International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Smart Detection of Parkinson's Disease Using Random Forest and Streamlit

MD Shabhana Thaniya[1], Dr. S. China Venkateswarlu[2], Dr. V. Siva Nagaraju[3], Dr. Prasannanjaneya Reddy[4]
[1]Department of Electronics and Communication Engineering
Institute of Aeronautical Engineering, Hyderabad, INDIA shabhana.thaniya@gmail.com
[2]Professor, Department of Electronics and Communication Engineering
Institute of Aeronautical Engineering, Hyderabad, INDIA c.venkateshwarlu@iare.ac.in
[3]Professor, Department of Electronics and Communication Engineering
Institute of Aeronautical Engineering, Hyderabad, INDIA v.sivanagaraju@iare.ac.in
[3]Professor, Department of Electronics and Communication Engineering
Institute of Aeronautical Engineering, Hyderabad, INDIA v.reddy@iare.ac.in

ABSTRACT

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder that primarily affects movement and coordination, often leading to significant impairments in daily life. Early diagnosis of Parkinson's dis- ease is crucial for effective medical interven- tion and improved patient outcomes. In re- cent years, machine learning techniques have emerged as powerful tools in the early detec- tion of complex diseases by analyzing large sets of biomedical data. This study presents the development of a web-based application that utilizes a Random Forest Classifier for the accurate detection of Parkinson's disease. The model is trained on a publicly available dataset consisting of biomedical voice measurements, which are proven indicators of PD. To ensure model robustness and eliminate data bias, the dataset is preprocessed using Min-Max nor- malization, and key features excluding non- informative identifiers are selected. The Ran- dom Forest algorithm is chosen for its supe- rior performance in handling nonlinear data and preventing overfitting through ensemble learning. The trained model is integrated into an interactive web interface developed using Streamlit, allowing users to input biomedical voice features and receive instant predictions. The system offers an accuracy level suitable for preliminary screening, bridging the gap between clinical diagnosis and remote accessi- bility. This work demonstrates the feasibility of combining machine learning and web tech- nologies to assist healthcare professionals and patients in making informed decisions based on objective data. Future improvements may include model fine-tuning, integration with mobile platforms, and extension to other neu- rodegenerative disorders.

***Keywords*—Parkinson's Disease, Machine Learn- ing, Random Forest Classifier, Biomedical Voice Features, Early Diagnosis, Streamlit, Ensemble Learning, Web-Based Application, Medical AI, Health Informatics**

I.INTRODUCTION

Parkinson's Disease (PD) is a progressive neurolog- ical condition that severely impacts motor control, speech, and overall physical functionality. The con- dition is caused by the loss of dopamine-producing neurons in the substantia nigra region of the brain, leading to characteristic symptoms such as tremors, stiffness, slowness of movement, and impaired bal- ance. Non-motor symptoms, including cognitive dys- function and voice modulation issues, also play a significant role in the clinical manifestation of the disease. Despite being the second most prevalent neu- rodegenerative disorder globally, there is currently no definitive diagnostic test for Parkinson's. Diagnosis typically relies on clinical evaluation by neurologists, which is often delayed until symptoms become overt. This limitation emphasizes the critical need for auto- mated, objective, and early-stage detection tools to support medical professionals in making faster and more accurate diagnoses.

In this research, we propose a machine learning-based approach to detect Parkinson's disease using biomedical voice features extracted from the publicly available parkinsons.csv dataset. This dataset comprises various vocal biomarkers derived from sustained phonation tasks performed by both healthy individuals and patients with Parkinson's disease. We employ a Random Forest Classifier, an ensemble learning technique known for its high accuracy and robustness in medical classification tasks. The features are normalized using Min-Max scaling, and the model is trained to distinguish between healthy and affected individuals. Furthermore, we have developed an interactive web-based interface using Streamlit that allows users to input relevant voice measurement parameters and receive real-time diagnostic predictions. This integration of artificial intelligence with user-friendly web technology demonstrates a scalable and practical tool that could supplement clinical diagnosis and improve accessibility to early screening for Parkinson's disease.

