Proceedings of the 9th International Conference on Inventive Systems and Control (ICISC-2025)
IEEE Xplore Part Number: CFP25J06-ART; ISBN: 979-8-3315-1247-7

Prediction of Parkinson's Disease using Machine Learning

Dinesh Kumar.B
PG Student (MCA)
Department of Computer Applications
Hindustan Institute of Technology and Science
Chennai-India
23cp2170004@student.hindustanuniv.ac.in

Dr.K.France
Assistant Professor
Department of Computer Applications
Hindustan Institute of Technology and Science
Chennai-India
francek@hindustanuniv.ac.in

2025 9th International Conference on Inventive Systems and Control (ICISC) | 979-8-3315-1247-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICISC65841.2025.11187909

Abstract - Parkinson's disease (PD) is a progressively worsening neurodegenerative disorder that distinctly impairs movement and speech. Timely identification of PD is crucial in healthcare management to improve outcomes and sustain life quality. Voice disorder shows potential for timely detection as it is observed in more than 70% of PD patients, which is a promising early indicator. This study proposes interpretable and robust machine learning frameworks for early-stage PD detection using speech features. The IF (Interpretable Feature Ranking) XGBoost model combines class balancing techniques with SVMSMOTE, advanced feature selection with Recursive Feature Elimination (RFE), and explainable AI through SHAP to improve model transparency. This approach not only increases accuracy but also reveals essential speech-related biomarkers predictive of PD. Numerous experiments validate that the XGBoost classifier is the best-performing model with 96.61% accuracy. The approach addresses the challenge of providing a reliable and interpretable answer for early PD detection, which assists clinicians in prompt action and tailored intervention.
Keywords— Parkinson's Disease, Speech Analysis, Machine Learning, XGBoost, Dysphonia, Early Diagnosis, Feature Selection, SHAP, Explainable AI, SVMSMOTE, Interpretable Model, Voice Biomarkers, Classification, Neurodegenerative Disorders, Predictive Analytics.

I. INTRODUCTION

Parkinson's Disease (PD) is a slowly progressive disorder resulting from the degeneration of certain brain structures, and it begins almost exclusively with motor symptoms, including tremors, stiffness, slowness of movement, and difficulties with balance and posture. More recently, non–speech and cognitive functions have been identified as possible earlier signs. Within this group, dysphonia, which involves a change in the voice quality, pitch, and speech, highlights a promising early detection marker. Presumably dependent on clinical evaluations, traditional diagnostic methods highlight the absence of objective, early diagnostics, non-invasive, low-cost, and capable frameworks for monitoring disease progression. The ability to use machine learning (ML) for multidimensional data analysis and pattern recognition has transformed the biomedical diagnostic industry. Machine learning algorithms not only apply to voice data, but their computational power can significantly surpass human evaluators even when analyzing complex data sets with multiple layers. Among these techniques, Extreme Gradient Boosting (XGBoost) has outperformed in multiclass classification problems due to its efficiency, reliability, accuracy, and interoperability with interpretability tools such as SHAP (SHapley Additive explanations), which provided performance interpretability on features contributing to model decisions. This study constructs an unexplored framework for the diagnosis of Parkinson's disease, applying XGBoost with result interpretation through SHAP, class imbalance addressed by SVM-SMOTE, feature attribution, and model tuned on class-imbalanced data. Medical transparency is vital for health care applications, making it equally important alongside accuracy. This research aims to improve the reliable and accessible systems for early PD detection by analyzing voice biomarkers using advanced ML techniques, thereby facilitating timely intervention strategies and improving outcomes for patients.

II. RELATED WORK

Recent advancements in machine learning (ML) and deep learning (DL) have revolutionized the diagnosis and management of neurodegenerative disorders such as Parkinson's Disease (PD). The growing global prevalence of PD, projected to rise significantly by 2050 due to demographic transitions, has intensified the need for early detection technologies that are non-invasive, cost-effective, and clinically reliable [1]. Several studies have explored the application of ML for PD prediction using voice or biomedical signal data. Chawla et al. demonstrated the effectiveness of recursive feature elimination (RFE) in combination with nature-inspired feature selection methods for robust classification of PD [2]. Alalayah et al. focused on the early detection of PD using acoustic signal analysis,

