Integrative Healthcare System AI-Driven Disease and Patient Diagnosis System

1st Abhinav
Dept. of Computer Science and Engineering
KIET Group of Institutions,
Ghaziabad, India

2nd Ayush Sachan
Dept. of Computer Science and Engineering
KIET Group of Institutions,
Ghaziabad, India

3rd Andril Omer
Dept. of Computer Science and Engineering
KIET Group of Institutions,
Ghaziabad, India

4th Omprakash Kushwaha (Asst. Prof.)
Dept. of Computer Science and Engineering
KIET Group of Institutions,
Ghaziabad, India

***Abstract*—Parkinson's Disease (PD) is a degenerative neurological condition affecting motor function and speech, providing diagnostic challenges due to mild early signs. This study explores computational voice analysis for PD detection using a public dataset of acoustic recordings from PD patients and healthy individuals. Key biomarkers like frequency variability, pitch instability (jitter), amplitude fluctuation (shimmer), harmonic distortion (noise to harmonics ratio), and nonlinear complexity indices are analyzed to identify pathological voice patterns. Seven supervised algorithms, including ensemble tree based methods (Random Forest, XGBoost), probabilistic classifiers(Naive Bayes), and kernel based models (SVM), are evaluated for diagnostic reliability. Performance metrics such as accuracy, sensitivity, F1 score, and ROC AUC are used to optimize early detection. Results highlight ensemble methods as robust solutions for imbalanced voice data. The research emphasizes the potential of voice based machine learning tools as non invasive screening aids in remote healthcare, enabling timely interventions to mitigate disease progression.**

Keywords: Parkinson's Disease, K-Nearest Neighbour, Support Vector Machines, Convolutional Neural Networks, Recurrent Neural Networks, Noise to Harmonics Ratio

I. Introduction

Parkinson's disease (PD) is one example of a neurodegenerative condition that evolves due to the gradual decay of dopamine producing neurons in the **substantia nigra**, which is important for motor control is formed. This type of neuron destruction shows itself through physical and non physical manifestations such as tremors, bradykinesia, muscle stiffness, balance problems, alongside speech issues, cognitive decline, mood disorders and depression. Mitigation of the ailment's progression heavily depends on early diagnosis, however, it is made difficult due to the diagnostic procedures that relies solely on the clinician's rudimentary evaluation checklist.

There is an urgent call for standardized diagnosis tools: objective and devoid of the clinician's discretion bias [6]. Dysphonia is an overlooked yet promising area for PD detection, often occurring before motor symptoms present themselves. Even at these early stages, patients soften their speech, become less articulate, and speak in a monotone which leads to more pronounced cases of breathiness [15].Even if these changes are easy to overlook, the new potential created by computational analytics permits measuring the assessment of voice patterns for early diagnosis [13]. Important features of the voice signal like variations of the Fo as well as non-uniform vocal event occurrences (jitter, shimmer) make reaching this goal achievable .

Further strengthen the armament for predictive analytics with probabilistic techniques including Naive Bayes, alongside proximity based K-Nearest Neighbours (KNN) and gradient boosted XGBoost models [15]. Performance metrics including accuracy, precision, recall, F1 score, and ROC AUC are employed to objectively assess these models so assuring a thorough evaluation of diagnostic dependability. This study evaluates the effectiveness of voice biomarkers for PD identification through a comparative analysis of machine learning algorithms [4]. The study aims to develop efficient frameworks for non-invasive, automated diagnosis by training classifier models on acoustic data and evaluating their accuracy [1].

Moreover, it analyses pragmatic problems in applying such systems inside clinical and telemedicine models, thereby maybe allowing remote monitoring and rapid therapies. These findings potentially alter early PD diagnosis, decrease reliance on subjective assessments, and boost acces- sibility to care, particularly in impoverished locations [11]. By bridging data driven innovation considering clinical needs, our work contributes to the rising f ield of digital biomarkers, enabling scalable techniques for neurodegenerative disease management.

II. Related Work

Parkinson's Disease (PD) is a progressive neurodegenerative disorder distinguished by the slow loss of dopaminergic neurons in the substantia nigra, a brain area critical for regulating movement. This neuronal degeneration disrupts motor function, resulting in hallmark symptoms such as resting

tremors (rhythmic shaking in limbs), bradykinesia (slowness of movement), muscle rigidity, and postural instability.These non motor symptoms, such as hypophonia (reduction in vocal volume) and dysprosody (loss of normal rhythm in speech), add to the challenge of diagnosing PD (Parkinson's disease).These emerging vocal impediments, even if subdued in initial stages, are of great importance for early diagnosis. They present a low risk diagnostic option as mentioned in citation [9].