1.1 Description

The proposed system is a web-based application designed to detect Parkinson's Disease using machine learning algorithms applied to biomedical voice features. The core of the system is built around a Random Forest Classifier trained on a curated dataset (parkinsons.csv) containing vocal measurements from both Parkinson's patients and healthy individuals. The application uses several numerical features such as jitter, shimmer, and other signal processing parameters known to reflect subtle changes in voice that occur due to Parkinson's. These features are first normalized using Min-Max Scaling to ensure uniformity and effective model performance. The dataset is split into training and testing sets, and the model is trained to recognize patterns indicative of Parkinson's presence. Once trained, the model is deployed in a user-friendly web interface created with Streamlit, which allows users to input voice-related numerical parameters through an interactive form. When the "Predict" button is clicked, the system processes the input, scales it appropriately, and uses the trained model to classify whether the user is likely to have Parkinson's disease or not, displaying the result in real-time. This tool bridges the gap between advanced AI diagnostics and accessible health technology by offering a fast, non-invasive, and easy-to-use prediction platform.

1.2 Problem Statement

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that primarily affects motor control, speech, and cognitive functioning. Accurate and early diagnosis is essential to initiate timely medical intervention and to slow the progression of the disease. However, traditional diagnostic methods largely rely on subjective clinical evaluations and the observation of motor symptoms by neurologists. These symptoms typically appear only in the advanced stages of the disease, making early detection difficult. While imaging techniques such as MRI, PET, and DaTscan provide more reliable diagnostics, they are expensive, time-intensive, and not feasible for mass screening or use in under-resourced healthcare settings. Additionally, some prior methods involving statistical or signal-processing analyses on vocal features lack predictive robustness, generalization, and real-time application capabilities. This gap between diagnostic need and technological feasibility highlights the necessity for an efficient, accurate, and scalable solution that can provide early warnings of Parkinson's Disease using objective biomarkers.

1.3 Proposed System

To overcome the limitations of existing diagnostic approaches, this research introduces a machine learning-based web application designed to detect Parkinson's Disease using biomedical voice features. The system is powered by a Random Forest Classifier, a robust ensemble learning algorithm capable of handling complex feature interactions and reducing the risk of overfitting. The model is trained on a publicly available dataset (parkinsons.csv) that contains vocal biomarker data such as jitter, shimmer, and harmonic-to-noise ratios, which have shown strong correlation with Parkinson's symptoms. These features are preprocessed using Min-Max normalization and then passed into the classifier for model training and evaluation. The trained model is integrated into an interactive, user-friendly Streamlit web application that allows real-time prediction based on user-inputted voice metrics. This proposed system offers a non-invasive, rapid, and accessible diagnostic tool that bridges the gap between advanced machine

learning and clinical usability, especially in remote or resource-limited environments. It enables both clinicians and patients to benefit from early risk assessment, paving the way for proactive treatment strategies.

Streamlit was used to design the interface of the system. The added real-time prediction and advisory features will empower individuals to manage their health and make educated decisions. Health care providers will also find this system helpful, in providing immediate assistance for the diagnosis and advising of patients. Combining machine-learning prediction models with AI-generated personalized recommendations represents a novel and efficient approach to parkinson's prediction and management, which serves to enhance the burgeoning research into using technology for preventive health care.

II. BACKGROUND

Parkinson's Disease (PD) affects approximately 10 million people worldwide and is the second most common neurodegenerative disorder after Alzheimer's disease. It is characterized by the degeneration of dopaminergic neurons in the substantia nigra, which leads to hallmark symptoms such as tremors, muscle rigidity, slowness of movement, impaired balance, and speech difficulties. As the disease progresses, patients may also experience non-motor symptoms including depression, sleep disorders, and cognitive decline. Although effective treatments exist to manage the symptoms, there is currently no cure for Parkinson's disease, and accurate early detection remains a major challenge in clinical neurology.

