Parkinson's disease prediction based on analysis of voice segments using ML

Suchithra J
23MCS0056
MTech, CSE, SCOPE
VIT Vellore, TN, India
suchithra.j2023@vitstudent.ac.in

Nikhil Sharma
23MCS0020
MTech, CSE, SCOPE
VIT Vellore, TN, India
nikhil.sharma2023a@vitstudent.ac.in

Bhavesh Kumar Modi
23MCS0045
MTech, CSE, SCOPE
VIT Vellore, TN, India
bhavesh.kumar2023a@vitstudent.ac.in

Dr. JaiSankar N
Professor, SCOPE
VIT Vellore, TN, India
njaisankar@vit.ac.in

Abstract: **Parkinson's disease (PD) is a neurological illness that affects a person's quality of life significantly. So, to the people who are suffering from PD need a prompt intervention and customized treatment. Early diagnosis and ongoing monitoring is essential for PD. In this research, we test the potential of machine learning methods to detect PD from speech characteristics. In addition to Logistic regression, Support Vector Machine, k-Nearest Neighbors, Random Forest, and Decision tree, we implement the Recursive Feature Elimination, because of its Cross-Validation Integration, Performance and Compatibility with different models for acquiring strong results for computation. To train and evaluate our models, we use a dataset consisting of audio recordings of patients and normal people. By employing voice-based machine learning models to enable early diagnosis and ongoing monitoring of Parkinson's disease, the study's findings may have an effect on the healthcare sector.**

Keywords: ***Voice analysis, Machine learning (ML), disease detection, Voice segments, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Tree, Recursive Feature Elimination (RFE), voice recordings, healthcare.***

I. INTRODUCTION

Parkinson's Disease is one of the long-term, degenerative neurological condition that primarily impacts movement. It is caused due to loss of dopamine-producing cells in the brain, resulting in symptoms such as tremor, stiffness, slow movement, balance, and coordination. Additionally, individuals with Parkinson may experience other non-movement-related symptoms, such as depression, anxiousness, difficulty with bowel movements, sleep disorders, alterations in cognitive function, speech impediments, and difficulty swallowing. These symptoms can become more severe with time, and can have a negative impact on daily life, movement, and overall well-being. It ranks as the second most prevalent type of dementia following Alzheimer's disease.

This study aims to investigate how speech of person with PD is examined to make it easier to identify Parkinson's disease at an early stage. Machine learning allows for the analysis and extraction of complex patterns within voice data, which may not be immediately apparent to the human eye. The aim is to create models that are capable of distinguishing between people with Parkinson's and healthy controls, thus allowing for early intervention and enhancing the quality of life of those suffering from the condition.

The importance of this study is reinforced by its emphasis on the integration and optimization of machine learning algorithms, and in particular, the use of RFE as a feature selection technique. RFE is a key factor in increasing the predictive capabilities of the model, as it systematically identifies and eliminates less relevant features, thereby increasing the model's precision, readability, and effectiveness. The study aims to examine the predictive capability of RFE in selecting the most relevant features to identify if someone has Parkinson's disease. The RFE model is compared to independent machine learning models in terms of accuracy, recency and other key performance parameters.

II. Literature review

This study presented a novel approach for Parkinson's disease (PD) detection based on voice classification and feature selection techniques [19]. The authors conducted a comparing the results of different machine learning algorithms for detecting PD based on voice disorders analysis. This paper used two feature selection methods, maximum relevance minimum redundancy (mRMR) and ReliefF, to select the most relevant attributes to the target class and provide the feature subset containing less and minimum redundant features as possible. The authors found that the SVM and KNN algorithms performed well in identifying PD based on voice disorders analysis.

Research of this paper aimed to explore the use of traditional and end-to-end classifier architectures to predict Parkinson's Disease (PD) from speech, utilizing glottal features estimated by iterative adaptive inverse filtering and quasi-closed phase glottal inverse filtering methods [17]. The paper also delves into the use of deep learning models trained on both raw speech waveforms and raw voice source waveforms. The results of the study show that the end-to-end approach using glottal features outperforms the traditional pipeline approach and previous research on detecting PD from speech.

This research described a smartphone-based system for early detection of Parkinson's disease (PD) through voice analysis. "The system utilizes passively-captured voice samples from routine phone calls to extract voice features. The system employs machine

learning techniques, including multiple- and single-instance learning classifiers, to predict and classify individuals. The study involved a multilingual cohort of 498 subjects, including 392 self-reported healthy controls and 106 PD patients" [1]. The best-performing models achieved high accuracy in classifying PD patients and healthy controls [14].

Parkinson's disease (PD) is classified using the algorithm based on the analysis of isolated words [6]. The algorithm aims to automatically assess speech impairment in PD patients, which can contribute to early diagnosis and remote monitoring of the disease [12].