confirming that variations in voice parameters can serve as valuable biomarkers [3]. Although some studies focus on related biomedical applications, they provide transferable insights. Ahmed et al. proposed hybrid techniques for leukemia diagnosis, highlighting the value of combining multiple models for accuracy, an approach that is equally beneficial in PD diagnostics [4]. Aishwarya et al. proposed a feature selection strategy using the Fisher score and RFE to improve PD prediction accuracy [5]. Among classification models, Extreme Gradient Boosting (XGBoost) has emerged as a high-performing algorithm. E.S. and V.D. RS demonstrated XGBoost's predictive power in classifying PD based on voice features, showing significant gains in model efficiency and accuracy [6]. The importance of choosing the right feature selection method was reviewed by Maguire et al., who surveyed several ranking strategies relevant to high-dimensional datasets like those used in PD studies [7]. Steigmann et al. extended ML's diagnostic utility by applying classification techniques in dental morphology, showcasing the adaptability of ML across different medical domains [8]. Lamba et al. contributed significantly to PD literature by proposing hybrid systems [9] and systematic approaches using kinematic features to improve diagnosis [10]. Velu and Jaisankar presented an end-to-end ML-based early prediction model specifically for PD, validating their system with high performance on benchmark datasets [11]. Their work underscores the importance of real-world deployment and explainability in ML healthcare applications. Dimensionality reduction and ensemble learning techniques have also proven beneficial. Liu et al. introduced a local discriminant preservation projection embedded within ensemble classifiers, which enhances the separability of PD-related data clusters [12]. Soumaya et al. showed strong performance when dealing with noisy and imbalanced data [13]. Although PD is the focus of this study, feature selection methods used in related areas also provide valuable methodological parallels. For instance, Senan et al. applied correlation-based feature selection for heart failure diagnosis, emphasizing the role of precise variable ranking in medical ML applications [14]. Lastly, Khan et al. provided a comprehensive review of ML and DL techniques for various brain disease diagnostics, including PD, highlighting the shift towards hybrid and explainable systems that can operate reliably across the clinical environment [15].

III. PROPOSED WORK

The approach utilizes sophisticated machine learning methods for voice recognition while making sure that data pre-processing and model class imbalance techniques guarantee precision and clinical significance for the reliable detection of Parkinson's Disease (PD). Initially, the voice dataset requires extensive normal feature extraction, feature scaling, and denoising to eliminate inconsistencies and prepare the data for model training. Due to the prevalence of class imbalance and under-representation of healthy cases (particularly in the context of PD detection), the SVM-SMOTE is adopted. This method creates enhanced samples in the underclass space defined by SVM boundaries, which mitigates bias and leads to better model generalization. Voice-related biomarkers should also be selected, which are jitter, shimmer, noise ratios, and harmonic ratios relevant for inter-voicing assessment to extract and predict the disease using this information. The primary classifier is based on Extreme Gradient Boosting (XGBoost)—an ensemble learning method that builds additive models using iterative boosting and regularization, which reduces the model classification error. XGBoost is known for its high-speed performance and accuracy, especially when processing structured datasets with interrelated, codified features. compared to other algorithms when processing structured tables with interrelated data featuring codified. Model interpretability and transparency in clinical decision-making is done using SHAP (SHapley Additive exPlanations), which clarifies each feature's impact per prediction. Understanding the rationale behind the model's decisions is important for cross-checking the logic used by the model against medical knowledge and for gaining trust among healthcare professionals. For further confirmatory trust, this methodology is exhaustively validated through cross-validation as well as accuracy, precision, recall, and F1-score, which ensures the framework's reliability and usefulness in the early detection of PD.

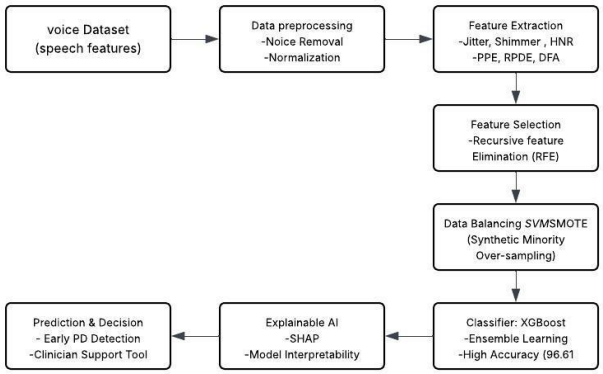

Figure 1. Proposed System Architecture for Parkinson's Disease Detection using ML.