Utilization of machine learning for diagnosing case of PD has grown tremendously, with supervised learning models spearheading the analysis of voice patterns [5]. Early breakthroughs by Little et al.(2009) [2] indicated the efficacy of Support Vector Machines(SVM) in categorising PD patients using dys phonia measures, reaching great accuracy by differentiating patho logical voice patterns. Subsequent research broadened this paradigm: Sakar et al. (2017) [2] underlined the merits of ensemble methods such as Random Forests in comparison to more primitive approaches like single Decision Trees, attributing their robustness to overfitting and generalization capabilities. However, simpler models such as Logistic Regression and Naive Bayes, as studied by Chen et al. (2020) [7], encountered challenges in modeling sophisticated relationships in multi-dimensional sound recordings and revealed the need for refined computation approaches.

Model performance has been reported to benefit from feature engineering [10]. PCA identifies distinguishing features of a voice and simplifies datasets which enables focus to be placed on more relevant aspects of the signal like background noise (jitter and shimmer) being supressed. Nonlinear analysis, like entropy and fractal dimension analysis has identified chaotic speech patterns associated with Parkinson's disease and has shed more light on the concept of illness specific dysphonia. [12].

Recent innovations in deep learning have further revolutionized PD detection. For instance, Xie et al.(2020) harnessed Convolutional Neural Networks(CNNs) to automatically extract spectral features from voice recordings, coupled with Recurrent Neural Networks(RNNs) to model temporal dependencies, outperforming conventional models in cross-validation studies. Hybrid frameworks, such as SVM paired with genetic algorithms for feature selection, have also demonstrated enhanced accuracy by prioritizing biomarkers like noise to harmonics ratio(NHR) and fundamental frequency variation. Despite these advancements, the field lacks consensus on optimal model selection and feature relevance.

Building on these foundations, this study adopts a systematic framework to evaluate seven machine learning algorithms Decision Trees, Random Forests, Logistic Regression, SVM, Naive Bayes, K-Nearest Neighbors(KNN), and XGBoost, while dissecting the contribution of individual vocal features to classification outcomes [5]. By synthesizing insights from accuracy, precision, F1 score, and ROC AUC metrics, the research addresses methodological gaps in existing literature [8]. Notably, it investigates the clinical viability of deploying these models in telemedicine platforms, where rapid, voice-based screening could democratize access to early diagnosis. This methodological gap specifically, the absence of standardized benchmarks for PD voice analysis motivates the study's comparative approach.

III. Methodology

A. *Dataset Description*

The UCI Parkinson's dataset consists of 195 instances with 23 attributes derived from voice recordings. These attributes include:

- Fundamental Frequency Features – MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz)
- Jitter (Frequency Variation) – MDVP:Jitter(%), MDVP:RAP, MDVP:PPQ
- Shimmer (Amplitude Variation) – MDVP:Shimmer, Shimmer:APQ3, MDVP:APQ
- Noise-to-Harmonics Ratio – NHR, HNR
- Nonlinear Dynamical Complexity Measures – RPDE, DFA, spread1, spread2, D2, PPE
- Target Variable – status (1 = Parkinson's, 0 = Healthy)

article graphicx booktabs [a4paper,margin=1in]geometry

Feature	**Description**
Subject Identifier	Unique label combining the subject's ID and session number, used for tracking individuals.
Maximum Pitch (Fo)	The highest pitch reached during vocal activity, showing how much the voice can vary.
Jitter	Captures small, rapid changes in pitch from one vibration cycle to the next (pitch instability).
Shimmer	Represents amplitude fluctuations between consecutive voice cycles (volume instability).
Noise to Harmonics Ratio (NHR)	Measures the ratio of noise components to harmonic components in the voice, indicating clarity or distortion.
Complexity Measures	Looks at irregular vocal signal patterns using entropy and nonlinear metrics to detect abnormalities.
Spread (1 & 2)	Quantifies how much the pitch deviates from the average (e.g., Spread1 for overall variation, Spread2 for local fluctuation).

This dataset provides rich voice-based biomarkers, making it suitable for computational PD diagnosis.