This paper proposes a solution to the problem of early detection of Parkinson's disease (PD) through voice analysis. The introduction explains that PD is a neurodegenerative disorder that affects dopamine-producing neurons in the brain, and its cause is not well understood [4]. The proposed solution involves supervised learning algorithms to learn the patterns of PD patients associated with their voice. The researchers used large data of voice recordings of PD patients and for each recording 70 features were obtained. They then evaluated four machine learning. Additionally, to improve the performance of the classifiers they evaluated two dimensionality reduction techniques.

The results showed that the Support Vector Machine (SVM) model, using a High Correlation Filter, was the model with the best performance, achieving an accuracy of 88% [4]. Therefore, the proposed solution is based on supervised learning algorithms that can aid in the early detection of Parkinson's disease through voice analysis. Experimental research concludes by presenting the possibility of incorporating voice analysis as a digital biomarker to diagnosis PD [4].

The findings in this research aimed to detect Parkinson's Disease (PD) via voice features. They used a dataset of 195 voice records from 147 PD and 48 healthy controls. Machine learning algorithms, which is mentioned were employed. The dataset was split, and SMOTE addressed imbalances, enhancing performance. Voice could be a faster, cost-effective PD biomarker [19].
In summary, the study employed ML techniques to detect PD using voice features, achieving high accuracy (SVM 95%, MLP 98.31%). This model promises early, cost-effective diagnosis and potential enhancements, benefiting medical students and healthcare professionals.

III. Limitations

Recent research papers on detecting Parkinson's disease with the help of machine learning have some limitations in terms of feature selection methods. Many research papers have used univariate feature selection methods, such as mRMR and ReliefF, which consider each feature separately. This can lead to the selection of redundant features and the exclusion of important features.
Additionally, many research papers have evaluated their feature selection methods on a single dataset, without validating them on external datasets to ensure that they are generalizable to unseen data.

Finally, many research papers have not compared their feature selection methods to other state-of-the-art methods, such as RFE, which is making it difficult to evaluate the effectiveness of the proposed methods.

This research used a univariate feature selection method (SelectKBest) to select the eight best features of the dataset, but they did not explain the rationale or criteria for choosing this number of features [19]. This particular study reffered a correlation-based feature selection method to select the most relevant and non-redundant features for PD prediction, but they did not compare or justify the use of this method with other possible methods, such as wrapper methods, embedded methods, or hybrid methods [6].
Across these papers, the chosen feature selection techniques, such as "SelectKBest, F-MDI, F-PER, F-CORR, mRMR, ReliefF, and correlation-based feature selection" [19], were typically used without substantial justification or comparison to other potential methods.

In addition, there was no statistical significance reporting, leaving the door open for performance variations to be attributed to random fluctuation rather than to feature selection. Finally, the lack of external validation makes the model's generalizability uncertain, potentially leading to overfitting or biases in the data set. Taken together, these limitations highlight the need for more systematic, comparative, and rigorous feature selection and statistical reporting, as well as external validation, in the quest for robust and robust Parkinson's disease detection models.

IV. Description of the Dataset

To measure the effectiveness of the REF, Parkinson dataset from UCI was used. This dataset consists of several speech segments collected from 31 people, 23 of whom have Parkinson's disease. between healthy individuals and those with PD. Using this dataset after preprocessing it is used to measure the performance of different parameters using distinct ML algorithms with feature selection as REF.

V. Methodology

To detect the PD efficiently is one of the most important purposes of this research. Research indicates that Recursive Feature Elimination (RFE) is a reliable choice for addressing the issues mentioned in earlier studies. By following best practices, it enhances the dependability, strength, and effectiveness of our method for identifying Parkinson's disease. By choosing Recursive Feature Elimination (RFE), we not only address the drawbacks of previous ways of picking features but also help progress research in this field.

RFE's capacity to lessen duplication in feature selection is one of its main benefits. The ability of RFE to find the solution to the external validation and generalizability, which was mainly lacking in the earlier studies, is perhaps its most significant asset. RFE can be seamlessly integrated with robust cross-validation techniques, enabling us to evaluate the model's generalizability and its performance when applied to diverse datasets or populations. So, including this REF in this research is the aim to efficiently detect PD. The figure 1 shows the Methodology of this research.

I. Dataset Construction

The research involves selecting and preparing a clean dataset for Parkinson's disease detection technology, including a variety of voice recordings from both Parkinson's disease-diagnosed and non-diseased individuals, to build and validate

the technology.

II. Data preprocessing

To ensure that the data is of high quality and prepared for analysis, the preprocessing stage is essential. First, we'll take care of any missing data points and standardize the format of the data. To ensure that the audio recordings are free of any discrepancies or anomalies, this stage is crucial. In addition, we will apply noise reduction methods to improve the fidelity of the voice recordings, guaranteeing that our models are constructed using accurate and trustworthy data.

III. Feature Selection

Feature selection is a pivotal stage in our methodology. We have chosen Recursive Feature Elimination (RFE) as the preferred method. RFE systematically analyzes the data, ranking features by their relevance and eliminating those that contribute less to the model's predictive power. The result is a streamlined dataset that includes only the most informative and non-redundant features, enhancing the model's precision and efficiency.

