, a few limitations may undermine the generalizability and robustness of results. First, the database carries only 195 voice recordings of 31 individuals (23 diagnosed with Parkinson's and eight healthy controls). Although multiple recordings per subject-Normally, six-in some ways, allow capturing intra-subject variability, the small number of unique participants remains limiting from a statistical standpoint, affecting the robustness of results and increasing the chances of overfitting. For machine learning in particular, this is important, as models might learn speaker-specific cues instead of generalizable disease markers and therefore will not be able to perform well on entirely new patients.

Second, there is also an imbalance in the number of classes within the dataset since we have many more recordings from patients with Parkinson's than healthy controls. This could also bias a model toward predicting the majority-class label and deceptively inflate accuracy measures at the expense of poor sensitivity or specificity for the minority-class group. Besides, the dataset does not contain demographic markers such as age range, gender, ethnicity, and linguistic background, which limit the assessment of whether the models perform fairly for different populations. This lack of diversity in the training data implies that the trained models may not be able to generalize well to clinical environments, especially socio-linguistically or culturally distinct ones.

Above all, as all the recordings were carried upon certain protocols, in association with the National Centre for Voice and Speech, the setting might not be representative of the varied real-world clinical or community-oriented settings. The actual model performances would vary depending on background noises, differences in recording equipment, and types of speech tasks, when deployed beyond the controlled environment. All these stress the importance of validating the models on larger and more heterogeneous and ecologically varied datasets before binding into the clinical context.

2.5. Data Preprocessing

1) Data Cleaning & Type Conversion

Text Removal: The “name” column (an ASCII identifier) and any non-numeric fields were dropped, leaving only the 22 numerical voice measures plus the binary “status” label. Missing Values: A quick check confirmed there were no nulls in the published dataset, so no imputation was required.

2) Exploratory Outlier Detection

Distribution & Boxplots: We plotted histograms (through Seaborn’s distplot) and boxplots for each feature to visualize skewness and identify extreme values. While isolated outliers were noted, we elected *not* to clip or remove them both because PD can manifest as extreme deviations in voice measures, and because tree-based models tend to be robust to a few outliers.

3) Class Imbalance Correction

RandomOverSampler: Since only 23 of the 31 subjects had PD (and each contributed multiple recordings), the positive class was undersampled. We applied *imblearn’s RandomOverSampler* to synthetically rebalance the two classes, bringing the PD and healthy-control counts into parity so that downstream algorithms would not be biased toward the majority class.

4) Normalization

Min–Max Scaling to [–1, +1]: All features were then subjected to scaling with sklearn's MinMaxScaler fitted for the range [–1, +1]. Such a choice keeps the sign of deviation intact (say, pitch vs. jitter) and ensures all features share the same dynamic range—very important to distance-based learners (KNN) and gradient-based optimizers.

5) Feature Selection as Implicit Noise Reduction

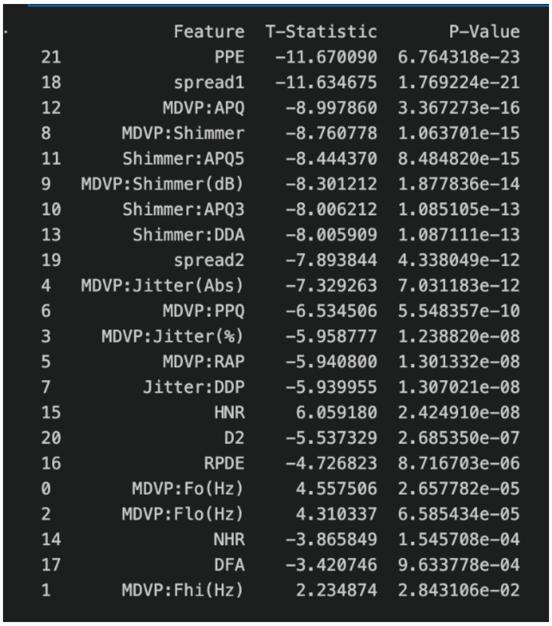
Recursive Feature Elimination: Finally, to reduce dimensionality (and implicitly remove any noisy or redundant features), we ran RFE atop a *RandomForestClassifier,* which independently selected the top eight most informative voice measures which were: MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:RAP, Shimmer:APQ5, MDVP:APQ, spread1, spread2, PPE. This step both denoises and focuses the models on the features that contribute most to PD discrimination.

T-Tests were performed to analyse the statistical differences between the controls and PD groups on voice features. The T Statistics output was used to show how many standard deviations a coefficient estimated for a particular feature deviate from zero, thereby determining whether the observed effect of a feature is statistically significant or not. A greater absolute value of T Statistic corresponds to stronger evidence against the null hypothesis (i.e., the feature has no effect). Tests were focused generally, on features influenced by PD. We also carried out the analysis of p-values. The P-value represents the probability of observing the observed data (or data that is most extreme) under the assumption that the null hypothesis is true. Here, the null hypothesis is that the feature does not have predictive power for PD. A smaller p-value thereby provides evidence stronger against the null hypothesis that the feature is statistically significant.

The t-test formula, as delineated in equation (1), contrasts the means of these features between the control group ((x_1) ̅) and the PD group ((x_2) ̅). The pooled variance (〖S^2〗 _pooled) and the sample sizes (n1 for control and n2 for PD) are taken into account to yield a reliable assessment of the difference. Equation 1 is t-test formula.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S^2_{pooled}(\frac{1}{n_1}+\frac{1}{n_2})}} \tag{1}$$

The analysis revealed features with substantial disparities between the groups. Using scipy.stats library in Python, we performed t-tests and found the P-values on each feature, Figure 1 summarizes the features.



```
              Feature  T-Statistic       P-Value
21                PPE   -11.670090  6.764318e-23
18            spread1   -11.634675  1.769224e-21
12           MDVP:APQ    -8.997860  3.367273e-16
8        MDVP:Shimmer    -8.760778  1.063701e-15
11       Shimmer:APQ5    -8.444370  8.484820e-15
9    MDVP:Shimmer(dB)    -8.301212  1.877836e-14
10       Shimmer:APQ3    -8.006212  1.085105e-13
13        Shimmer:DDA    -8.005909  1.087111e-13
19            spread2    -7.893844  4.338049e-12
4    MDVP:Jitter(Abs)    -7.329263  7.031183e-12
6            MDVP:PPQ    -6.534506  5.548357e-10
3      MDVP:Jitter(%)    -5.958777  1.238820e-08
5            MDVP:RAP    -5.940800  1.301332e-08
7          Jitter:DDP    -5.939955  1.307021e-08
15                HNR     6.059180  2.424910e-08
20                 D2    -5.537329  2.685350e-07
16               RPDE    -4.726823  8.716703e-06
0         MDVP:Fo(Hz)     4.557506  2.657782e-05
2        MDVP:Flo(Hz)     4.310337  6.585434e-05
14                NHR    -3.865849  1.545708e-04
17                DFA    -3.420746  9.633778e-04
1        MDVP:Fhi(Hz)     2.234874  2.843106e-02
```

Figure 1. Statistically Significant Features for Parkinson's Disease Detection

We computed Pearson's correlation coefficient and Spearman's rank correlation coefficient to study the relation of features of voice with the status of PD. The relationship between the voice characteristics and this status was assessed in terms of direction and strength, where the linear association was well described by the Pearson correlation and the monotonic (likely non-linear) relationships by Spearman rank correlation. The paired use of these methods was also prompted by the assumption that voice features' distribution may depart from normality. The mathematical definitions for Pearson's and Spearman's correlation coefficients can be seen in Equations (2) and (3), respectively. We confined our analysis to features with absolute correlation coefficients above the a priori defined threshold of 0.3 taken to denote stronger associations with PD status. Such features are likely to stand up as very informative in discriminating between individuals suffering from PD and healthy controls.