B. *Data Preprocessing*

To ensure robust data quality and model efficiency, we implement a systematic workflow. First, we address incomplete data by filling gaps to preserve dataset integrity. Next, we normalize all features to a 0–1 scale, eliminating bias from variable magnitudes. The data is then split into training 80% and testing 20% subsets to validate performance objectively. Finally, we identify and retain only the most impactful features using statistical relevance and iterative elimination, stripping away noise and redundancy. This approach sharpens model accuracy while streamlining computational demands, ensuring practicality for real-world clinical use.

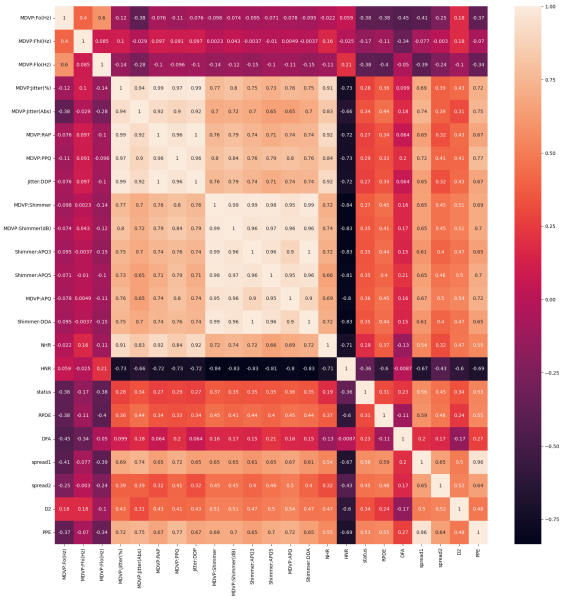

Fig. 1. correlation heatmap

C. Machine Learning Models

To achieve a comparative analysis in Parkinson's Disease detection, we implement seven different machine learning models, each offering distinct classification capabilities. The Decision Tree Classifier follows a hierarchical rule based learning approach, making decisions through a series of logical conditions. The Random Forest Classifier, an ensemble method, combines numerous decision trees to minimize variance and enhance predicted accuracy. Logistic Regression, a prob- abilistic binary classifier, calculates the likelihood of a given input belonging to a specified class. Support Vector Machine optimizes the separation margin between classes, ensuring effective classification even in high-dimensional spaces. The Naive Bayes classifier, a fast and efficient probabilistic model, operates under the assumption of feature independence. The K-Nearest Neighbors algorithm classifies instances based on proximity to neighboring data points, utilizing distance metrics. Lastly, the XGBoost Classifier, an optimized gradient boosting model, enhances performance by minimizing errors through iterative improvements.

To evaluate model performance, various statistical metrics are utilized. Accuracy is calculated as the proportion of correctly classified cases, determined by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP(True Positives) and TN(True Negatives) represent correctly predicted Parkinson's and non-Parkinson's cases, respectively, while FP (False Positives) and FN (False Negatives) denote misclassifications. Precision, defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

measures the proportion of correctly identified Parkinson's cases among all predicted positive cases. Recall (Sensitivity), computed as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

assesses the model's ability to detect actual Parkinson's cases. The F1 score, given by:

$$F1 - score = \mathbf{2} \times \frac{Prescision \times Recall}{Precision + Recall} \tag{4}$$

provides a harmonic mean between precision and recall, balancing the trade off between false positives and false negatives. Finally, the ROC AUC score analyzes classification quality by calculating the area under the receiver operating characteristic curve, demonstrating the model's ability to discriminate between Parkinson's and healthy individuals at varied probability levels.

D. Voice Feature Extraction Process

This work employs an automated approach to diagnose Parkinson's Disease (PD) using speech analysis. The workflow involves three stages:

- **Audio Preprocessing:** A user provides a voice sample, which is cleaned to remove background interference using noise reduction algorithms. The audio is split into short segments to analyze subtle speech changes over time, and open-source audio analysis tools (e.g. Librosa) identify key vocal markers.
- **Feature Computation:** The system calculates metrics linked to PD symptoms, such as pitch variations (average, high, and low ranges), irregular voice vibrations ("jitter"), volume fluctuations ("shimmer"), and clarity (noise to harmonics ratio). Complexity metrics, like entropy-based measures and signal stability indices, are also derived to detect chaotic speech patterns.
- **Model Inference Prediction:** Algorithms that have been trained previously Evaluate the obtained measures to allocate the example to either 'PD likely' or 'healthy'. Best performers (XGBoost for example) concentrate on precision. How well claims are made for dependability, especially with clinical work, determines how useful these claims are for using in actual practice. This also eases the burden of first stage PD screening by turning voice anomalies input to actionable steps devoid of drilling procedures.