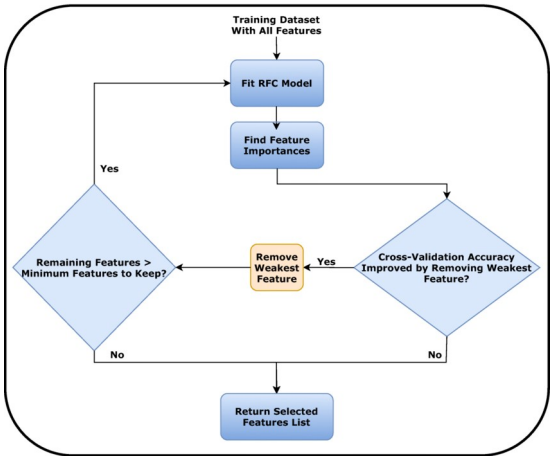

Fig 1 *: Recursive Feature Elimination with Cross-Validation (RFECV)*

IV. SMOTE

Imbalance of classes is a common problem in medical datasets, where one class (e.g., individuals with Parkinson's Disease) may be underrepresented. To address this, we will implement the Synthetic Minority Over-sampling Technique (SMOTE) [30]. This is used to generates data points to the minority class and ensures that our models are not biased towards the majority class. This technique mitigates the risk of an imbalanced dataset negatively impacting the model's accuracy.

\

V. Classification Models

We will employ a selection of machine learning algorithms, each with unique strengths, to construct our predictive models. These models include:

- ***Logistic Regression:*** Logistic Regression is a linear model which designed to predict the probability of a binary outcome. Employing the logistic function, it transforms linear combinations of input features into a range between 0 and 1. The resulting probability is utilized for making predictions, and a specified threshold determines the assignment of the class.
- ***Support Vector Machine (SVM)***: The Support Vector Machine (SVM) is a flexible algorithm suitable for both linear and non-linear classification tasks. Its operation involves identifying a hyperplane that optimally divides classes within a feature space of high dimensions. SVM excels in managing intricate decision boundaries and proves effective in situations where there is a distinct margin between classes.
- ***K-Nearest Neighbor (KNN):*** KNN, short for k-nearest neighbors, is a non-parametric learning algorithm that operates on instances. It assigns a class to a data point by considering the majority class of its closest neighbors in the feature space. The fundamental concept behind KNN is rooted in the notion that data points exhibiting similarities in proximity are probably part of the same class.
- ***Random Forest:*** Random Forest is an ensemble learning algorithm that creates multiple decision trees during training, each trained on a subset of the data. The predictions of these trees are then aggregated through voting or averaging, leading to improved model accuracy and generalization.
- ***Decision Tree (DT):*** Decision Trees consist of hierarchical structures that iteratively divide data according to feature values. At each node, a decision is taken, forming a tree-shaped arrangement. These trees are user-friendly, easily understandable, and proficient in capturing intricate relationships within the data.

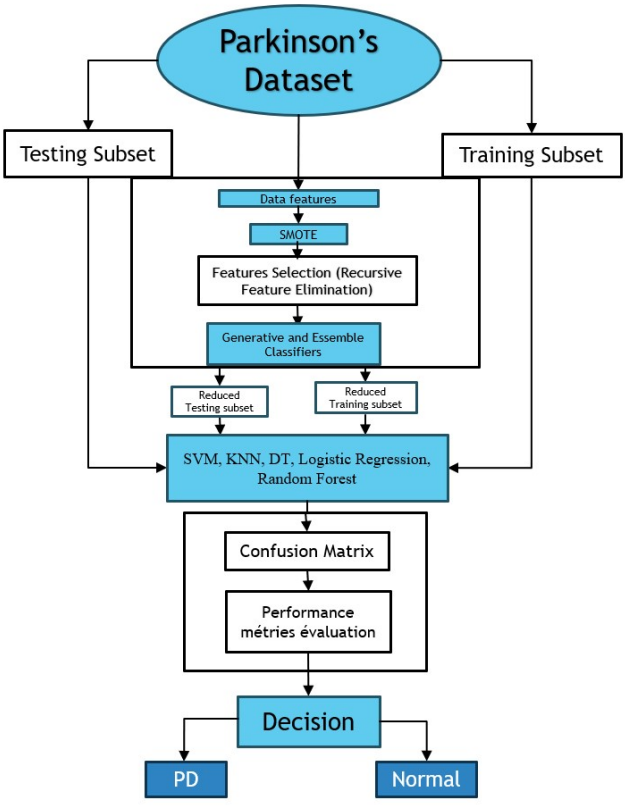

Fig 2 *: The flowchart of the global steps of the adopted model*

VI. Enhanced Classification

The improved classification strategy for detecting Parkinson's disease integrates a diverse array of models, including

traditional classifiers like Logistic Regression, decision-focused models, proximity-based K-Nearest Neighbor, robust Support Vector Machines, Bagging, AdaBoost, and a Voting Classifier. Additionally, Gaussian Mixture Models are employed for a thorough analysis. This fusion aims to enhance accuracy and handle dataset complexities, providing a more nuanced and dependable system for medical diagnosis. The evaluation of system performance includes comprehensive metrics to rigorously assess its effectiveness.