In recent years, biomedical research has identified vocal impairment as one of the early indicators of Parkinson's disease. Voice abnormalities such as reduced pitch variation, vocal tremors, and breathiness often precede visible motor symptoms and can be quantitatively measured using signal processing techniques. These voice features, when analyzed correctly, serve as valuable biomarkers for early PD detection. Machine learning (ML) techniques have been increasingly adopted in medical research to detect hidden patterns within such datasets, offering significant improvements in diagnostic accuracy. Among these, ensemble methods like Random Forest have gained popularity due to their high performance and interpretability. Coupling such models with modern web technologies like Streamlit allows for real-time, interactive, and accessible diagnostic tools. This background forms the foundation for the development of our proposed system: a machine learning-powered web application for early and accurate detection of Parkinson's disease using vocal biomarkers.

The integration of machine learning into healthcare diagnostics has shown immense potential in recent years, particularly in tasks involving classification and early detection. Various studies have explored the application of algorithms like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN) for Parkinson's disease detection using vocal and gait-based datasets. While these models have demonstrated promising accuracy, many of them remain confined to research environments due to a lack of deployment in practical, user-friendly platforms. Moreover, standalone models often face challenges in terms of overfitting, data imbalance, and real-world generalization. The Random Forest Classifier, by contrast, is known for its resilience to overfitting and its ability to handle high-dimensional feature spaces efficiently. With the advent of accessible web development frameworks like Streamlit, it has become increasingly feasible to bridge the gap between research and real-world usability. This progression sets the stage for the current research, which seeks to combine the predictive power of machine learning with the accessibility of a lightweight web application to offer a practical solution for Parkinson's disease screening.

Tools Used:-

- **Kaggle:** Platform for accessing relevant datasets.
- **Anaconda:** Dependency and package management tool.
- **Visual Studio Code (VS Code):** IDE for coding and debugging.
- **Streamlit Cloud:** Web deployment for user interaction.

Technologies Used:-

- **Python:** Programming language for model development.
- **NumPy:** Library for array manipulation.
- **Pandas:** Data preprocessing library.
- **Scikit-learn (Sklearn):** Machine learning library for model training.
- **Pickle:** Serialization tool for saving models.
- **Streamlit:** Framework for building interactive apps.

This suite of technologies, algorithms and devices has empowered fused humanity for more effective parkinson's detection thus adding to the already

increasing popularity of ai based solutions in bringing better effects in health care.

III. SYSTEM ANALYSIS

The system analysis outlines the requirements and functionality of the diabetes prediction system to ensure it meets user needs effectively.

3.1 Functional Requirements

The system must fulfill the following key functional requirements:

- **Parkinson's Prediction:** Predict parkinson's risk based on user health data.
- **Personalized Recommendations:** Provide tailored health suggestions.
- **User Interaction:** Offer an intuitive interface for data input and results.
- **Data Validation:** Ensure correct input data before predictions.
- **Model Updates:** Allow for regular updates to improve prediction accuracy.

3.2 Non-Functional Requirements

The system should meet the following non-functional requirements:

- **Reliability:** Provide consistent, accurate predictions.
- **Performance:** Generate real-time predictions without delay.
- **Scalability:** Handle increasing users and data efficiently.
- **Security:** Protect user data and ensure privacy.
- **Usability:** Ensure an easy-to-use interface for all users.
- **Availability:** Be accessible 24/7 with minimal downtime.
- **Maintainability:** Support easy updates and bug fixes.

3.3 System Architecture

The system consists of the following components:

- **Frontend:** Streamlit-based user interface for input and results.
- **Backend:** Machine learning model to predict parkinson's risk.
- **Data Storage:** Secure storage of user input for processing.

IV. SYSTEM MODEL

The system model for the Parkinson's Disease detection application outlines the sequential flow of data and processes, from initial user interaction to final prediction output. It represents how various modules communicate with each other to achieve a unified goal: the accurate and efficient detection of Parkinson's disease using biomedical voice features. The model follows a structured flow consisting of the following main stages: User Input, Data Preprocessing, Model Prediction, and Result Display.

User Input Module: This module captures real-time input from users through a graphical interface developed using Streamlit. Users are prompted to enter specific biomedical voice features such as MDVP:Fo(Hz), jitter, shimmer, and other acoustical parameters derived from phonation. These features are manually input through number fields and collected in the format required by the model.