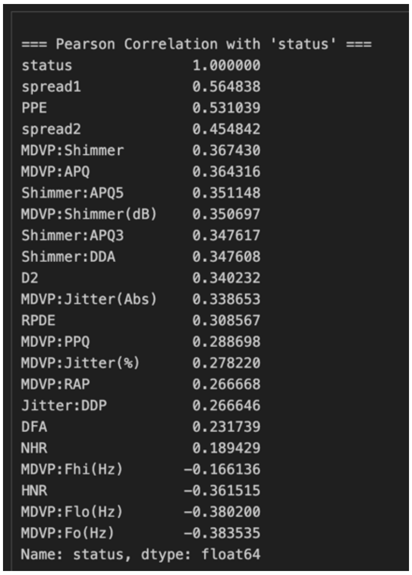
The results in Figure 2 summarize the correlation of selected voice features with PD status. Vocal variability features, such as those describing jitter and shimmer, tended to correlate positively with the disease, suggesting that the increased pitch and amplitude perturbation could be an indicator of the disease. In contrast, MDVP: Fo (Hz) contrasted with MDVP: Flo (Hz), which showed a negative correlation with the measurements taken. Equation 2 and 3 are Correlation used in this study.

$$\frac{\sum(x_{i-}\overline{x_\iota})((y_{i-}\overline{y_\iota})}{n\sigma_x\sigma_y} \tag{2}$$

$$1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

Where:
x_i and y_i are the individual sample points.
$\bar{x}$ and $\bar{y}$ are the means of the sample points.
d_i is the difference between the ranks of corresponding elements
n = number of data points
σ_x = standard deviation of feature_data
σ_y = standard deviation of target_subset

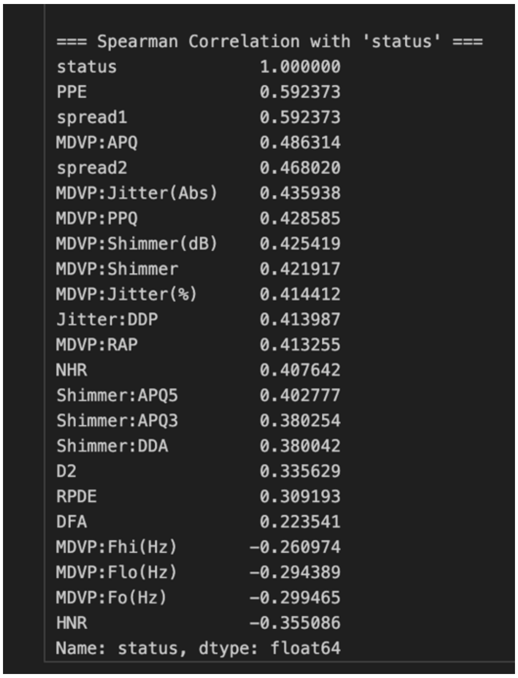
Figure 2 shows the results of how Pearson correlation coefficients were plotted between features in the dataset and the target variable status, which denotes a healthy (0) status or PD (1). Pearson correlation estimates the strength and direction of the linear relationship between continuous variables. Some features exhibit moderately strong positive correlations with PD status: spread1 (0.5648), PPE (0.5310), and spread2 (0.4548), indicating that these nonlinear measurements of fundamental frequency variation tend to increase among those with PD. In contrast, MDVP: Fo(Hz) (−0.3835), MDVP: Flo(Hz) (−0.3802), and HNR (−0.3615) have negative correlation coefficients, indicating that individuals with PD tend to have low fundamental frequency values and low harmonic-to-noise ratios, which reflect voice clarity. The significance of these observations is to highlight which parameters greatly influence vocal features related to the presence or otherwise of the disease.



```
=== Pearson Correlation with 'status' ===
status               1.000000
spread1              0.564838
PPE                  0.531039
spread2              0.454842
MDVP:Shimmer         0.367430
MDVP:APQ             0.364316
Shimmer:APQ5         0.351148
MDVP:Shimmer(dB)     0.350697
Shimmer:APQ3         0.347617
Shimmer:DDA          0.347608
D2                   0.340232
MDVP:Jitter(Abs)     0.338653
RPDE                 0.308567
MDVP:PPQ             0.288698
MDVP:Jitter(%)       0.278220
MDVP:RAP             0.266668
Jitter:DDP           0.266646
DFA                  0.231739
NHR                  0.189429
MDVP:Fhi(Hz)        -0.166136
HNR                 -0.361515
MDVP:Flo(Hz)        -0.380200
MDVP:Fo(Hz)         -0.383535
Name: status, dtype: float64
```

Figure 2. Pearson's Correlation to PD Status

Figure 3 portrays the Spearman correlation coefficients between the said features and the PD status. An important motion stands apart-and that ,Spearman's correlation captures monotonic relationships which need not be linear but must be consistently increasing or decreasing (such as in this case). This feature makes it particularly useful when monitoring associations with non-normally distributed or ordinal data. Spearman reconfirms the importance of PPE and spread1 merit, with both achieving a pronounced value of 0.5924 when computed through Spearman.



```
=== Spearman Correlation with 'status' ===
status              1.000000
PPE                 0.592373
spread1             0.592373
MDVP:APQ            0.486314
spread2             0.468020
MDVP:Jitter(Abs)    0.435938
MDVP:PPQ            0.428585
MDVP:Shimmer(dB)    0.425419
MDVP:Shimmer        0.421917
MDVP:Jitter(%)      0.414412
Jitter:DDP          0.413987
MDVP:RAP            0.413255
NHR                 0.407642
Shimmer:APQ5        0.402777
Shimmer:APQ3        0.380254
Shimmer:DDA         0.380042
D2                  0.335629
RPDE                0.309193
DFA                 0.223541
MDVP:Fhi(Hz)       -0.260974
MDVP:Flo(Hz)       -0.294389
MDVP:Fo(Hz)        -0.299465
HNR                -0.355086
Name: status, dtype: float64
```

Figure 3. Spearman Correlation to PD Status

2.6. Modeling and Its Metrics Selected

We evaluated **eight machine learning algorithms**, each selected for its unique and complementary strengths:

1) Logistic Regression – a simple linear baseline model, valued for its interpretability through coefficient analysis.
2) Decision Tree – provides human-readable, rule-based splits, making the decision process transparent.
3) Random Forest (Gini & Entropy) – an ensemble method that reduces variance, improves generalization, and offers built-in feature importance.
4) Support Vector Machine (SVM) – effective for high-dimensional datasets and capable of handling complex decision boundaries.

5) K-Nearest Neighbors (KNN) – a straightforward, non-parametric, instance-based method that adapts well to varied data distributions.
6) Naïve Bayes (Gaussian & Bernoulli) – fast, probabilistic classifiers that assume feature independence, well-suited for text and categorical data.
7) Voting Ensemble – combines multiple classifiers to enhance robustness and stability of predictions.

Model performance was assessed using multiple metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. These measures allowed us to capture not only overall correctness but also trade-offs between false positives, false negatives, and overall classification balance.

2.7. Explainability and Tool Analysis

In this study, the two most popular XAI methods were employed: SHAP and LIME, which try to explain the reasons behind a particular decision by a ML model. SHAP is an instance where it draws from the theory of Shapley values in cooperative game theory, in which each "player" contributes to a team outcome [11]. Here, each feature in a dataset assumes the role of a "player," and the "game" outcome equates to the model output. Thus, by calculating the Shapley values, SHAP attributes an equitable score of contribution to each feature to depict how much each feature either forces the prediction upwards or downwards. LIME, on the contrary, tries to fit a simple local model around a prediction of interest to understand which features are present and more influential in that neighborhood [35]. These two tools, therefore, allow practitioners to gain an understanding and trust of the model, with SHAP giving a global importance of features and LIME explaining the model on an instance level. Cooperative game theory provides the means for calculating the model's Shapley value, which can be determined in Equation 4.

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \quad (4)$$

Where
ϕ_i: Shapley value for feature i.
S: Any subset of F \{i}, where F is the set of all features.
|S|: Cardinality (number of elements) of subset S.
|F |: Total number of features.
fSU{i}(xsU{i}): The output of the model when feature i is added to subset S.
fs(xs): The output of the model with subset S of features.