This approach streamlines early PD screening by translating voice irregularities into actionable insights without invasive procedures.

IV. Results and Discussion

After evaluating the models, XGBoost and Random Forest consistently delivered the strongest results. Their success stems from their capacity to model complex, nonlinear relationships within voice data, using combined tree based strategies to minimize errors and improve reliability. SVM

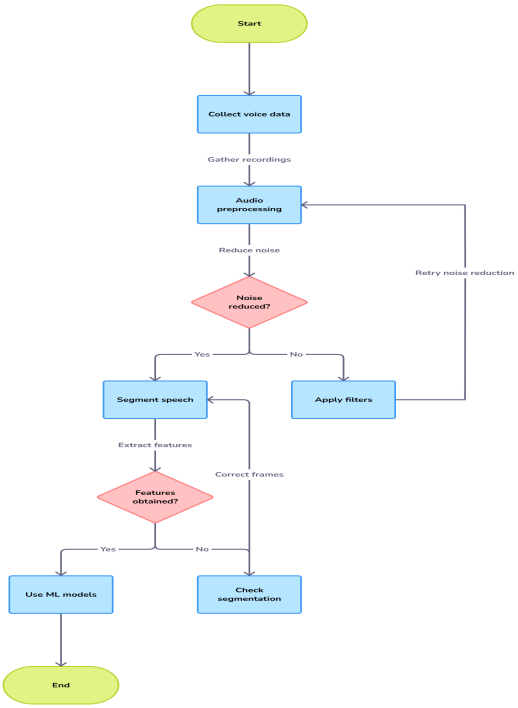

Fig. 2. Performance Metrics Visualization

also performed well, particularly in ambiguous cases, thanks to its ability to define clear decision boundaries between classes. Simpler models like Logistic Regression and Naive Bayes lagged behind, struggling to interpret intricate patterns in vocal features. Naive Bayes faced additional hurdles due to its reliance on oversimplified assumptions about feature independence, which clashed with the interconnected nature of voice biomarkers. KNN's performance suffered further, as noise and high dimensional data amplified its limitations, making it less practical for real world diagnostics. To understand what drives predictions, we examined key vocal markers. Models like XGBoost and Random Forest highlighted frequency related features such as average, maximum, and minimum pitch as critical for detecting vocal tremors, a common PD symptom. Nonlinear measures, including entropy based and signal stability metrics, also stood out for their ability to differentiate patients from healthy individuals. Metrics capturing vocal instability, like irregular pitch and volume fluctuations, further reinforced the diagnostic value of voice analysis. These insights validate voice based tools as practical, non invasive methods for early PD detection, aligning with clinical observations of speech degradation in patients.

A. *Confusion Matrix and ROC Curve*

To evaluate how effectively the models distinguish between Parkinson's and healthy cases, we analyze their predictions using a results breakdown table. This table categorizes outcomes into four groups:(1) correct Parkinson's detections, (2) correct identifications of healthy individuals, (3) healthy cases mistakenly flagged as Parkinson's, and (4) Parkinson's cases overlooked by the model. Examining these categories helps us to find how often the model differs between the two groups, thereby giving priority to its capacity to prevent crucial mistakes. For clinical trust and patient safety, missing actual Parkinson's cases or triggering false alarms is absolutely crucial.

We use a threshold sensitivity graph to depict the decision-making behavior of the model and hence evaluate performance. This graph shows how well the model balances its tendency to misclassify healthy people (false positives) as detection thresholds change against its accuracy in Parkinson's disease (true positives). This balance is quantified by a performance score between 0 and 1; numbers nearer 1 indicate almost perfect accuracy and lowest error.

By integrating these methods, we identify models that excel in real-world circumstances, where eliminating diagnostic errors is crucial. This technique delivers accurate, actionable insights for healthcare applications, where accuracy directly influences treatment decisions and patient well-being.