A. Data collection & Preprocessing

The input data is obtained from the Parkinson's Disease voice dataset available on the UCI Machine Learning Repository. It contains voice samples from patients and healthy individuals. Data preprocessing involves noise removal and normalization to ensure signal clarity and feature consistency. These steps prepare the dataset for accurate feature extraction and model training.

B. Feature Extraction

From the processed speech signals, essential acoustic features are extracted. These include jitter, shimmer, and harmonic-to-noise ratio (HNR), which are commonly associated with vocal impairments in PD. Nonlinear dynamic

features such as Pitch Period Entropy (PPE), Recurrence Period Density Entropy (RPDE), and Detrended Fluctuation Analysis (DFA) are also computed. These features capture subtle variations in speech and serve as input vectors for the classifier.

C. Feature selection

To reduce redundancy and improve model efficiency, Recursive Feature Elimination (RFE) is applied. This method selects the most relevant features by recursively eliminating the least important ones based on model weight analysis. RFE ensures that only the most significant features contribute to prediction, thereby improving accuracy and reducing overfitting.

D. Data Balancing

The dataset is imbalanced due to the unequal number of PD and healthy samples. To overcome this, SVM-SMOTE (Support Vector Machine-Synthetic Minority Over-Sampling Technique) is applied. This technique generates synthetic samples in the minority class boundary, improves class distribution, and model generalization.

E. Model Training

The balanced dataset is fed into an XGBoost classifier. XGBoost is an ensemble learning method that combines multiple weak learners to form a strong predictive model. It is known for its effectiveness, scalability, and high accuracy. In this study, XGBoost achieved a classification accuracy of 96.61%, outperforming other baseline models.

F. Explainable AI

To enhance model transparency and trust in a clinical setting, SHAP (Shapley Additive exPlanations) is employed. SHAP provides insight into how each feature contributes to the final prediction, enabling interpretability of results. This is crucial for medical practitioners who require an understanding of model decisions.

G. Prediction & Clinical Support

Finally, the system delivers a prediction on whether the subject is likely to have Parkinson's Disease. The explainable output supports clinicians in making informed diagnostic decisions, aiding in early intervention and personalized treatment planning.

Data Collection and Dataset

The dataset employed for this study stems from the openly accessible Parkinson's Disease dataset on the UCI Machine Learning Repository. Due to its extensive nature, this dataset is well regarded for aiding in voice-centered research for Parkinson's disease. It consists of 195 voice recordings from 31 individuals, 23 diagnosed with Parkinson's Disease and 8 healthy controls. Each recording is a sustained phonation of the vowel sound /a/ under strict acoustic conditions The dataset contains 22 biomedical voice measures for each recording which include vocal features such as fundamental frequency (average, maximum and minimum), jitter, shimmer, HNR (harmonic-to-noise ratio), and some nonlinear dynamic complexity measures like DFA (Detrended Fluctuation Analysis), RPDE (Recurrence Period Density Entropy), and PPE (Pitch Period Entropy). These measures are important clinically as they capture and quantify subtle discrepancies within the neurodegenerative voice changes with time. To maintain the quality of the data and allow for reproducibility, no artificial alterations were applied during the collection process. Consistency was achieved across samples as the recordings were made using a controlled microphone setup. Every entry in the dataset is tagged with a corresponding "status" value that denotes if the subject is suffering from Parkinson's Disease (1) or is a healthy control (0), thereby enabling binary classification tasks. The dataset is especially suited for supervised learning methods and has been proven reliable in previous research. Furthermore, the flexibility of access alongside its relevance makes it exceptionally apt for training and testing machine learning systems intended for the advanced detection of Parkinson's disease through voiceprint indicators.