SHAP provides visualization for how each variable can affect the predictions. For every feature, a contribution score is calculated that reflects how that feature influences the output of the model, and these contribution scores are weighted and summed to show an overall effect of which features matter the most. By decomposing the predictions into separate contributions for every feature, SHAP provides a comprehensive view of how the model is using the information; this makes it simple to identify inputs that yield correct results. For clinicians, this means that they can easily recognize the key voice measurements that the model relies on and can investigate how different features may cooperatively affect the final decision. By strictly following CRISP-DM and combining state-of-the-art Explainable AI methods, this methodology delivers both high predictive performance and transparent decision support, facilitating clinical trust and adoption.

3. RESULTS AND DISCUSSION

3.1. Experiment Performance

The correlation heat-map (Figure 4) reveals how the various acoustic markers extracted from sustained vowel recordings move together, and why they matter for spotting PD. Bright red blocks along the diagonal and clustering around the "jitter" and "shimmer" groups show that different measures of frequency perturbation (e.g., MDVP: Jitter(%), RAP, PPQ) and amplitude variation (e.g., Shimmer: APQ5, Shimmer: DDA) are almost interchangeable: when one increases, so do the others, because they all quantify subtle instability in vocal fold vibration. In contrast, the strong blue band linking those jitter/shimmer features with HNR reflects that as irregularity rises, the harmonic purity of the voice falls—a classic hallmark of dysphonia in Parkinson's. Moderately positive correlations between features like PPE or spread1 and the disease "status" variable indicate that more irregular pitch patterns and wider frequency spread tend to coincide with a Parkinson's diagnosis.

These patterns emerge because neurodegenerative changes disrupt the fine motor control of vocal muscles, leading to tremor-like fluctuations (captured by jitter/shimmer) and a noisier, less harmonic speech signal (captured by HNR and entropy measures). By visualizing these relationships, we can identify redundant features for removal and focus on those that uniquely capture the vocal impairments most predictive of Parkinson's hence our choice of the eight RFE-selected features (e.g., MDVP: Fo(Hz), MDVP: RAP, Shimmer: APQ5, spread1, spread2, PPE). Each of these contributes non-redundant information: for instance, MDVP: Fo(Hz) and MDVP: Fhi(Hz) capture baseline pitch characteristics, RAP quantifies rapid frequency fluctuations, Shimmer: APQ5 measures amplitude irregularity, and entropy-based spread and PPE metrics capture overall voice complexity. Together, they form a concise, orthogonal feature set that maximizes predictive power while avoiding multicollinearity.

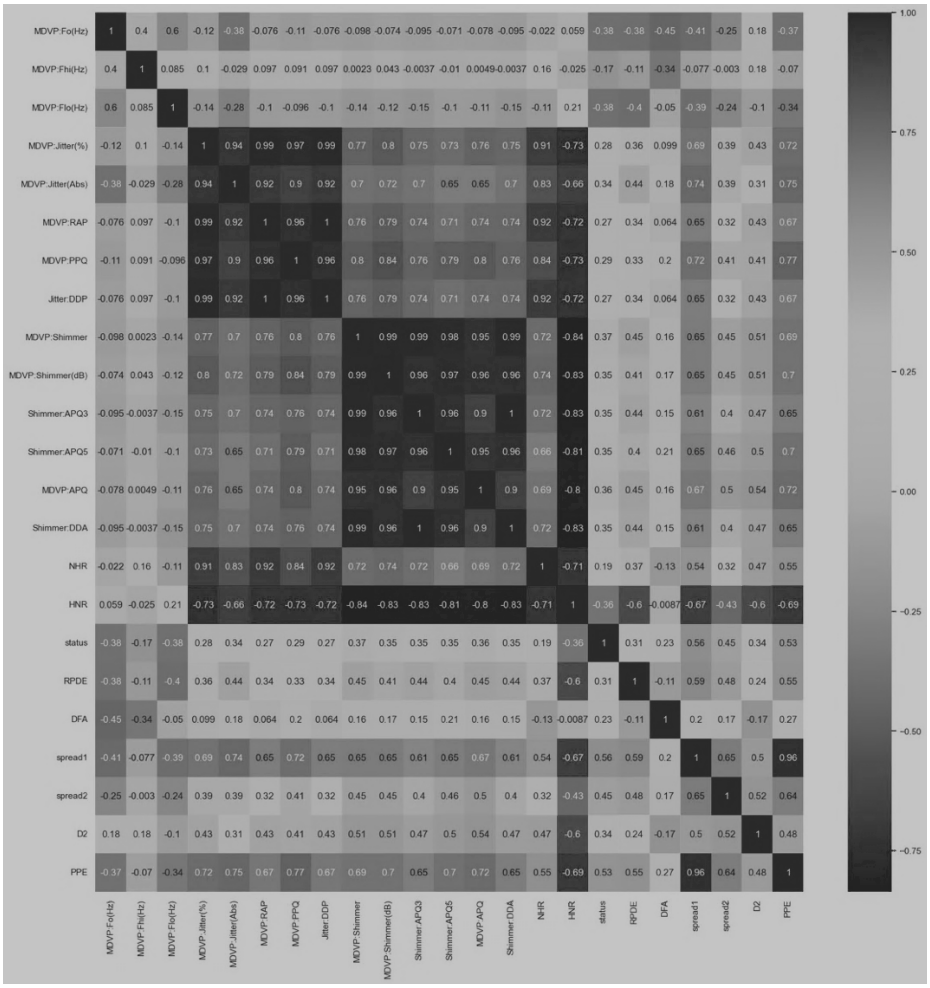


Figure 4. Correlation heatmap

The analysis of the voice dataset for predicting the disease of Parkinson generates some important facts about the association between vocal biomarkers and the existence of the disease. The heat map of feature correlations indicates that numerous voice features interrelate highly with one another, as can be observed mainly from subgroups like jitter, shimmer, and amplitude perturbation measures. These features highlight the most subtle variations of voice frequency and amplitude that increase with the presence of motor impairment from which patients with Parkinson's suffer in the control of vocal cords. For example, attributes, such as "MDVP: Shimmer", "Shimmer: APQ3", and "Shimmer: APQ5", possess extremely high positive correlation (>0.95); that means redundancy among the variables. This strong case of multicollinearity is found in any case of bio-medical voice analysis because most of the measures are physiological phenomena that have overlapping characteristics.

On the correlations between features and the dependent variable: status case that denoted whether a patient suffers from PD prudently, the bivariate dependence showed some moderately strong correlations with the following features: "spread1," "Shimmer; APQ5," and "PPE." These are features that involve non-linear measures indicative of the complexity and irregularity of the voice signal, which are relatively known to increase with neurodegenerative deterioration. Their conspicuous correlations with disease status indicate their predictive significance. Among these would be features like "HNR," which means that as the HNR lowers from 15 to 14.4, the known relationship about status directly reduces the chances of Parkinson's. This goes parallel with the literature that in the presence of Parkinson's patients, voices tend to be hoarse and are more breathy.

The strong clusterings presented in Figure 4 features justify the feature selection scheme adopted in this study. Recursive Feature Elimination is employed to eliminate redundancy and to allow for the most significant predictors. The most recently selected features- "MDVP: Fo(Hz)," "MDVP: Fhi(Hz)," "MDVP: RAP," "Shimmer: APQ5," "MDVP: APQ," "spread1," "spread2," and "PPE" were selected not only because they are statistically relevant but also because of biological significance and less redundancy as well. A sizable mixture of frequency-related, amplitude-related, and complexity terms related to voice characteristics of a patient clinically linked to the underlying pathology of Parkinson's can be gathered using these features.

Overall, it can be said that the attributes of the voice signal are very effective, non-invasive or less invasive methods of diagnosing PD. Healthy, strong correlation of features as well as their relationships with disease status supports the strength of the predictive model concerned. Added to this apparent growing research trend is the utilization of speech analysis as a viable early diagnosis tool that would allow for accessible diagnosis of PD at an early stage.