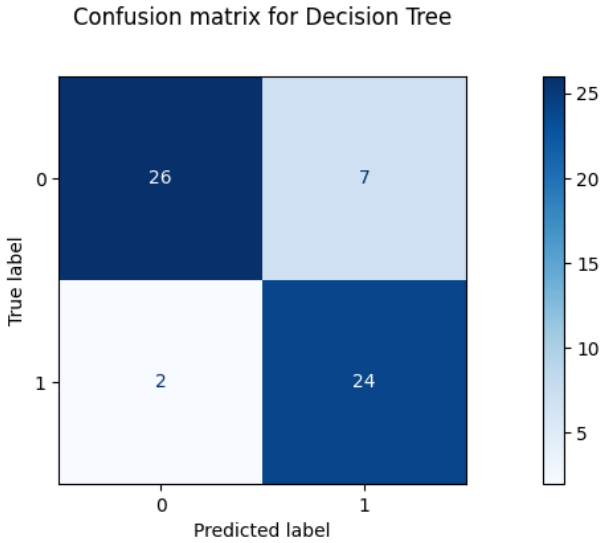

Fig. 3. Decision Tree Confusion Matrix

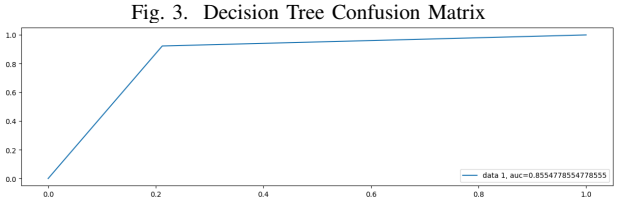

Fig. 4. Decision Tree ROC Curve

V. Final Comparative Analysis

Their incisive ideas and assistance have considerably assisted to the effective completion of this project. We also express our thanks to the researchers and developers of the UCI Parkinson's dataset for making their data publicly available, which played a significant role in our work.