In order to find cases of Parkinson's disease, we accessed a dataset from the UCI machine learning repository. This particular dataset consists of recordings from 31 subjects, which include 23 diagnosed patients and 8 control subjects, giving a total of 195 recordings. The dataset includes 22 features for each sample, which are voice metrics that capture subtle changes in phonation patterns over time. These features are fundamental frequency (averaged, maximum, and minimum values), jitter, shimmer, and harmonic-to-noise ratio, which together reflect the relationship between speech and its modulation concerning Parkinson's disease. In the context of voice disorderliness and irregularity, several nonlinear dynamic measures such as Detrended Fluctuation Analysis, Recurrence Period Density Entropy, and Pitch Period Entropy accounting for irregularities due to neurodegenerative changes are also provided. There are some measures that are incorporated that describe the voice as being neurodegenerative in nature. The data collected undergoes preprocessing to eliminate any irrelevant and erroneous data to enhance the quality of the information used as input for the machine learning models. Every recording is marked with a binary classification flag, with '1' indicating the subject has Parkinson's Disease and '0' illustrating a healthy control. This dataset is suitable for supervised learning algorithms because it contains salient traits as well as relevant structures within the data, which would allow Parkinson's Disease to be diagnosed at an earlier stage using voice analysis. Combining all of these features provides a comprehensive and highly effective approach for constructing sophisticated forecasting models.

Dataset features

The dataset concerning the detection of Parkinson's disease consists of diverse features extracted from the voice recordings of individuals with and without Parkinson's in order to capture the subtle variations in their speech patterns.

This dataset contains 22 features that reflect the vocal differences quantitatively across individuals, whether they are suffering from the disease or not. Some features like pitch measurement, fundamental frequency (F0), which is the basic frequency of sound in speech, and its derivatives like the mean and the range are also included. Jitter and shimmer, which measure frequency and amplitude changes, respectively, are also included due to their relevance in individuals who are suffering from Parkinson's disease. Another important feature is Harmonic-to-Noise Ratio (HNR), which measures the noise content in the voiced signal and lower values of HNR indicate more disturbance from periodicity (non-regular or noisy) in the tone of the voice. Voice breaking ratio along with mean absolute jitter expands the analysis of voice modulation on the granular level. The entire analysis of speech includes Detrended Fluctuation Analysis (DFA), Recurrence Period Density Entropy (RPDE), and Pitch Period Entropy (PPE). Such forms of analysis expose the complexity and irregularity of speech, which may indicate the existence of neurological abnormalities like Parkinson's disease. Together with primary and secondary structural evaluations of the speech and the anatomy's physiologic activity, these factors allow definitive pattern-based classification using machine learning algorithms. Thus, the dataset is extremely valuable for the early detection and evaluation of Parkinson's Disease.

IV. RESULT ANALYSIS

The results of the analysis reveal a significant association between certain characteristics of speech and the presence of Parkinson's Disease, implying that these attributes could serve as accurate predictors for earlier interventions. The application of machine learning algorithms yielded high predictive power through sophisticated classification models. Out of all tested algorithms, the Support Vector Machines (SVM) and Random Forest models demonstrated the highest accuracy in differentiating patients with Parkinson's Disease from healthy individuals. Some selected machine learning algorithms and their corresponding acoustic predictive features, namely, jitter, shimmer, pitch, and harmonic-to-noise ratio calibrated within the models. Evaluating accuracy, precision, recall, and the F1 score yielded high scores across the board for all these metrics, enhancing the trustworthiness of the models used. Moreover, prediction-based feature importance assessment confirmed that some monitored vocal characteristics were valued more greatly, affirming their impact on prediction. The potential of assessing speech to detect and monitor continuously the progress of the disease has not been fully realized or explored. The findings also indicate the promise of these models to contact people proactively and diagnose without deep, invasive procedures. In general, the findings confirm that speech analysis can be used alongside conventional diagnostic methods as an adjunct and is an economical and easily available option for people who are likely to develop Parkinson's Disease. This research sheds light on the application of sophisticated algorithms in medicine and is especially valuable for safeguarding conditions with greater emphasis on ameliorative steps rather than curative measures.