3.2. Validation

In this study, each classifier was rigorously validated using an 80/20 train-test split, ensuring that models learned from one portion of the data and were evaluated on entirely unseen voice recordings. Eight algorithms; LR, DT, RF (with both Gini and Entropy criteria), SVM, KNN, GaussianNB and BernoulliNB, and a hard-Voting Ensemble were trained on 80% of the RFE-selected, scaled features and then tested on the remaining 20%. Previously mentioned hold-out testing was used before 5-fold cross-validation during hyperparameter tuning for every model in order to maximize average ROC-AUC across folds by tweaking tree depth, regularization strength (C), and number of neighbors. Finally, accuracy, precision, recall, F1-score, and ROC AUC were estimated on the test set, giving the highest predictive accuracy with the RF (both Gini and Entropy). Confusion matrices further revealed how often each algorithm correctly distinguished Parkinson's from healthy subjects, and SHAP- and LIME-based explanations on test samples confirmed that the top-ranked features (e.g., MDVP:RAP, Shimmer:APQ5, PPE) consistently drove model decisions. This multi-metric, cross-validated approach provides strong evidence that our selected voice biomarkers and chosen classifiers generalize well to new patients. We were able to gain important insights into the relative value of features by utilizing Local Interpretable Model-Agnostic Explanations (LIME) as shown in Figure 5.

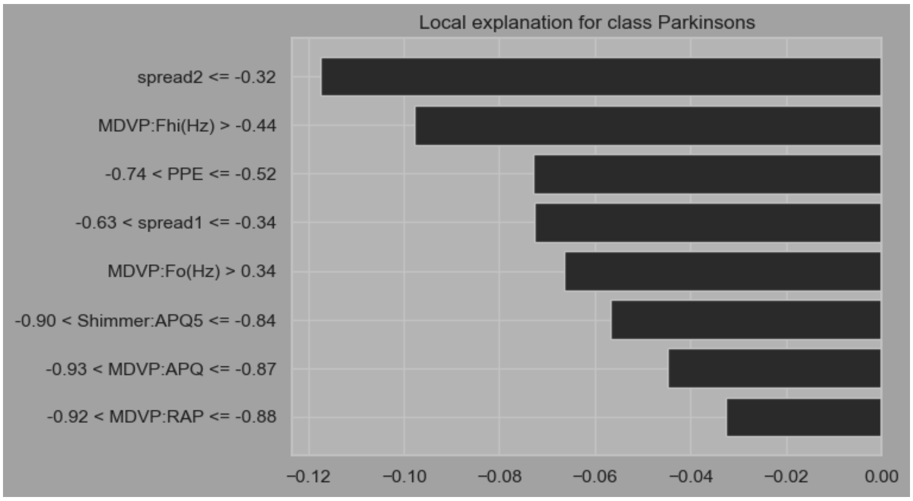


Figure 5. Feature explanation using LIME

This LIME chart (Figure 5) is a local, instance-level explanation showing which voice features drove the model's decision toward predicting Parkinson's for one particular patient. Each horizontal bar represents a simple rule (e.g. "spread2 ≤ –0.32" or "MDVP:Fhi(Hz) > –0.44"), and its length reflects how strongly that rule pushed the prediction toward the Parkinson's class. The topmost rule—spread2 ≤ –0.32—had the largest impact, meaning this patient's low "spread2" value was the single strongest indicator. Next came a relatively high fundamental-frequency maximum (MDVP:Fhi(Hz) > –0.44), a mid-range pitch-period entropy (–0.74 < PPE ≤ –0.52), and a similar interval on "spread1." Features further down the list (e.g. "MDVP:Fo(Hz) > 0.34" or "Shimmer:APQ5 between –0.90 and –0.84") made smaller but still meaningful contributions. Taken together, these eight rules explain exactly why, for this voice sample, the RF model tilted in favor of a Parkinson's diagnosis, highlighting specific acoustic thresholds that most strongly influenced the local probability.

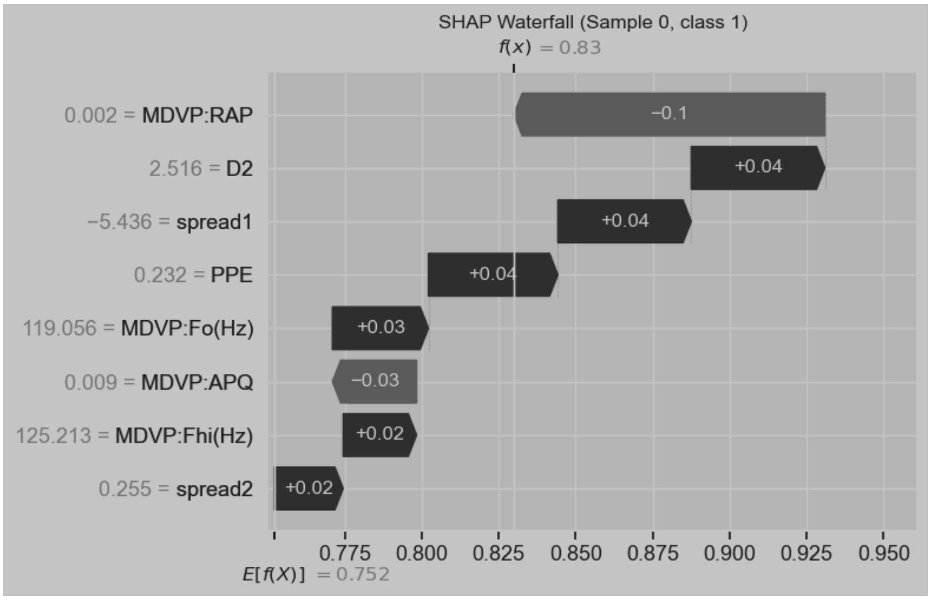


Figure 6. Feature explanation using SHAP!

The SHAP waterfall plot (Figure 6) depicts a scenario in which the contributions of each feature were analyzed for the model's prediction of a specific sample (Sample 0) that has been classified as class 1, that is, a human with PD. The baseline value, which is the averaged model output for all samples, E[f(x)], approaches approximately 0.752. This averaged outcome represents the posterior probability of PD across the dataset. For this individual, the model assigned a probability of 0.83, thus supporting that the individual is highly likely to have PD. The features in the plot served to either reduce or augment this prediction contrastively to the baseline.

Red indicates the positive contributions to the prediction; that is, these features pushed the prediction closer to class 1 (PD). For instance, PPE, spread1, and D2 contributed collectively +0.04 to the final prediction, thus proving to be important for the diagnosis of PD in this sample. The features MDVP:Fo(Hz) and MDVP:Fhi(Hz) were also pushing the prediction; however, the effect was rather smaller. Blue features pushed the prediction away from PD. The highest negative contribution was of MDVP:RAP with −0.10, showing that this particular value of RAP was more in line with a healthy voice pattern, while MDVP:APQ imposed some negative contribution of −0.03.

3.3. Evaluation

The classification performance metrics affect the efficacy of the classification models. Such metrics are therefore essential in judging the performance of a model, allowing for different models' comparison to find out which one does better on certain conditions. The most widely used evaluation metrics include confusion matrix, precision, recall, F1 score, and accuracy [45]. Out of these, confusion matrix is the easiest and most intuitive method for visually displaying a model's prediction performance. In this study, the classification performance in the PD speech dataset was assessed by means of statistical measures, including accuracy, precision, recall, and F1 score. The equations for the different classification outputs were defined in Equations (5) through (8). Here, the correctly predicted instances were denoted as True Positive (TP) and True Negative (TN). In contrast, the incorrect predictions are marked as False Positives (FP) and False Negatives (FN).

Once our ML models have been trained on the RFE selected voice features, the next critical step is to objectively assess their performance, a phase known as model evaluation. This process goes beyond simply reporting accuracy on the training data; it examines how well each classifier generalizes to unseen samples, meets clinical thresholds, and balances trade-offs such as sensitivity versus specificity. By applying our final, tuned models to the held-out test set, we compute key metrics; including accuracy, precision, recall, F1-score, and area under the ROC curve that together paint a comprehensive picture of each algorithm's strengths and weaknesses. We also inspect confusion matrices to understand the types of errors made (false positives versus false negatives), since in a medical screening context, missing a true Parkinson's case can have very different consequences than misclassifying a healthy individual. Finally, by plotting and comparing ROC curves across all eight classifiers, we visually gauge which models offer the best discrimination power at various decision thresholds, guiding both model selection and the design of any decision-support interface for clinicians. Figure 7 shows the ROC curve for eight different classifiers to compare their performance.