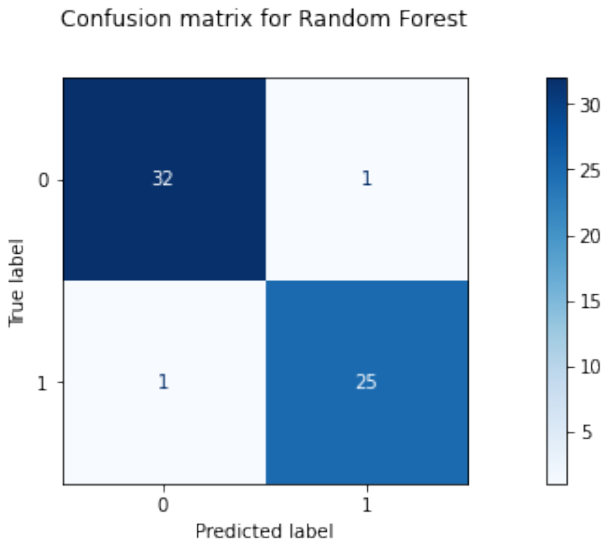

Fig. 5. Random Forest Confusion Matrix

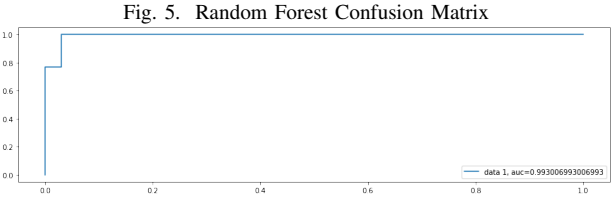

Fig. 6. Random Forest ROC Curve

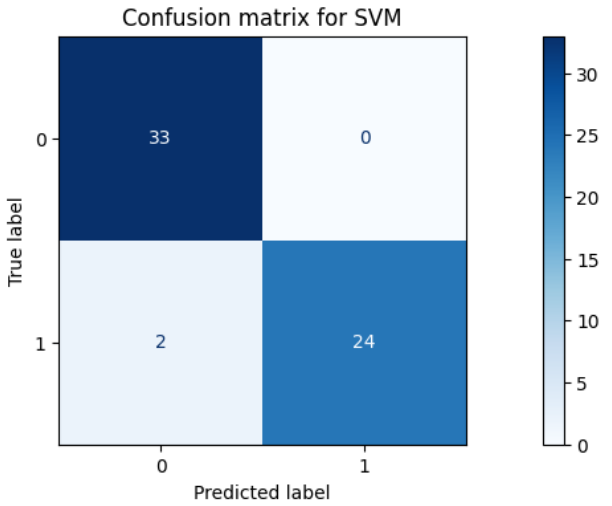

Fig. 9. SVM Confusion Matrix

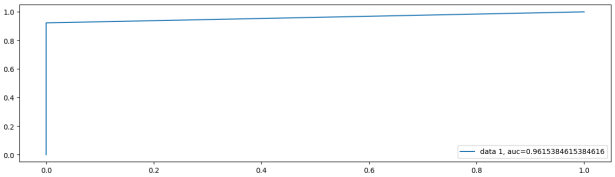

Fig. 10. SVM ROC Curve

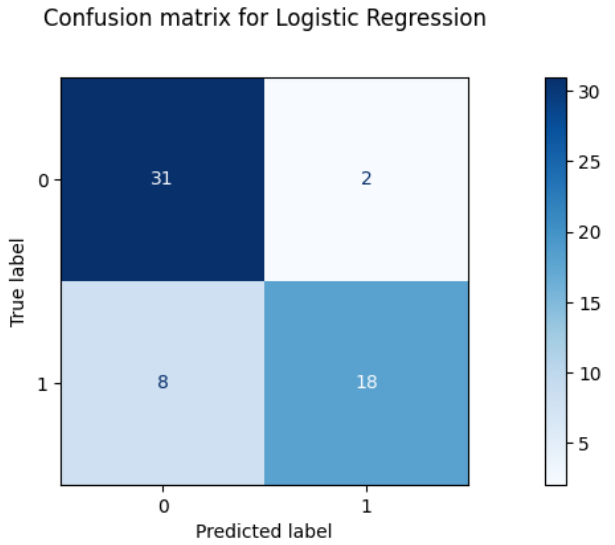

Fig. 7. Logistic Regression Confusion Matrix

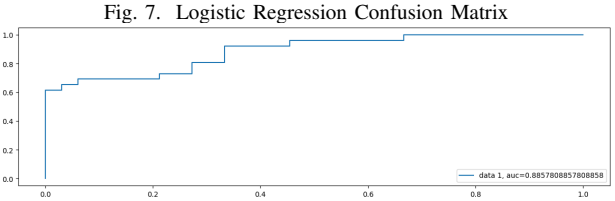

Fig. 8. Logistic Regression ROC Curve

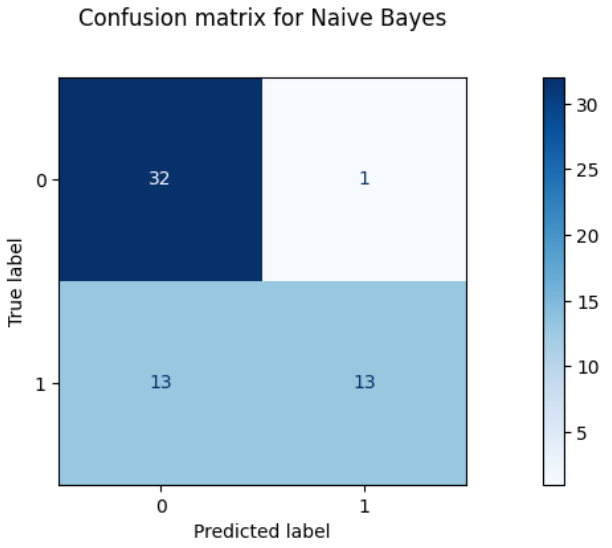

Fig. 11. Naive Bayes Confusion Matrix

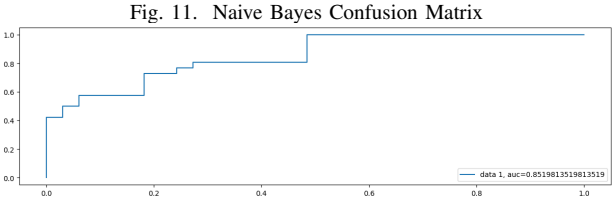

Fig. 12. Naive Bayes ROC Curve

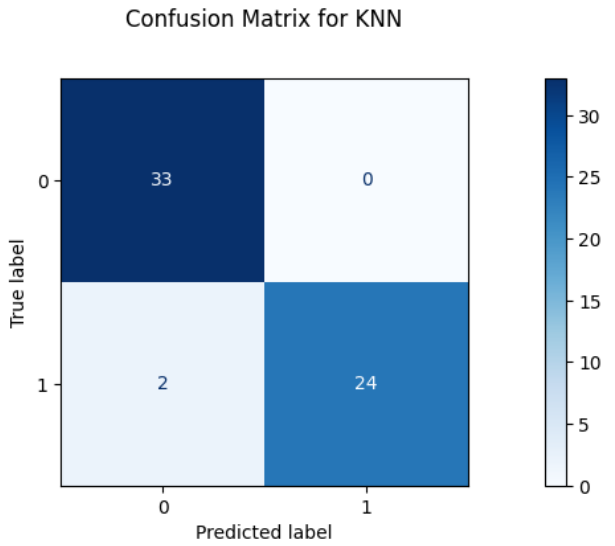

Fig. 13. KNN Confusion Matrix

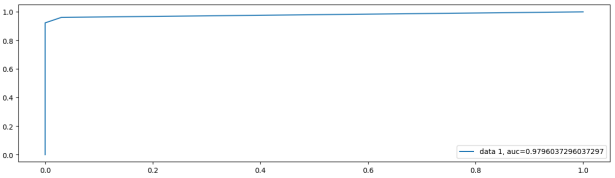

Fig. 14. KNN ROC Curve

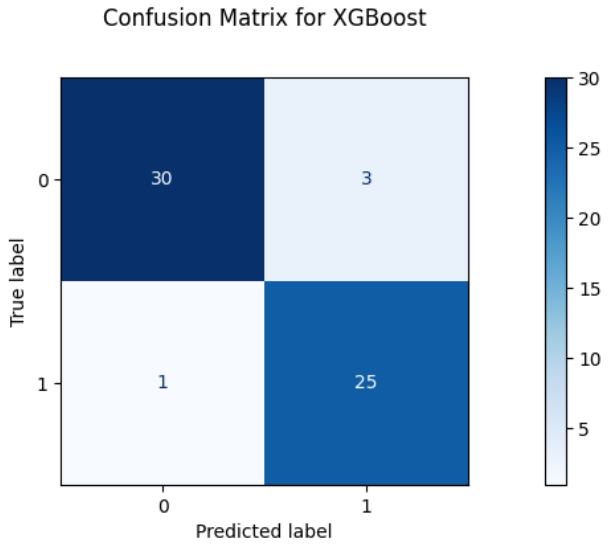

Fig. 15. XGBoost Confusion Matrix

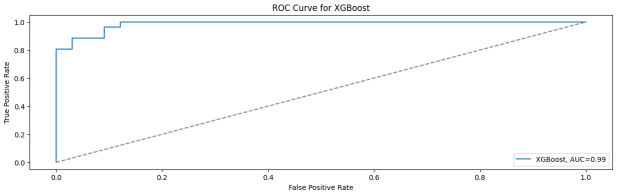

Fig. 16. XGBoost ROC Curve

Metric	DT	RF	LR	SVM	NB	KNN	XGB
Accuracy	0.847	0.966	0.831	0.966	0.763	0.966	0.932
F1-Score	0.842	0.961	0.783	0.960	0.650	0.960	0.926
Recall	0.923	0.961	0.692	0.923	0.500	0.923	0.962
Precision	0.774	0.961	0.900	1.000	0.929	1.000	0.893
R2-Score	0.381	0.862	0.312	0.862	0.037	0.862	0.725

TABLE I
PERFORMANCE METRICS FOR DIFFERENT MACHINE LEARNING MODELS

VI. CONCLUSION AND FUTURE WORK