- **Bagging (Ensemble):** Bagging is an ensemble technique that creates multiple instances of a base estimator (e.g., Decision Trees) trained on different portions of data after it is trained. The predictions of individual methods are then aggregated to reduce overfitting and improve model stability.

- **AdaBoost (Ensemble):** AdaBoost is an adaptive boosting algorithm that combines weak learners into a strong classifier. It assigns weights to data points, with higher weights given to misclassified points. Sequential training of weak learners focuses on improving the model's performance on previously misclassified instances.

- **Voting Classifier (Ensemble):** Voting Classifier combines predictions from multiple models, either through 'hard' voting (selecting the majority class label) or 'soft' voting (weighted average of predicted probabilities). It's a versatile ensemble technique that can include different types of classifiers.

- **Gaussian Mixture Model (GMM):** GMM is a probabilistic model that represents a mixture of Gaussian distributions. In the context of clustering, GMM identifies clusters by fitting Gaussian distributions to the data. It provides insights into the underlying structure of the data, capturing both the mean and covariance.

VII. Performance Evaluation

Our comprehensive performance evaluation will encompass key metrics including accuracy, sensitivity, recency, and AUC. These metrics will be rigorously applied to assess the efficacy to predict the Parkinson's Disease.
Accuracy will provide an overall measure of the model's correctness, sensitivity will evaluate its capability to exactly identify patients affected with Parkinson's Disease, specificity will assess its precision in recognizing non-Parkinson's individuals. This thorough evaluation ensures that our system is not only accurate but also reliable, making a significant contribution to the field of medical diagnosis.