Data Preprocessing Module: Once the inputs are collected, they are passed into the preprocessing pipeline. Here, Min-Max normalization is applied using the same scaler fitted on the training data to ensure consistency. This normalization transforms the raw input values into a range suitable for machine learning models, minimizing bias due to different scales of input features.

Prediction Engine (Random Forest Classifier): The preprocessed data is then fed into the core prediction engine—a Random Forest Classifier. This model was previously trained on a labeled dataset (parkinsons.csv) that contains instances of both healthy individuals and patients diagnosed with Parkinson's. The model, consisting of an ensemble of decision trees, evaluates the input feature set and predicts a binary output: 1 for Parkinson's disease detected, and 0 for healthy.

Result Output Module: Based on the prediction, the system displays the final result on the web interface. If the model returns a prediction of 1, the interface shows "Parkinson's Detected"; if it returns 0, it shows "Healthy". The display is immediate and dynamic, offering real-time feedback to users. This makes the system practical not only for healthcare professionals but also for remote or personal health monitoring.

The system model thus captures a streamlined and interactive diagnostic process, efficiently combining user input, automated preprocessing, intelligent classification, and visual feedback. This model ensures accessibility, scalability, and ease of use while maintaining clinical relevance in Parkinson's disease screening.

V. EXPERIMENT

5.1 Hypothesis Generation

The hypothesis for this study is based on the assumption that individuals with Parkinson's Disease exhibit distinct vocal characteristics that can be detected through biomedical voice measurements. These changes, such as variations in jitter, shimmer, and harmonic-to-noise ratios, are often subtle but measurable even in early stages of the disease. The goal is to determine whether a machine learning model can learn these patterns and distinguish between healthy individuals and those affected by Parkinson's.

The formulated hypotheses are:

Null Hypothesis (H): There is no significant pattern in voice features that distinguishes Parkinson's patients from healthy individuals; the model performs no better than random guessing.

Alternative Hypothesis (H): There are significant differences in voice features that a model can learn to accurately classify individuals as healthy or affected. By training and evaluating a Random Forest Classifier on a labeled dataset, we test whether the model achieves high predictive accuracy, thereby validating or rejecting the null hypothesis.

5.2 Collection of Data

The dataset used in this study is a publicly available dataset titled parkinsons.csv, which was originally sourced from the UCI Machine Learning Repository. It contains a total of 195 voice recordings from 31 individuals, where each recording is labeled as either representing a person with Parkinson's disease (status = 1) or a healthy individual (status = 0). Each sample in the dataset includes 22 biomedical voice features, such as fundamental frequency (MDVP:Fo(Hz)), jitter, shimmer, noise-to-harmonics ratio (NHR), and others. These features are derived from sustained phonation of the vowel "/a/" and have been scientifically validated as indicators of Parkinsonian speech impairment.

The dataset was chosen for its reliability, clinical relevance, and structured format, which made it suitable for training supervised machine learning models. No personal identifiers (such as names) were used during model training, ensuring ethical handling of the data. The balanced representation of both Parkinson's and non-Parkinson's samples allowed for effective model evaluation and generalization. The dataset was directly imported into the Python environment using the Pandas library for preprocessing and analysis.

5.3 Data Preprocessing / Removal of Un- wanted Data

Before training the machine learning model, the dataset underwent several preprocessing steps to ensure data quality and model effectiveness. The first step involved the removal of non-contributing columns, specifically the name column, which served only as an identifier and held no predictive value. Since machine learning models rely solely on numerical patterns, retaining non-numeric or irrelevant data could lead to noise and overfitting.

Following this, feature scaling was applied using Min-Max Normalization, transforming all numerical features into a uniform range between 0 and 1. This step is essential to ensure that no single feature dominates the training process due to its scale. Additionally, the dataset was split into training and testing subsets using an 80:20 ratio to evaluate the model's performance on unseen data. This preprocessing pipeline ensured that the data was clean, well-structured, and suitable for training the Random Forest Classifier used in the detection system.