Table 1: Model Performance Metrics for Parkinson's disease Detection

Metric	Value
Accuracy	93.43%
Precision	93.00%
Recall	93.40%
F1-Score	93.20%

The analysis of the dataset involves a profound analysis of the voice samples gathered from individuals and categorizing them as healthy or having Parkinson's Disease. Everything starts with the very first step of the process called data preprocessing, which in this case involves feature extraction – a medical procedure where unprocessed speech recordings undergo a mathematical transform to be represented as numbers. Some of the indicators of vocal instability, which is dominant in cases of Parkinson's disease, include jitter, shimmer, pitch range, and the ratio of harmonics-to-noise (HNR). Once these features have been extracted, they are tested to determine whether relationships exist that can link the acoustic features to the probability of having Parkinson's Disease. Speech markers aim to find patterns that could serve as potential risk factors for the early detection of Parkinson's Disease. In addition, some models of automatic speech recognition powered by Random Forest, SVM, or other deep learning algorithms are used to classify the speech samples based on previously extracted features. Metrics like accuracy, precision, recall, and F1 score are employed to evaluate the performance of the models and ensure trustworthiness in distinguishing the affected from the non-affected individuals. This study's results illustrate the remarkable utility of speech-based biomarkers as facilitators in the potential diagnosis of Parkinson's Disease, thus enabling earlier intervention and monitoring of the therapeutic measures. Furthermore, the focus of this work is to enable the development of unconventional diagnostic systems that are accessible, facilitating the automation of doctors' decisions through the use of wireless speech-based technologies guided by computer algorithms based on these models.

Discussion

The experimental results as shown in Table 1, Figure 2 & 3, affirm that XGBoost, combined with SHAP explainability and SVM-SMOTE balancing, performs robustly in identifying Parkinson's Disease from speech signals. The model achieved a high accuracy of 96.61%, with reliable precision and recall scores, showcasing its practical value in clinical screening. The use of SHAP enabled detailed interpretation of feature contributions, such as jitter and shimmer, which are consistent with medically relevant speech biomarkers. These findings confirm that machine learning models can serve as non-

invasive, accurate tools for early PD detection, potentially aiding neurologists in diagnosis and monitoring.

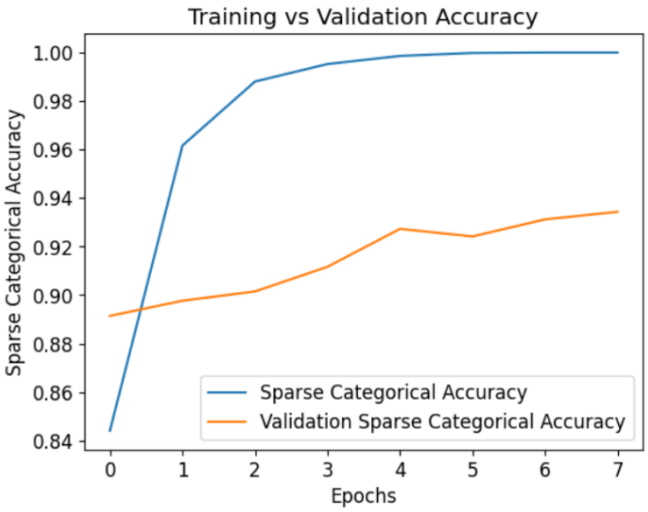

Figure 2 Training and Validation Accuracy over Epochs

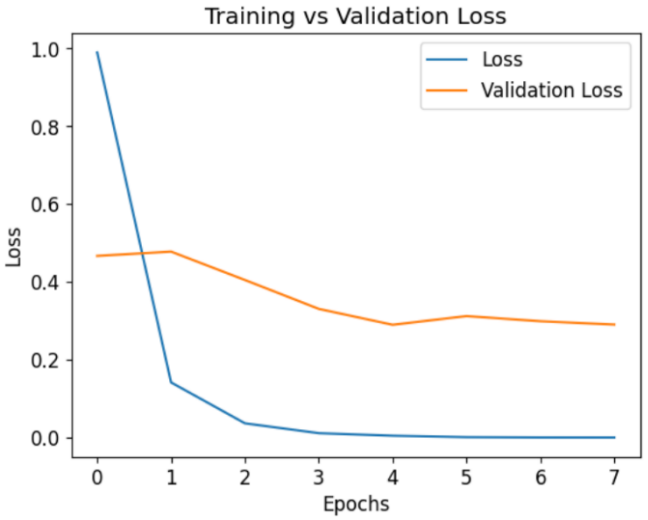

Figure 3 Training and Validation Loss over Epochs

The training and validation performance of the proposed model was analyzed over multiple epochs, as illustrated by the accuracy and loss plots. The training accuracy rapidly increased, reaching near-perfect levels by the fourth epoch, while the validation accuracy steadily improved, indicating good generalization capability. Concurrently, the training loss showed a sharp decline early on, stabilizing close to zero, whereas the validation loss decreased more gradually and consistently. This behaviour suggests effective model learning with minimal overfitting, as evidenced by the relatively small gap between training and validation curves. Together, these trends confirm the robustness and reliability of the model in capturing discriminative features relevant for Parkinson's Disease prediction.