Fig 3*: Performance measure using Logistic Regression:*

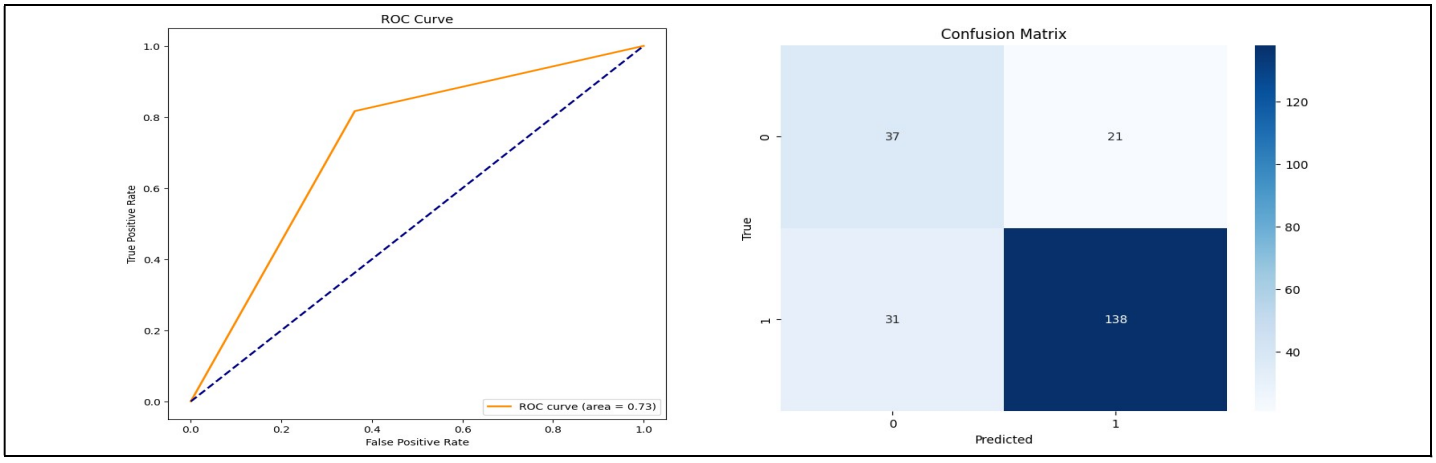

Fig 4: ***Performance measure using Support Vector Machine***

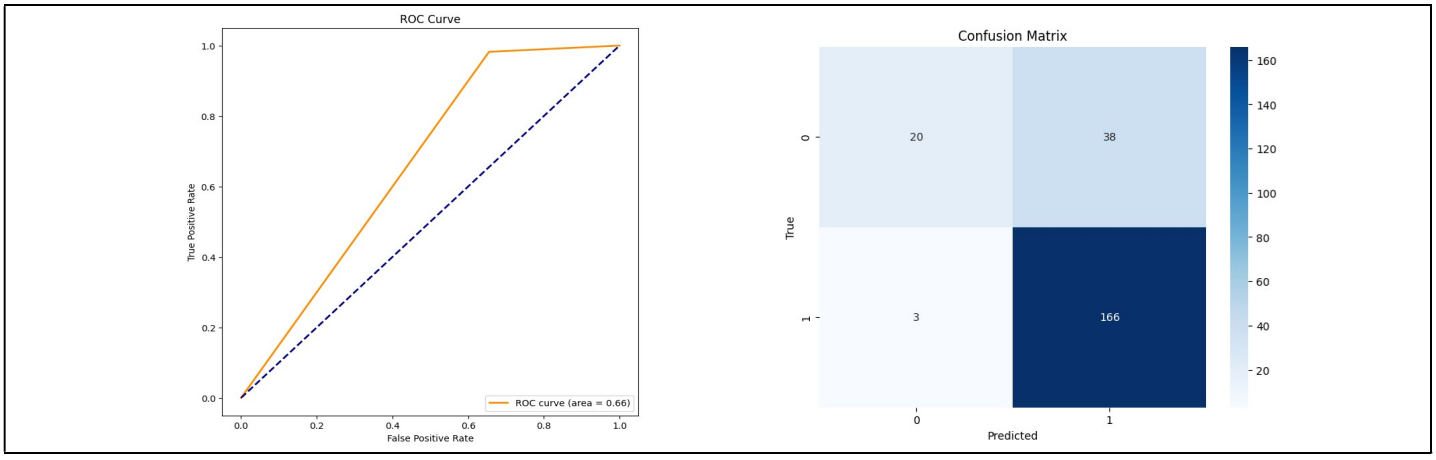

Fig 5: ***Performance measure using K-Nearest Neighbor***

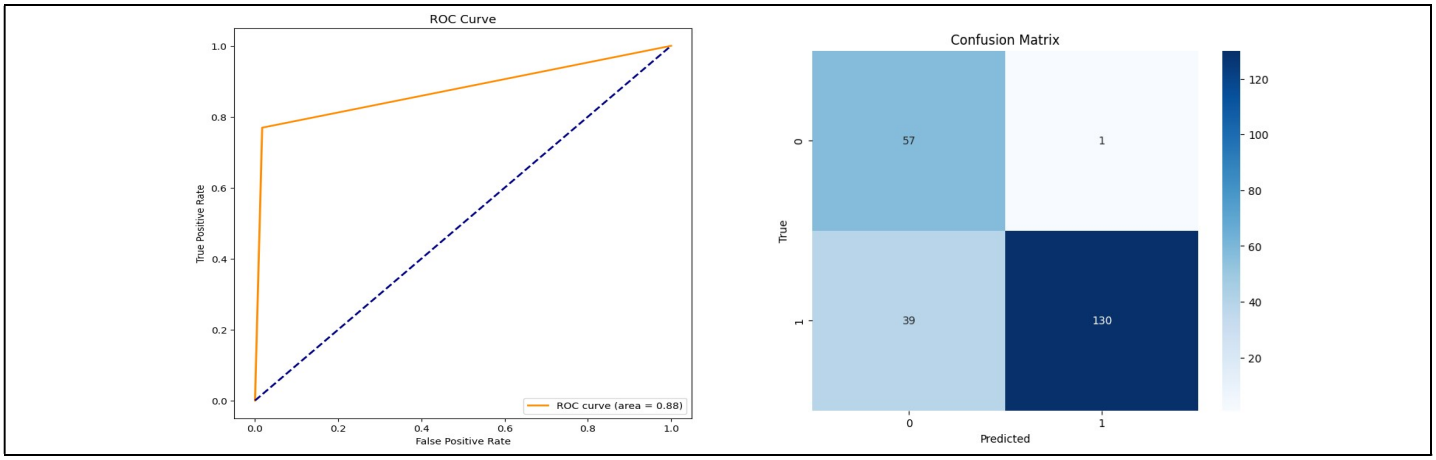

Fig 6: ***Performance measure using Random Forest***

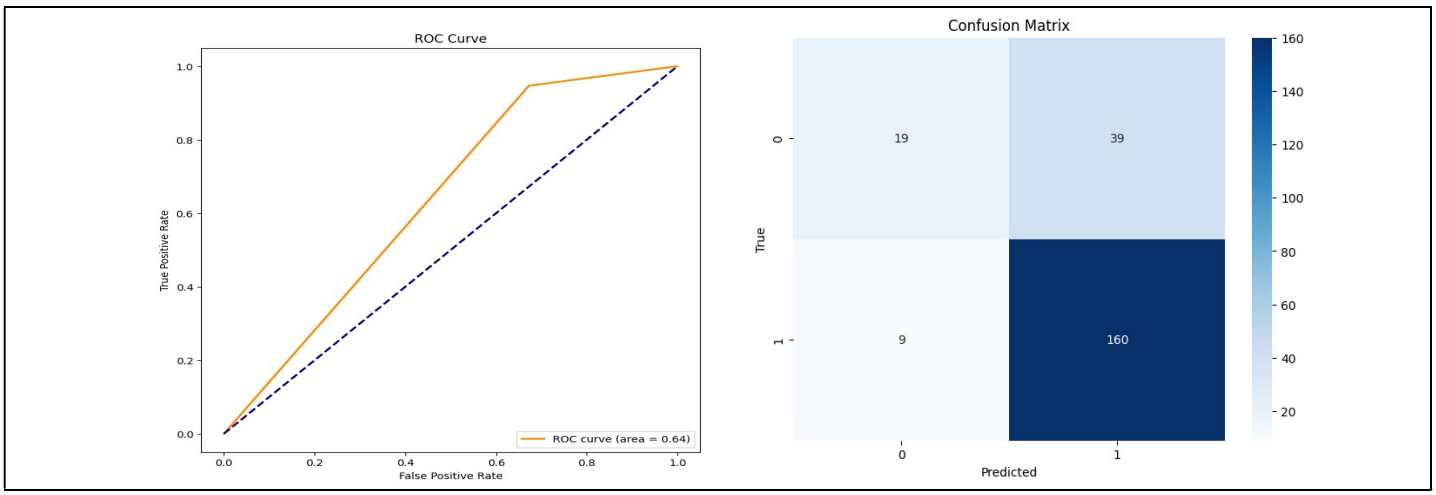

Fig 7: ***Performance measure using Decision Tree***

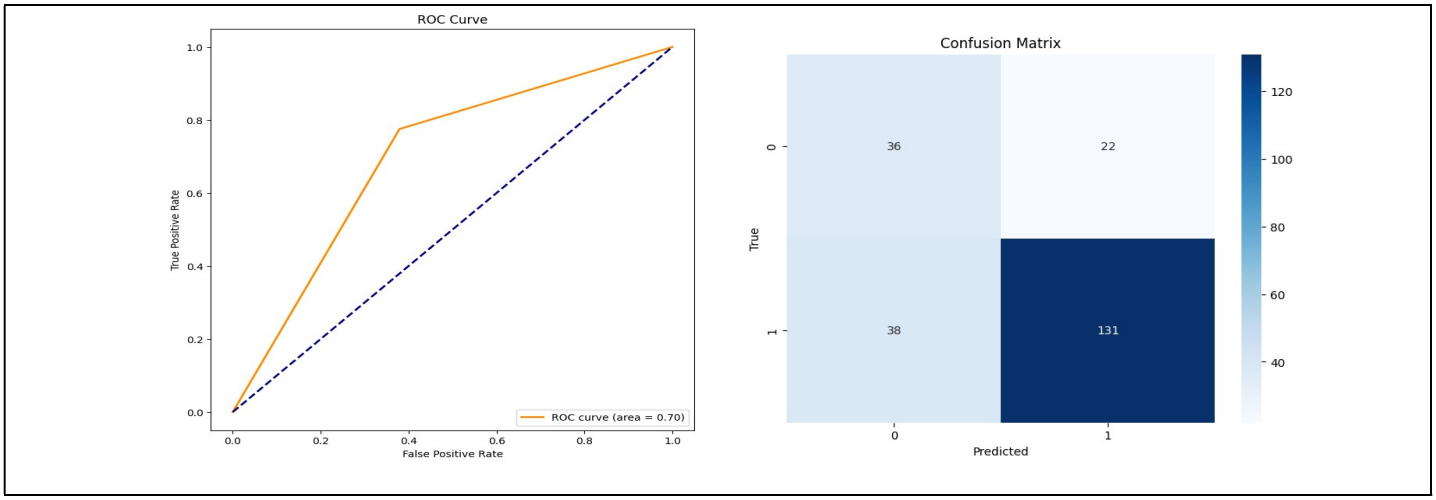

Table 1: ***Comparison of Different ML models performance in predicting PD***

Classifier	Accuracy	Sensitivity	Specificity	AUC
Random Forest	94.4%	94.7%	32.8%	63.7%
Logistic Regression	88.6%	81.7%	63.8%	72.7%
k-Nearest Neighbor	83.9%	76.9%	98.3%	87.6%
Decision Tree	79.9%	77.5%	62.1%	69.8%
Support Vector Machine	73.4%	98.2%	34.5%	66.4%

VI. Conclusion

The evaluation of various classifiers for Parkinson's disease detection highlights distinctive strengths and weaknesses among the models. Random Forest achieves the highest accuracy at 94.4%, closely trailed by k-Nearest Neighbors at 83.9%. However, a comprehensive assessment should consider additional metrics.