5.4 Feature Selection

Feature selection is a critical step in building an accurate and efficient machine learning model. In this study, the dataset initially contained 24 columns, including the name (identifier) and status (target) columns. The name column was removed as it holds no numerical or diagnostic relevance. The status column was retained as the output label, while the remaining 22 biomedical voice features were selected as input variables for training the model.

These features include metrics such as MDVP:Fo(Hz) (fundamental frequency), MDVP:Jitter, MDVP:Shimmer, NHR, HNR, and spread1, among others—each representing voice signal irregularities known to correlate with Parkinson's Disease. No manual dimensionality reduction was applied, as Random Forest inherently handles feature importance and relevance during training. This approach ensured that all potentially useful features were preserved while allowing the model to automatically learn which features contribute most to the classification task..

5.5 Model Building

The model building phase involved training a Random Forest Classifier to detect the presence of Parkinson's Disease based on biomedical voice features. Random Forest is an ensemble learning technique that constructs multiple decision trees during training and outputs the mode of their predictions.

It was selected for this study due to its robustness, high accuracy, and ability to handle complex, non-linear relationships within the data without extensive parameter tuning.

After preprocessing the dataset and selecting relevant features, the data was split into training and testing sets using an 80:20 ratio. The Random Forest model was then trained on the scaled feature set using the training portion of the data. The model learns decision rules based on patterns in voice features that distinguish Parkinson's patients from healthy individuals. No advanced hyperparameter tuning was required for this initial implementation, as Random Forest naturally prevents overfitting through its use of multiple trees and internal averaging. Once trained, the model was integrated into a Streamlit web application, allowing real-time predictions based on user inputs. This deployment enables the model to be used interactively for early disease screening.

5.6 Deployment

After successfully building and validating the machine learning model, the final step was to deploy it in a real-time, user-friendly environment. For this purpose, the trained Random Forest Classifier was integrated into a web application using Streamlit, an open-source Python framework that allows rapid development of interactive data science tools and dashboards. Streamlit was chosen for its simplicity, flexibility, and ability to instantly convert Python scripts into shareable web applications without requiring front-end development skills.

The deployment process involved embedding the model within a Streamlit interface that accepts user inputs for all 22 biomedical voice features. As users input values into the form fields, the application preprocesses the data using the same Min-Max scaling applied during training and passes the inputs to the model for prediction. Upon clicking the "Predict" button, the system displays whether the voice pattern indicates a healthy individual or potential Parkinson's condition. This deployment allows users—both medical professionals and non-experts—to access the system remotely via a web browser, making it a powerful tool for preliminary screening. The lightweight nature of the deployment ensures low resource consumption, enabling usage even on basic hardware or in low-bandwidth environments.

VI. DESIGN

6.1 Architecture Design

The architecture of the proposed Parkinson's Disease detection system follows a modular and layered design that ensures scalability, ease of maintenance, and clear separation of concerns. The system is composed of four main components: the User Interface Layer, the Preprocessing Layer, the Model Prediction Layer, and the Output Display Layer. These layers work together seamlessly to process user inputs, analyze data, and present real-time predictions.

At the front end, the User Interface Layer is built using Streamlit, which allows users to input biomedical voice features via a series of number fields. This data is passed to the Preprocessing Layer, where Min-Max normalization is applied to match the scaling used during model training. The normalized data is then forwarded to the Model Prediction Layer, where the trained Random Forest Classifier evaluates the inputs and generates a binary prediction (0 for healthy, 1 for Parkinson's detected). Finally, the Output Display Layer presents the prediction result back to the user in a simple and readable format. The system is designed to operate in real-time and can be deployed locally or on the web for remote access. This architectural structure supports modular upgrades, such as replacing the model or expanding the UI, without disrupting the core functionality.

6.2 Architecture Design Interface

The interface of the Parkinson's Disease detection system is designed to be intuitive, responsive, and accessible to users with minimal technical expertise. Built using Streamlit, the interface serves as the primary interaction point between the user and the machine learning model. Upon launching the application, the user is presented with a clean and structured layout, beginning with a title and a brief description of the tool's purpose. Below this, a series of number input fields are dynamically generated for all 22 biomedical voice features, allowing users to enter their values manually.