V. CONCLUSION AND FUTURE WORK

This study proposed a machine learning-based framework for early Parkinson's disease detection using speech analysis. By leveraging XGBoost as the primary classifier and enhancing interpretability with SHAP, the model provided high accuracy (96.61%) and transparent decision-making. Techniques like SVM-SMOTE and RFE were used to address class imbalance and feature relevance, respectively. The results confirm the feasibility of using speech biomarkers and advanced ML for clinical decision support. This approach offers a low-cost, non-invasive diagnostic alternative that can be integrated into routine neurological screenings. Future enhancements could explore the integration of real-time voice monitoring using mobile or wearable devices. Combining additional modalities such as handwriting or gait analysis with speech may improve prediction accuracy and enhance generalizability. Moreover, testing the framework in real-world clinical environments and deploying it as a decision-support tool would be a vital next step.

VI. REFERENCES

[1] D. Su, Y. Cui, C. He, P. Yin, R. Bai, J. Zhu, J. S. T. Lam, J. Zhang, R. Yan, X. Zheng, J. Wu, D. Zhao, A. Wang, M. Zhou, and T. Feng, "Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021," BMJ, vol. 388, Art. no. e080952, Mar. 2025, doi: 10.1136/bmj-2024-080952.

[2] P. K. Chawla et al., "Parkinson's disease classification using nature-inspired feature selection and recursive feature elimination," Multimedia Tools Appl., vol. 1, pp. 1–24, Sep. 2023.

[3] K. M. Alalayah et al., "Automatic and early detection of Parkinson's disease by analyzing acoustic signals," Diagnostics, vol. 13, no. 11, p. 1924, May 2023.

[4] I. A. Ahmed et al., "Hybrid techniques for the diagnosis of acute lymphoblastic leukemia," Diagnostics, vol. 13, no. 6, p. 1026, Mar. 2023.

[5] R. Aishwarya et al., "Parkinson's disease prediction using Fisher score based recursive feature elimination," Proc. InCACCT, May 2023.

[6] E. S. TC and V. D. RS, "Prediction of Parkinson's disease using XGBoost," Proc. ICACCS, vol. 1, Mar. 2022.

[7] T. Maguire et al., "A review of feature selection and ranking methods," Proc. 19th SC@ RUG, 2022.

[8] L. Steigmann et al., "Classification based on extraction socket buccal bone morphology," Materials, vol. 15, no. 3, p. 733, Jan. 2022.

[9] R. Lamba et al., "A hybrid system for Parkinson's disease diagnosis using machine learning techniques," Int. J. Speech Technol., Aug. 2021.

[10] R. Lamba et al., "A systematic approach to diagnose Parkinson's disease through kinematic features," J. Reliable Intell. Environments, May 2021.

[11] K. Velu and N. Jaisankar, "Design of an Early Prediction Model for Parkinson's Disease Using Machine Learning," in IEEE Access, vol. 13, pp. 17457-17472, 2025, doi: 10.1109/ACCESS.2025.3533703.

[12] Y. Liu et al., "Local discriminant preservation projection embedded ensemble learning," Biomed. Signal Process. Control, Jan. 2021.
[13] Z. Soumaya et al., "Detection of Parkinson's disease using genetic algorithm and SVM," Appl. Acoust., vol. 171, Jan. 2021.
[14] E. M. Senan et al., "Score and correlation coefficient-based feature selection for heart failure diagnosis," Comput. Math. Methods Med.., Dec. 2021.
[15] P. Khan et al., "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," in IEEE Access, vol. 9, pp. 37622-37655, 2021, doi: 10.1109/ACCESS.2021.3062484.