This study focuses on the effectiveness of machine learning in detecting Parkinson's Disease (PD) through voice analysis. The results illustrate that ensemble methods, especially XGBoost and Random Forest, offer the best accuracy because of their non-linear dimensional subclass speech feature expressions. These models surpass conventional classifiers like Logistic Regression and Naive Bayes, which do not cope with the complex interdependencies of speech features and voice pattern intricacies. The results support the importance of pd diagnostics using machine learning models for early and non-invasive detection.

The results are promising, but further research is needed to enhance model robustness and ensure real-world applicability. Several key areas for future improvement include:

- **Advanced Computational Methods:** Investigating deep learning approaches such as Convolutional Neural Networks(CNNs) and Long Short-Term Memory(LSTMs) for automated feature extraction and enhanced classification accuracy.
- **Feature Engineering:** Expanding the set of voice based features by incorporating additional acoustic and prosodic parameters, which could provide a more comprehensive representation of speech impairments in PD patients.
- **Real-Time Detection Systems:** Developing a mobile or web based application capable of real time voice analysis for continuous PD monitoring, making the technology accessible for early detection and patient follow-up.
- **Clinical Integration:** Collaborating with neurologists and speech therapists to test the model's effectiveness in real-world diagnostic situations, confirming its practicality for medical application.