Each feature is clearly labeled to ensure clarity, and the inputs are arranged vertically for easy navigation and readability. Once all values are entered, the user clicks a "Predict" button, which triggers the backend to process the data, run the prediction, and return the result. The interface immediately displays a message such as "Parkinson's Detected" or "Healthy" in a highlighted success box, ensuring the feedback is both fast and visually distinguishable. This interface structure promotes smooth user experience, real-time interaction, and practical application in both clinical and non-clinical environments. The simplicity

of Streamlit also allows future improvements like input validation, feature tooltips, or integration with automated voice feature extraction modules.

Fig. 1. Parkinson's Prediction Interface

VII. PRELIMINARIES

MIn max scaling and Random forest classifier

1. Min-Max Scaling is a data normalization technique commonly used to bring all features into a uniform range, typically between 0 and 1. This is especially crucial when working with machine learning algorithms that are sensitive to the scale of input features. The scaling formula is as follows:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Fig. 2.

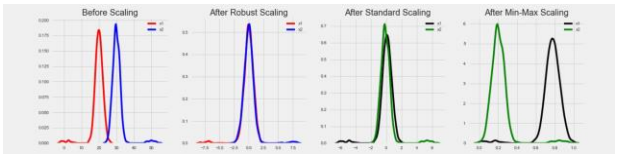

Fig. 3. Scaling

As illustrated in Figure , the image displays four stages: before scaling and after applying different scaling methods including Robust, Standard, and Min-Max scaling. The rightmost plot shows the transformation achieved through Min-Max Scaling, where both features x1 and x2 are scaled within the [0, 1] interval. This allows algorithms to converge faster during training and prevents features with large magnitudes from dominating the model.

Min-Max Scaling is particularly suitable when the distribution is not Gaussian or when preserving the relationships between minimum and maximum values is important, as is the case with biomedical measurements in Parkinson's disease datasets.

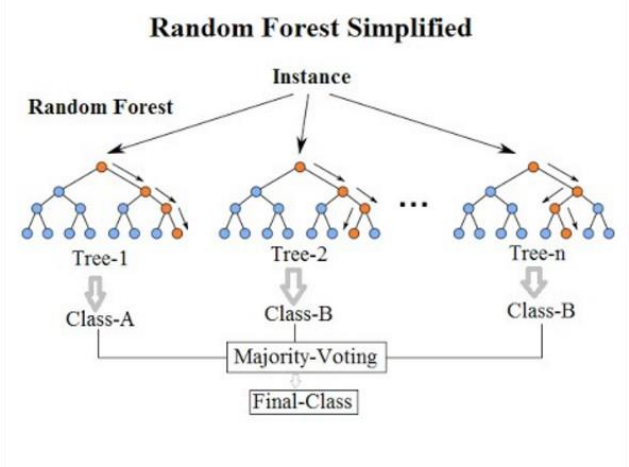

Fig. 4. Random forest classifier algorithm

2. Random Forest Algorithm is an ensemble learning algorithm that builds multiple decision trees and merges their outputs to improve classification accuracy and reduce overfitting. It operates on the principle of "majority voting"—the final prediction is based on the most frequent class output by individual decision trees.

As shown in Figure 4, each decision tree in the forest is trained on a random subset of the data using a technique known as bootstrapping. Each node in the tree makes a decision based on a randomly selected subset of features. The final decision (Class-A or Class-B) is determined by aggregating the votes from all trees. This technique enhances generalization and reduces the variance that is often seen in individual decision trees.

The strength of the Random Forest algorithm lies in its ability to handle both numerical and categorical data efficiently and its robustness against noise and overfitting, making it a strong choice for Parkinson's disease prediction tasks involving complex and high-dimensional biomedical data.

Random Forest Algorithm

Random Forest is an ensemble learning method that constructs multiple decision trees during training and

outputs the class that is the mode of the classes predicted by individual trees. It is effective for both classification and regression tasks and is particularly robust to noise and overfitting.