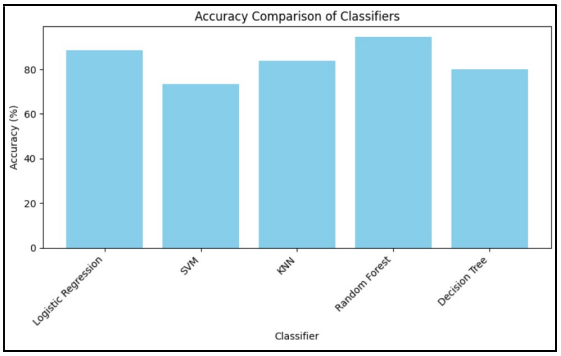

Fig 8: ***Accuracy comparison***

In terms of Sensitivity, denoting the capability to accurately detect the patient with PD, Support Vector Machine excels with an impressive 98.2%, while Random Forest and k-Nearest Neighbors also display notable Sensitivity values of 94.7% and 76.9%, respectively.

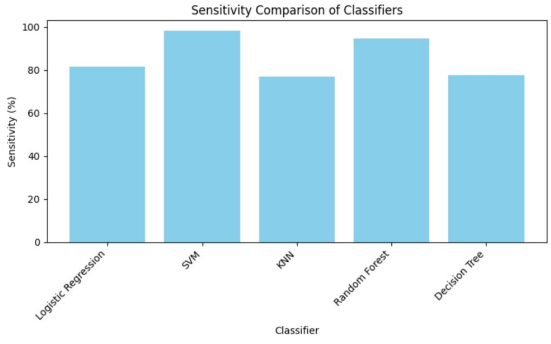

Fig 9: ***Sensitivity comparison***

For Specificity, indicating precision in recognizing non-Parkinson's individuals, k-Nearest Neighbors leads at 98.3%, whereas Support Vector Machine shows lower Specificity at 34.5%.

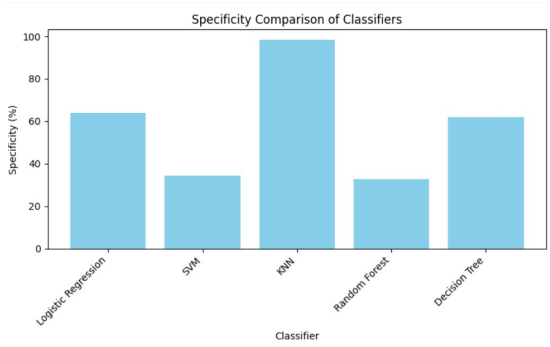

Fig 10: ***Specificity comparison***

In the AUC metric, considering overall classifier performance, k-Nearest Neighbors takes the lead at 87.6%.

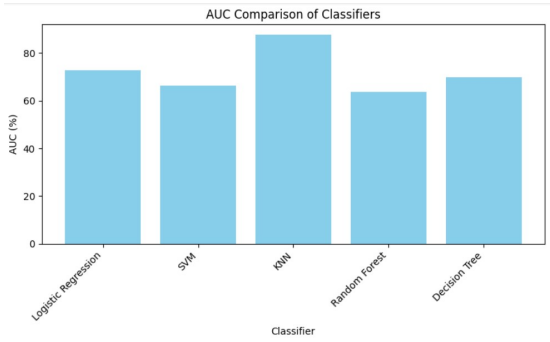

Fig 11: ***AUC comparison***

Random Forest stands out for high accuracy, while k-Nearest Neighbors excels in Sensitivity and Specificity. The optimal classifier choice depends on specific goals, balancing the importance of correctly identifying cases (Sensitivity) with precision in classification (Specificity). Further refinement may be considered to tailor model selection based on specific clinical requirements and trade-offs.

In conclusion, Recursive Feature Elimination (RFE) emerges as a well-founded choice for addressing the limitations identified in prior research, aligning with best practices, and enhancing the credibility, robustness, and generalizability of our Parkinson's disease detection model. By opting for RFE, we not only mitigate the shortcomings of previous feature selection methods but also contribute to the advancement of research in this domain.

VII. Future Scope

The evaluation of classifiers for Parkinson's disease detection reveals promising avenues for future research as well. Exploring the optimization of ensemble methods like Random Forest and k-Nearest Neighbors through careful fine-tuning and configuration adjustments could unlock their full potential for increased accuracy and resilience. Further refinement of algorithms, particularly in the case of Support Vector Machine, may enhance specificity through continued tuning and exploration of advanced algorithmic variations. A sustained focus on feature engineering remains a pivotal aspect for ongoing investigation, offering the opportunity to introduce new relevant features or derive innovative ones from existing data. These efforts have the possibility not only to improve the interpretability of models but also to enhance overall performance in the field of detecting Parkinson's disease.

VIII. References

[1] C. Laganas, “Parkinson’s disease detection based on running speech data from phone calls,” *IEEE Transactions on Biomedical Engineering,* vol. 69, no. 5, pp. 1573-1584, 2021.

[2] S. Rahman, “Classification of Parkinson’s Disease using Speech Signal with Machine Learning and Deep Learning Approaches.,” *European Journal of Electrical Engineering and Computer Science,* vol. 7, no. 2, pp. 20-27, 2023.

[3] K. A. Shastry, “An ensemble nearest neighbor boosting technique for prediction of Parkinson’s disease.,” *Healthcare Analytics,* vol. 3, 2023.