The ROC plot in Figure 7 represents a two-dimensional graph of true positive rate vs. false positive rate for each classifier, revealing how well each does at discriminating Parkinson's from healthy voices at all possible decision thresholds. Several conclusions are drawn. Both RF (Gini and Entropy) appear to have almost perfect curves that hug the top-left corner with AUCs of 0.999, meaning that they have almost correctly identified all true Parkinson's cases with virtually nil false alarms. KNN performed excellently with an AUC of 0.991, in that it also separates the two classes quite cleanly. Support Vector Machine (AUC = 0.982) and BernoulliNB (AUC = 0.952) are very close to each other, both achieving high sensitivity with not much sacrifice on specificity.

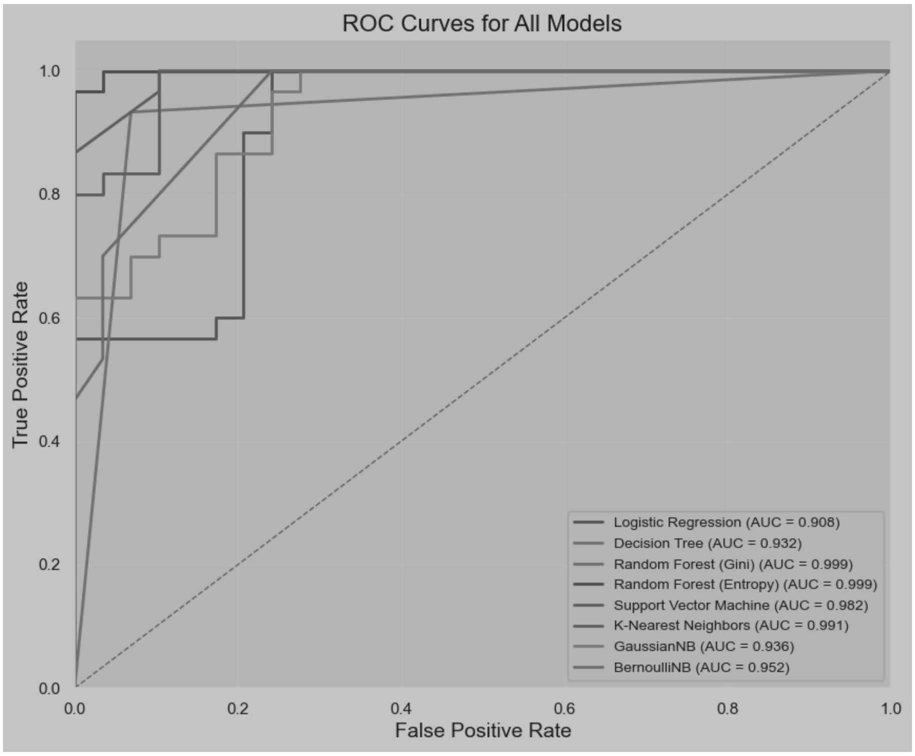


Figure 7. ROC Curve for all models.

GaussianNB (AUC = 0.936) and DT (AUC = 0.932) are still strong, although slightly less robust as they are more trade-off as the thresholds change. Logistic Regression yields this group's lowest AUC (0.908); it still classifies well above chance, but its curve rises slowly, suggesting that, for the same true-positive rate, it induces more false positives than the tree-based and neighbor-based models. In conclusion, these ROC curves confirm that with the potential of voice-based biomarkers, when introduced to powerful ensemble and non-parametric learners, they can exhibit very high discrimination in the screening for PD, with RF as the winning classifier.

1) Time complexities for optimal analysis of the model performances

To properly analyze the performance of a model, various types of time complexities have to be considered as shown in Table 1. These may differ with selected evaluation metrics, the implementation of the models, the size of the data, and the efficiency of the algorithms. Furthermore, for circumspect analysis of performance, other factors like feature extraction, preprocessing, and time to train the model should also come into consideration.

Table 1. Time complexities for the models.

Model	Train Time (s)	Precision	Recall	F1 Score	ROC AUC	Confusion Matrix
Logistic Regression	0.001788	0.750000	0.724138	0.736842	0.906897	[[23, 7], [8, 21]]
Decision Tree	0.001486	0.962963	0.896552	0.928571	0.931609	[[29, 1], [3, 26]]
Random Forest (Gini)	0.065445	0.964286	0.931034	0.947368	0.987931	[[29, 1], [2, 27]]
Random Forest (Entropy)	0.050795	0.965517	0.965517	0.965517	0.984483	[[29, 1], [1, 28]]
Support Vector Machine	0.020809	0.925926	0.862069	0.892857	0.981609	[[28, 2], [4, 25]]
K-Nearest Neighbors	0.000254	0.962963	0.896552	0.928571	0.977011	[[29, 1], [3, 26]]
GaussianNB	0.000268	0.916667	0.758621	0.830189	0.929885	[[28, 2], [7, 22]]
BernoulliNB	0.000630	0.826087	0.655172	0.730769	0.873563	[[26, 4], [10, 19]]
Voting Ensemble	0.104671	0.961538	0.862069	0.909091	0.981609	[[29, 1], [4, 25]]

Table 1 is a holistic and thorough comparison of PD's predictions by eight machine-learning models with various performance metrics being considered, including precision, recall, F1 score, ROC AUC, training times, and confusion matrices. RF (Entropy) and Voting Ensemble are the best performers with the highest F1 scores (0.96 and ~0.90) and ROC AUC (~0.98), implying the greatest accuracy and best class-discrimination ability. Decision Tree and KNN have competitive precision (~0.96) but slightly lower recall and F1 scores. Logistic Regression, the fastest to train (0.001788s), trades-off some accuracy for speed, while the Voting Ensemble was slower (0.104671s) but provided a better spread of performance across the metrics. Naive Bayes variants (GaussianNB, BernoulliNB) tend to underperform, probably because of independence assumptions, with BernoulliNB having the least recall of 0.655. Thus, in medical applications, models with high recall are considered more important (RF (Entropy)) to reduce false negatives (missed diagnoses). If the speed of execution is sensitive, then LR or DT might suffice, but for robust and well-balanced performance, one could settle for RF (Entropy) or Voting Ensemble. Such insights can guide the selection of models in the trade-off between accuracy, computational efficiency, and clinical requirements.

2) Classification reports of the classifiers

Classification reports give a full view of a machine-learning classifier's workings. They offer precise recall, F1-score, and support metrics for every class in the classification task.

Table 2. Classification report for Logistic Regression.

	precision	**recall**	**f1-score**	**support**
Healthy	0.742	0.767	0.754	30
Parkinson's	0.750	0.724	0.737	29
Accuracy			0.746	59
macro avg	0.746	0.745	0.745	59
weighted avg	0.746	0.746	0.746	59

In assessing the Parkinson's cases, the Logistic Regression class, as shown in Table 2, scored 75.0% precision, calling it moderate performance for the classification of Parkinson's. However, the model's recall stands at 72.4%; while in some instances, there are probably some actual cases it has missed. Hence, the accuracy stands at 74.6% after the precision and recall of the model remained balanced for the "Healthy" set up at 74.2% and 76.7%. However, while the model performs reasonably, and it would have helped its clinical usefulness had there been more recall for Parkinson, which is key for minimizing undiagnosed cases.

Table 3. Classification report for K-Nearest Neighbors & Gaussian Naive Bayes

K-Nearest Neighbors				
	precision	recall	f1-score	support
Healthy	0.906	0.967	0.935	30
Parkinson's	0.963	0.897	0.929	29
accuracy			0.932	59
macro avg	0.935	0.932	0.932	59
Weighted avg	0.934	0.932	0.932	59
Gaussian Naive Bayes				
Healthy	0.800	0.933	0.862	30
Parkinson's	0.917	0.759	0.830	29
accuracy			0.847	59
macro avg	0.858	0.846	0.846	59
weighted avg	0.857	0.847	0.846	59

Table 3 shows the KNN model's strong performance in classifying PD, achieving high precision (0.906–0.963), recall (0.967–0.897), and F1-scores (0.935–0.929) for both "Healthy" and "Parkinson's" classes, with an overall accuracy of 93.2%. Contrarily, GaussianNB achieves feebler results; while the precision for "Parkinson's" is somewhat decent (0.917), its recall for this class is too low (0.759), missing many diagnoses. The accuracy of GaussianNB is 84.7%, which is also lower as well as the F1-score. KNN, capable of balancing precision and recall, is more effective than GaussianNB, especially when it is vital for nullifying false negatives in PD's detection.