By integrating these advancements, voice based computational analysis has the potential to become a valuable tool in early PD detection, enabling timely medical intervention and improved patient outcomes.

REFERENCES

[1] Amato, F., Borzì, L., Olmo, G. et al. An algorithm for Parkinson's disease speech classification based on isolated words analysis. Health Inf Sci Syst 9, 32 (2021). https://doi.org/10.1007/s13755-021-00162-8

[2] Srinivasan, S., Ramadass, P., Mathivanan, S. et al. Detection of Parkinson disease using multiclass machine learning approach. Sci Rep 14, 13813 (2024). https://doi.org/10.1038/s41598-024-64004-9

[3] Ali, L., Javeed, A., Noor, A. et al. Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network. Sci Rep 14, 1333 (2024). https://doi.org/10.1038/s41598-024-51600-y

[4] Emamzadeh FN and Surguchov A (2018) Parkinson's Disease: Biomarkers, Treatment, and Risk Factors. Front. Neurosci. 12:612. https://doi.org/10.3389/fnins.2018.00612

[5] Alshammri R, Alharbi G, Alharbi E and Almubark I (2023) Machine learning approaches to identify Parkinson's disease using voice signal features. Front. Artif. Intell. 6:1084001. https://doi.org/10.3389/frai.2023.1084001

[6] Majhi, B., Kashyap, A., Mohanty, S.S. et al. An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm. BMC Med Imaging 24, 156 (2024). https://doi.org/10.1186/s12880-024-01335-z

[7] Suppa A, Costantini G, Asci F, Di Leo P, Al-Wardat MS, Di Lazzaro G, Scalise S, Pisani A and Saggio G (2022) Voice in Parkinson's Disease: A Machine Learning Study. Front. Neurol. 13:831428. https://doi.org/10.3389/fneur.2022.831428

[8] Swain K, Samal S, Ravi V, Nayak S, Alahmadi T, Singh P, Diwakar M. Towards Early Intervention: Detecting Parkinson's Disease through Voice Analysis with Machine Learning. Open Biomed Eng J, 2024; 18: e18741207294056. http://dx.doi.org/10.2174/0118741207294056240322075602

[9] Iyer, A., Kemp, A., Rahmatallah, Y. et al. A machine learning method to process voice samples for identification of Parkinson's disease. Sci Rep 13, 20615 (2023). https://doi.org/10.1038/s41598-023-47568-w

[10] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machin e Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, https://doi.org/10.1109/ACCESS.2020.3016062

[11] Luna-Ortiz I, Aldape-Pérez M, Uriarte-Arcia AV, Rodríguez-Molina A, Alarcón-Paredes A, Ventura-Molina E. Parkinson's Disease Detection from Voice Recordings Using Associative Memories. Healthcare (Basel). 2023 May 30;11(11):1601. https://doi.org/10.3390/healthcare11111601

[12] Mei Jie, Desrosiers Christian, Frasnelli Johannes, (2021), "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature", in Frontiers in Aging Neuroscience, vol. 13, https://doi.org/10.3389/fnagi.2021.633752

[13] Zeng, Taisheng and Yang, Tong and Liu, Peizhong and Zhu, Daxin, Parkinson Disease Prediction Using Machine Learning–Based Features from Speech Signal. http://dx.doi.org/10.2139/ssrn.4201020

[14] Luna-Ortiz I, Aldape-Pérez M, Uriarte-Arcia AV, Rodríguez-Molina A, Alarcón-Paredes A, Ventura-Molina E. Parkinson's Disease Detection from Voice Recordings Using Associative Memories. Healthcare (Basel). 2023 May 30;11(11):1601. https://doi.org/10.3390/healthcare11111601

[15] Abdullah H. Al-Nefaie, Theyazn H. H. Aldhyani and Deepika Koundal. Developing System-based Voice Features for Detecting Parkinson's Disease Using Machine Learning Algorithms. JDR. 2024. Vol. 3(1). https://doi.org/10.57197/JDR-2024-0001