[4] Y. C. Tai, “A voice analysis approach for recognizing Parkinson’s disease patterns,” *IFAC-PapersOnLine,* vol. 54, no. 15, pp. 382-387, 2021.

[5] D. Francisco Santos, “Parkinson's Disease Detection using XGBoost and Machine Learning,” vol. 2023, no. 10, 2023.

[6] Amato, “An algorithm for Parkinson’s disease speech classification based on isolated words analysis.,” *Health Information Science and Systems,* vol. 9, pp. 1-15.

[7] Yadav, “Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers,” *National Conference on Computing and Communication Systems. IEEE,* pp. 1 - 8, 2012.

[8] Yaman, “Automated Parkinson’s disease recognition based on statistical pooling method using acoustic features.,” *Medical Hypotheses,* p. 135, 2020.

[9] S. Dabiri, “Inferring transportation modes from GPS trajectories using a convolutional neural network.,” *Transportation research part C: emerging technologies,* vol. 86, pp. 360-371, 2018.

[10] Colliot, “Machine Learning for Brain Disorders,” *Springer Nature,* vol. 197, 2023.

[11] Anter, “A robust intelligence regression model for monitoring Parkinson’s disease based on speech signals.,” *Future Generation Computer Systems,* pp. 316-327, 2023.

[12] Er, “Parkinson’s detection based on combined CNN and LSTM using enhanced speech signals with variational mode decomposition,” *Biomedical Signal Processing and Control,* vol. 70, 2021.

[13] Hassanien, “Proceedings of the International Conference on Advanced Intelligent Systems and Informatics,” *Springer,* vol. 533, 2016.

[14] Mohaghegh, “Identifying Parkinson’s disease using multimodal approach and deep learning,” *6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA). IEEE,* pp. 1 - 6, 2021.

[15] Singh, “Advances in computing and data sciences,” *Proc. of the 3rd International Conference, ICACDS,* 2019.

[16] Govindu, “Early detection of Parkinson's disease using machine learning.,” *Procedia Computer Science,* pp. 249-261, 2023.

[17] Narendra, “The detection of Parkinson's disease from speech using voice source information.,” *IEEE,* pp. 1925-1936, 2021.

[18] Pragadeeswaran, “An Adaptive Intelligent Polar Bear (AIPB) Optimization-Quantized Contempo Neural Network (QCNN) model for Parkinson’s disease diagnosis using speech dataset,” *Biomedical Signal Processing and Control,* vol. 87, 2024.

[19] Alshammri, “Machine learning approaches to identify Parkinson's disease using voice signal features,” *Frontiers in Artificial Intelligence,* 2023.

[20] Ouhmida, “A Novel Approach for Parkinson's Disease Detection Based on Voice Classification and Features Selection Techniques.,” *Int. J. Onl. Eng.,* vol. 17, no. 10, p. 111, 2012.

[21] P. M. Granitto, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemometrics and Intelligent Laboratory Systems,* vol. 83, no. 2, pp. 83-90, 2006.

[22] X. -w. Chen, “Enhanced recursive feature elimination,” *Cincinnati,* vol. 2007, no. 35, pp. 429-435, 2007.

[23] X. Lin, “A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information,” *Journal of Chromatography B,* vol. 910, pp. 149-155, 2012.

[24] A. M, “Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems,” *Journal of Sensor and Actuator Networks.,* vol. 12, no. 5, p. 67, 2023.

[25] C. Y. Freyte, “Recursive Feature Elimination with Cross Validation for Alzheimer’s Disease Classification using Cognitive Exam Scores,” *2023 Intelligent Methods, Systems, and Applications (IMSA),* pp. 327-332, 2023.

[26] J. H, “Hybrid-Recursive Feature Elimination for Efficient Feature Selection,” *Applied Sciences,* vol. 10, no. 9, p. 3211, 2020.

[27] P. M. Granitto, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemometrics and intelligent laboratory systems,* vol. 83, no. 2, pp. 83-90, 2006.

[28] L. V. Kalia, “Parkinson's disease,” *The Lancet,* vol. 386, no. 9996, pp. 896-912, 2015.

[29] A. Keller, “SMOTE and ENN based XGBoost prediction model for Parkinson’s disease detection,” *International Conference on Smart Electronics and Communication (ICOSEC), IEEE,* pp. 839-846, 2021.

[30] B. N. Nakkaş, “Feature Selection and SMOTE Based Recommendation for Parkinson's Imbalanced Dataset Prediction Problem,” *IEEE,* pp. 1-4, 2022.

[31] B. Thomas, “ Parkinson's disease,” *Human molecular genetics,* vol. 16, no. R2, pp. R183-R194, 2007.

[32] S. Shetty, “SVM based machine learning approach to identify Parkinson's disease using gait analysis,” *International conference on inventive computation technologies (ICICT), IEEE,* vol. 2, pp. 1-5, 2016.

[33] I. Bhattacharya, “SVM classification to distinguish Parkinson disease patients,” *In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India,* pp. 1 - 6, 2010.