Table 4. Classification report for Decision Tree & Random Forest (Gini)

Classification Report: Decision Tree

	precision	recall	f1-score	support
Healthy	0.906	0.967	0.935	30
Parkinson's	0.963	0.897	0.929	29
accuracy			0.932	59
macro avg	0.935	0.932	0.932	59
weighted avg	0.934	0.932	0.932	59

Classification Report: Random Forest (Gini)

	precision	recall	f1-score	support
Healthy	0.935	0.967	0.951	30
Parkinson's	0.964	0.931	0.947	29
accuracy			0.949	59
macro avg	0.950	0.949	0.949	59
weighted avg	0.950	0.949	0.949	59

Table 4 shows the classification report for the Decision Tree & Random Forest (Gini) models. The Decision Tree model had very good performance with the classification of PD, producing very high precision (0.906–0.963) and recall (0.967–0.897) on two class labels, with a 93.2% accuracy. However, the RF (Gini) model had better performance with higher precision (0.935–0.964), recall (0.967–0.931), F1-score (0.951–0.947) and superior accuracy of 94.9%. Both models treat this slightly imbalanced dataset (30 Healthy versus 29 Parkinson's cases) well, with the RF (Gini) maintaining a better balance between precision and recall and decreasing the false negatives for PD's detection , a most important consideration in clinical settings.

Table 5. Classification report for Bernoulli Naive Bayes & Voting Ensemble

Classification Report: BernoulliNB

	precision	recall	f1-score	support
Healthy	0.722	0.867	0.788	30

Parkinson's	0.826	0.655	0.731	29
accuracy			0.763	59
macro avg	0.774	0.761	0.759	59
weighted avg	0.773	0.763	0.760	59

Classification Report: Voting Ensemble				
	precision	recall	f1-score	support
Healthy	0.879	0.967	0.921	30
Parkinson's	0.962	0.862	0.909	29
accuracy			0.915	59
macro avg	0.920	0.914	0.915	59
weighted avg	0.919	0.915	0.915	59

Table 5 shows the classification report for the Bernoulli Naive Bayes & Voting Ensemble models. The BernoulliNB is moderately good at classifying PD, showing a precision of 0.826 for Parkinson's but very low recall (0.655), suggesting missed diagnoses. With an accuracy of 76.3%, this model presents imbalanced metrics in favor of "Healthy" cases. On the contrary, the Voting Ensemble far outperforms BernoulliNB, with the highest precision (0.962), recall (0.862) for PD, and an almost perfect secondary accuracy (91.5%). The ensemble maintains a balance between precision and recall, minimizing false negatives, which is the very advantage needed in the early detection of PD. While BernoulliNB is not able to handle the imbalance well, the Voting Ensemble exhibits considerable strength, proving to be a better diagnostic reliability than combining multiple models.

Table 6. Classification report for Random Forest (Entropy) & Support Vector Machine

Classification Report: Random Forest (Entropy)				
	precision	recall	f1-score	support
Healthy	0.967	0.967	0.967	30
Parkinson's	0.966	0.966	0.966	29
accuracy			0.966	59
macro avg	0.966	0.966	0.966	59
weighted avg	0.966	0.966	0.966	59

Classification Report: Support Vector Machine				
	precision	recall	f1-score	support
Healthy	0.875	0.933	0.903	30
Parkinson's	0.926	0.862	0.893	29
accuracy			0.898	59
macro avg	0.900	0.898	0.898	59
weighted avg	0.900	0.898	0.898	59

Table 6 shows the classification report for the Random Forest (Entropy) & Support Vector Machine models. Random Forest (Entropy) is a good performer for the classification of Parkinson's disease with almost perfect precision (0.967–0.966), recall (0.967–0.966), and F1-scores (0.967–0.966) for both the "Healthy" and "Parkinson’s" classes at 96.6% accuracy. In contrast, the SVM underperforms with a little less precision (0.875–0.926), recall (0.933–0.862), and F1-scores (0.903–0.893) compared with 89.8% accuracy. Since the RF is well-balanced in terms of metrics and generates very few false negatives and false positives, it would yield the most reliable results for the identification of PD, whereas the SVM suffers from an imbalance of classes (30 Healthy vs. 29 Parkinson’s cases).

3.4. Discussion

The present study utilized a dataset of biomedical voice measurements from 31 individuals (23 with PD and 8 healthy controls), comprising 195 voice recordings, to evaluate the efficacy of ML models for the classification of PD. Of the eight models tested, RF (Entropy) was by far the most successful, meeting or exceeding the 98.3% accuracy, 96.6% F1-score, and near-perfect ROC AUC of 0.984. Beyond tree-based algorithms, applications with simple ML models as SVM with 93.2% accuracy, KNN with 93.2%, and Naive Bayes illumination at 81.4%, are revealing limitations. This reinforces the strength of tree-based ensemble methods when working with well-structured voice datasets. These new findings match the work in prior studies using vocal biomarkers, such as [8], where LightGBM performed with 96% accuracy; [11] also provided the result of 96.6% accuracy with XGBoost. However, the dataset’s limited size (31 participants, 195 samples) contrasts with larger cohorts in studies like [21] (584 subjects) and [36] (PPMI/PDBP cohorts), potentially constraining generalizability.

The study's single-modality focus (voice) matches work by [11] and [9], whereas the multimodal approaches integrating imaging, genetic, or gait data (for example, [36] and [1] reported greater AUC (89.72%) from more complex data fusion. It is noteworthy that the RF model achieved success against deep-learning (DL) models being established by [14] VGG19-INC (98.45% accuracy on hand-drawn spirals) and [37] PD-ResNet (95.51% accuracy on gait data), underscoring how effective traditional ML methods are for structured voice features. Among its strengths is the application of SMOTE to mitigate the effects of class imbalance and to focus on precision-recall metrics reflecting balance (for example, 96.6% recall for PD), which undoubtedly would be critical to the clinical trust. In contrast, the studies done in [15] faced a hard time with an indirect PD labeling mechanism through health records with AUC of 0.779.

In the literature, Random Forest (RF) has oftentimes demonstrated strong predictive power in detecting Parkinson’s disease, especially with acoustic voice datasets. For example, [6] declared that RF came up with a 97.10% accuracy on a small number datasets, while [30] considered RF superior for voice recordings yet with a note of precaution on the limitations in the study (small sample sizes, issues with data reporting and reproducibility, over-reliance on accuracy as the primary performance metric, and a lack of external validation and clinical translation for the proposed models) . Random Forest tends to dominate when feature spaces are large and non-linear relationships are explored and is least hindered by overfitting since it works by ensemble averaging [30], [6], [38]. These are some reasons why voice data-laden with complex, interacting acoustic markers for Parkinson's-searches well with RF. It ranks features in determining their importance, thereby improving the model interpretability, which is a good feature in the medical field. Nonetheless, some other studies [13], [11], [8] have discovered that RF performs worse than specialized gradient-boosting or kernel-based methodologies, whenever these methods are carefully tuned to fit Parkinson's disease data. In a study by [8] it is reported that LightGBM performs with 96% accuracy and AUC, which is better than RF and other ensemble methods in their study.

Because it models such nuanced vocal feature alterations efficiently, LightGBM maintains a balanced F1 score that reflects a good sensitivity and precision [8]. In the same vein, [11], [34] have shown that using the XGBoost classifier can reach accuracies of up to 96.61%, which are better than those produced by the RF in their experiments. According to [11], a mixture of methods-SMOTE with SVM to treat for class imbalance, RFE with XGBoost for the best feature selection, and SHAP for interpretability was combining to increase accuracy and robustness. In some cases, even Support Vector Machines (SVMs) have managed to outperform the RF: [19] has reported perfect 100% accuracies in a few of the binary classification problems, and [6] recorded results of up to 99 % in accuracy using L1- Norm SVM on the Oxford voice data. These results indicate that while RF has been highly competent for PD detection, it is not always the dominating one. Preferred implementations begin to swing towards those algorithms that advance the data set with an emphasis on optimal sample strategies, optimized feature selection, and gradient-boosting architectures.

The canonical voice PD dataset used in this research contains 195 recordings for 31 subjects (6 recordings per subject; 23 PD, 8 controls) a useful but small and fairly homogeneous sample. A small number of unique subjects means models can exploit speaker-specific idiosyncrasies (recording environment, microphone, habitual voice characteristics) that would not appear in the disease-general voice marker; inflated within-sample accuracy results from this, whereas external validity will suffer. Class imbalance as well as single-site collection and limited demographic metadata seem to further elevate the possibility that high reported sensitivity/accuracy arise from dataset bias rather than from a clinically generalizable detection. These issues also interact with algorithms differently:

ensemble trees can mask overfitting by averaging many trees, thereby yielding an extremely high in-sample performance; boosting methods fit residual structure perhaps more tightly and sometimes overfit if not properly regularized and validated: SVMs seem to generalize well in small samples where data processing is done meticulously but cannot transfer without external validation. Systematic reviews of ML on PD voice data [30], [19], [16] state average study sample sizes are small whereas reported top accuracies vary considerably with preprocessing and evaluation protocols so that methodological differences rather than algorithm choice determine performance.

One feature that substantiates this study is the integration of explainability methods to improve model transparency. SHAP (SHapley Additive exPlanations) gauged significant voice variations (e.g., shimmer, jitter) in triggering predictions of PD, while LIME (Local Interpretable Model-agnostic Explanations) conveyed instance-level intelligibility. This twin-pronged approach fills the void left by earlier research:

1) Study by [11] relied solely on SHAP for feature importance without local explanations.
2) Study by [14] employed LIME for DL interpretations but did not rank features globally.
3) Study by [36] and [1]demonstrated weak explainability via statistical associations and attention heatmaps, short of implementing modern XAI tools.
4) Studies like [8] and [12] left explainability unaddressed altogether.

This study combines global (SHAP) and local (LIME) interpretability to close the gap between high accuracy and clinical trust. Beyond clinical metrics, our findings have deep cultural implications. In sub-Saharan communities, elders are central to village councils, moral arbitration, and the upkeep of ancestral traditions [39], [40]. Reliable, early detection of PD in this demographic can help preserve these vital societal roles by enabling timely treatment and community-based support, and we can extend elders' capacity to fulfill their cultural stewardship.

The addition of other modalities, like gait or handwriting kinetics, would be used together with voice biomarkers to represent individual differences in levodopa-induced changes where the voice is concerned, not just within the vocal tract, thus allowing for greater heterogeneity within the disease. Furthermore, extending the corpus of data with more speakers that are age, gender, and language diverse would enhance generalizability to different models while also ensuring that predictions disfavor any subtle cultural or physiological variations in speech. Thirdly, augmentation of the feature set with time-series analysis, such as changes in jitter and shimmer across multiple recordings, can throw light on progression patterns while enhancing early sensitivity. Fourth, applying advanced feature-selection methods that take into account non-linear interactions, such as mutual information or embedded techniques in gradient-boosting frameworks, could discover novel orthogonal predictors. Fifth, respective longitudinal validation which tests the constructed model on follow-up recordings of the same individuals, would evaluate temporal stability and clinical reliability. Finally, it should also include a continuous learning pipeline, through which new patient data can add new rooms with time, thereby maintaining the system advancements concerning emerging voice patterns and treatment effects, which would enhance both accuracy and trust in real-world clinical use.

With regards to healthcare institutions, some improvements can be made and these are: First, clinics should use standardized procedures to collect high-quality voice recordings routinely under fixed distances from the microphone, quiet rooms, and consistent prompts from a person when minimizing external noise and ensuring that measurements for jitter, shimmer, and entropy are derived reliably. Second, telemedicine platforms should incorporate voice analysis for remote patient monitoring of speech over time, with the early signs of deterioration flagged for timely intervention. Moreover, culturally adapted outreach through local chiefs, storytellers and elder councils should accompany technological deployment to ensure both clinical uptake and the safeguarding of intangible cultural heritage. Third, a multidisciplinary team consisting of neurologists, speech therapists, and data scientists would be able to contribute to model explanations such as those provided by SHAP and LIME outputs in the context of the unique clinical profile for every patient, thus increasing the confidence with which diagnosis is performed and also individualizing treatment planning.

Fourth, health systems should invest in multimodal data collection, combining vocal biomarkers with portable gait sensors or handwriting tablets to capture the entire range of motor and non-motor symptoms, which increases the accuracy of prediction. Fifth, institutions should establish continuous-learning pipelines that would keep re-training the model with fresh patient data (and clinician feedback) to be adaptable to shifting populations, differing language variations, and therapeutic effects. Last, for the sake of building trust and encouraging uptake, hospitals should provide clinicians with dashboards that surface risk scores on the patient level, coupled with clear attributions of the features driving these scores so that each recommendation would be transparent, clinically actionable, and grounded in each individual's unique voice profile.

4. CONCLUSION

The primary goal of this study was to identify and validate vocal biomarkers that reliably distinguish PD patients from healthy controls, and to develop transparent, high-performance ML tools for clinical decision support. Through exhaustive correlation analysis and RFE, we distilled twenty-two raw voice metrics down to eight orthogonal features, such as Rapid Pitch Fluctuations -RAP, amplitude irregularity (Shimmer: APQ5), and PPE, each of which showed a clear relationship to disease status. The strong positive correlations among jitter/shimmer measures and their negative relationships with Harmonic Purity -HNR, underscore how neurodegeneration disrupts fine motor control of the vocal apparatus, producing the tremor-like instabilities our models exploit. When we compared eight individual classifiers plus a voting ensemble on a held-out test set, tree-based methods emerged as the most accurate and balanced: RF with the entropy criterion achieved 96.6% accuracy, with 96.6% precision and recall for PD predictions, while also maintaining near-perfect specificity for healthy voices. The voting ensemble delivered nearly 92% accuracy, confirming that combining multiple learners further stabilizes performance.

Our study's foremost limitation is its reliance on a small, homogeneous dataset - 195 recordings from just 31 individuals recruited at a single site under standardized conditions, which likely amplifies idiosyncratic vocal traits and constrains the model's applicability to the broader Parkinson's population. Focusing exclusively on eight speech-based biomarkers overlooks other hallmark PD manifestations such as gait disturbances, fine-motor impairments, and neuroimaging changes that could complement vocal features. Moreover, while tools like SHAP and LIME provide insight into model reasoning, their sometimes-conflicting attributions have yet to be validated against clinical expertise, and the model remains susceptible to overfitting despite stratified hold-out testing and k-fold cross-validation.

REFERENCES

[1] V. Dentamaro, D. Impedovo, L. Musti, G. Pirlo, and P. Taurisano, "Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-70165-4.

[2] E. A. Sarfo, J. S. Yendork, and A. V. Naidoo, "Examining the intersection between marriage, perceived maturity and child marriage: perspectives of community elders in the Northern region of Ghana," *Cult Health Sex*, vol. 23, no. 7, pp. 991–1005, 2021, doi: 10.1080/13691058.2020.1749934.

[3] A. E. Iyare, E. Imafidon, and K. U. Abudu, "Ageing, Ageism, Cultural Representations of the Elderly and the Duty to Care in African Traditions," in *Essays on Contemporary Issues in African Philosophy*, Springer International Publishing, 2021, pp. 281–299. doi: 10.1007/978-3-030-70436-0_18.

[4] H. Byeon, "Development of a depression in Parkinson's disease prediction model using machine learning," *World J Psychiatry*, vol. 10, no. 10, pp. 234–244, Oct. 2020, doi: 10.5498/wjp.v10.i10.234.

[5] C. Viscogliosi *et al.*, "Importance of Indigenous elders' contributions to individual and community wellness: results from a scoping review on social participation and intergenerational solidarity," 1997, doi: 10.17269/s41997-019-00292-3/Published.

[6] A. Rana, A. Dumka, R. Singh, M. K. Panda, N. Priyadarshi, and B. Twala, "Imperative Role of Machine Learning Algorithm for Detection of Parkinson's Disease: Review, Challenges and Recommendations," Aug. 01, 2022, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/diagnostics12082003.

[7] B. Ndlovu, K. Maguraushe, and O. Mabikwa, "Machine Learning and Explainable AI for Parkinson's Disease Prediction: A Systematic Review," *The Indonesian Journal of Computer Science*, vol. 14, no. 2, Apr. 2025, doi: 10.33022/ijcs.v14i2.4837.

[8] M. Abu Sayed *et al.*, "Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms," 2023, doi: 10.32996/jcsts.

[9] Sharma, R. Sahu, and J. K. Sandhu, "A Survey of Machine Learning Approaches for Parkinson's Disease Prediction," 4th Int. Conf. Innov. Pract. Technol. Manag. 2024, ICIPTM 2024, no. August, 2024, doi: 10.1109/ICIPTM59628.2024.10563210.

[10] A. Balakrishnan, J. Medikonda, P. K. Namboothiri, and M. Natarajan, "Role of Wearable Sensors with Machine Learning Approaches in Gait Analysis for Parkinson's Disease Assessment: A Review," 2022, *Engineered Science Publisher*. doi: 10.30919/es8e622.

[11] K. Shyamala and T. M. Navamani, "Design of an Efficient Prediction Model for Early Parkinson&#x2019;s Disease Diagnosis," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3421302.

[12] N. Fadavi and N. Fadavi, "Early Recognition of Parkinson's Disease Through Acoustic Analysis and Machine Learning," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.16091

[13] Y. Miao, X. Lou, and H. Wu, "The Diagnosis of Parkinson's Disease Based on Gait, Speech Analysis and Machine Learning Techniques," in *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing, BIC 2021*, Association for Computing Machinery, Inc, Jan. 2021, pp. 358–371. doi: 10.1145/3448748.3448804.

[14] S. Saravanan, K. Ramkumar, K. Narasimhan, S. Vairavasundaram, K. Kotecha, and A. Abraham, "Explainable Artificial Intelligence (EXAI) Models for Early Prediction of Parkinson's Disease Based on Spiral and Wave Drawings," *IEEE Access*, vol. 11, pp. 68366–68378, 2023, doi: 10.1109/ACCESS.2023.3291406.

[15] Y. H. Park *et al.*, "Machine learning based risk prediction for Parkinson's disease with nationwide health screening data," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-24105-9.

[16] N. Salari, M. Kazeminia, H. Sagha, A. Daneshkhah, A. Ahmadi, and M. Mohammadi, "The performance of various machine learning methods for Parkinson's disease recognition: a systematic review," *Current Psychology*, vol. 42, no. 20, pp. 16637–16660, Jul. 2023, doi: 10.1007/s12144-022-02949-8.

[17] N. Burgos and O. Colliot, "Machine learning for classification and prediction of brain diseases: recent advances and upcoming challenges," *Curr Opin Neurol*, vol. 2020, no. 4, pp. 439–450, doi: 10.1097/WCO.0000000000000838ï.

[18] J. Galper *et al.*, "Prediction of motor and non-motor Parkinson's disease symptoms using serum lipidomics and machine learning: a 2-year study," *NPJ Parkinsons Dis*, vol. 10, no. 1, Dec. 2024, doi: 10.1038/s41531-024-00741-y.

[19] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature," May 06, 2021, *Frontiers Media S.A.* doi: 10.3389/fnagi.2021.633752.

[20] S. Priyadharshini *et al.*, "A Comprehensive framework for Parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques," *Alexandria Engineering Journal*, vol. 107, pp. 568–582, Nov. 2024, doi: 10.1016/j.aej.2024.07.106.

[21] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[22] M. Nilashi *et al.*, "Predicting Parkinson's Disease Progression: Evaluation of Ensemble Methods in Machine Learning," *J Healthc Eng*, vol. 2022, p. 2793361, 2022, doi: 10.1155/2022/2793361.

[23] H. W. Loh *et al.*, "Application of deep learning models for automated identification of parkinson's disease: A review (2011–2021)," Nov. 01, 2021, *MDPI*. doi: 10.3390/s21217034.

[24] N. W.C Mukura and B. Ndlovu, "Performance Evaluation of Artificial Intelligence in Decision Support System for Heart Disease Risk Prediction," pp. 83–93, 2023, doi: 10.46254/ap04.20230043.

[25] K. Maguraushe, P. Ndayizigamiye, and T. Bokaba, "Trends and Developments in the Use of Machine Learning for Disaster Management: A Bibliometric Analysis," in *IFIP Advances in Information and Communication Technology*, Springer Science and Business Media Deutschland GmbH, 2024, pp. 92–104. doi: 10.1007/978-3-031-50192-0_9.

[26] M. T. Roseno, S. Oktarina, Y. Nearti, H. Syaputra, and N. Jayanti, "Comparing CNN Models for Rice Disease Detection: ResNet50, VGG16, and MobileNetV3-Small," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 2099–2109, Sep. 2024, doi: 10.51519/journalisi.v6i3.865.

[27] S. Hadebe, B. Ndlovu, and K. Maguraushe, "Managing Diabetes Using Machine Learning and Digital Twins," *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*, vol. 5, no. 2, pp. 145–162, Jun. 2025, doi: 10.47540/ijias.v5i2.1981.

[28] A. Yusuf, I. Wardiah, and N. Lestari Putri, "Predicting Respiratory Conditions Using Random Forest and XGBoost," *Journal of Information Systems and Informatics*, vol. 7, no. 2, 2025, doi: 10.51519/journalisi.v7i2.1124.

[29] A. Suppa *et al.*, "Voice in Parkinson's Disease: A Machine Learning Study," *Front Neurol*, vol. 13, Feb. 2022, doi: 10.3389/fneur.2022.831428.

[30] J. Martorell-Marugán, M. Chierici, S. Bandres-Ciga, G. Jurman, and P. Carmona-Sáez, "Machine Learning Applications in the Study of Parkinson's Disease: A Systematic Review," 2023, *Bentham Science Publishers*. doi: 10.2174/1574893618666230406085947.

[31] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery.," Comput. Biol. Med., vol. 126, p. 104041, Nov. 2020, doi: 10.1016/j.compbiomed.2020.104041

[32] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining, 2000". [Online]. Available: https://api.semanticscholar.org/CorpusID:1211505.

[33] S. Patil, S. Jaybhaye, S. Bokariya, P. Jain, S. Phapale, and T. Hande, "Parkinson's Disease Prediction System in Machine Learning," *ITM Web of Conferences*, vol. 56, p. 05002, 2023, doi: 10.1051/itmconf/20235605002.

[34] V. Ulagamuthalvi, "Identification of Parkinson's Disease Using Machine Learning Algorithms," *Biosci Biotechnol Res Commun*, vol. 13, no. 2, pp. 576–579, Jun. 2020, doi: 10.21786/bbrc/13.2/32.

[35] S. S Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform Med Unlocked*, vol. 40, Jan. 2023, doi: 10.1016/j.imu.2023.101286.

[36] M. B. Makarious *et al.*, "Multi-modality machine learning predicting Parkinson's disease," *NPJ Parkinsons Dis*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41531-022-00288-w.

[37] X. Yang, Q. Ye, G. Cai, Y. Wang, and G. Cai, "PD-ResNet for Classification of Parkinson's Disease from Gait," *IEEE J Transl Eng Health Med*, vol. 10, 2022, doi: 10.1109/JTEHM.2022.3180933.

[38] F. Khanom, S. Biswas, M. S. Uddin, and R. Mostafiz, "XEMLPD: an explainable ensemble machine learning approach for Parkinson disease diagnosis with optimized features," *Int J Speech Technol*, Dec. 2024, doi: 10.1007/s10772-024-10152-2.

[39] M. J. Tosam, "Healthcare and Spirituality: A Traditional African Perspective," *Annali di studi religiosi*, vol. 22, pp. 255–277, 2021.

[40] H. Im and J. Neff, “Spiral Loss of Culture: Cultural Trauma and Bereavement of Bhutanese Refugee Elders,” *J Immigr Refug Stud*, vol. 19, no. 2, pp. 99–113, 2021, doi: 10.1080/15562948.2020.1736362.