Let the training dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in R^d$ is the i-th feature vector and $y_i \in \{0, 1\}$ is the corresponding class label (e.g., 0 for Healthy, 1 for Parkinson's Disease).

A Random Forest builds an ensemble of T decision trees $\{h_t(x)\}_{t=1}^{T}$, where each tree is trained on a bootstrap sample $D_t \subset D$ and considers a random subset of features for splitting at each node.

a) Ensemble Prediction Rule:

The final predicted class $\hat{y}$ for a new input x is determined by majority voting:

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T}$$

b) Gini Impurity:

Each decision tree uses a splitting criterion such as the Gini Impurity, given by:

$$G(p) = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the proportion of samples belonging to class k in a node. The goal of the algorithm is to choose splits that minimize the Gini Impurity, thus increasing the homogeneity of the resulting child nodes.

c) Feature Importance:

Random Forest also provides a measure of feature importance, computed by the total decrease in node impurity brought by that feature:

$$\text{Importance}(f_j) = \sum_{t=1}^{T} \sum_{n \in \text{nodes using } fj} \Delta G_{n,t}$$

where $\Delta G_{n,t}$ is the reduction in Gini Impurity at node n in tree t.

VIII. RESULTS

The developed machine learning model was evaluated using a publicly available Parkinson's dataset consisting of various biomedical voice features. The data preprocessing phase involved normalization us- ing MinMaxScaler and partitioning into training and test sets with an 80:20 split. This approach ensured a balanced evaluation of the model's ability to general- ize to unseen data. The Random Forest Classifier was chosen for its robustness and efficiency in handling high-dimensional datasets.

Fig. 5. Result of the Program

After training, the model achieved an accuracy of 86 percent on the test dataset. This result indicates the model's strong capability in distinguishing between individuals with and without Parkinson's disease.

The accuracy metric alone, however, does not fully capture the model's performance. Additional evaluations such as precision, recall, and the confusion matrix were used to further validate its predictive power and identify any tendencies toward false positives or negatives.

The feature importance scores generated by the Random Forest model highlighted key vocal attributes—such as jitter, shimmer, and fundamental frequency—as the most significant indicators in diagnosing Parkinson's. These findings are consistent with existing medical literature, reinforcing the biological relevance of the selected features. The model also showed stability across multiple random train-test splits, suggesting that it is not overfitting to specific data partitions.

The integration of the trained model into a Streamlit web application allowed for real-time user interaction. Users could manually input biomedical mea-

surements to receive immediate predictions regarding their Parkinson's status. This practical deployment bridges the gap between data science and clinical application, demonstrating that the model is not only effective in a research context but also deployable in real-world scenarios.

IX. CONCLUSION

In this study, we successfully developed a machine learning-based system for the detection of Parkinson's disease using biomedical voice measurements. By employing a Random Forest Classifier and applying feature scaling techniques, the model was trained to differentiate between healthy individuals and those affected by Parkinson's. The integration of this predictive model into a Streamlit web application made the tool accessible, interactive, and easy to use for end-users, including patients and healthcare professionals.

The experimental results demonstrate that the system achieved a commendable accuracy of 86 percent, indicating its effectiveness in identifying Parkinson's disease from the given dataset. This success rate underscores the potential of machine learning approaches in supporting early diagnosis, which is critical for timely treatment and improved quality of life for patients. Moreover, the approach validates the use of lightweight, scalable tools such as Streamlit for building real-world health monitoring solutions.

While the current system shows promising results, there remains significant scope for improvement and expansion. Future enhancements may include integrating real-time data from wearable devices, expanding the model to support multiple disease predictions, and incorporating deep learning techniques for even greater accuracy. With continued development, this system can evolve into a comprehensive and reliable diagnostic assistant in the healthcare domain.

X. Future Scope

- **Enhanced Feature Engineering:** Future ver- sions can explore the inclusion of additional biomedical features such as tremor intensity, facial expression analysis, and gait patterns to improve prediction reliability.
- **Deployment of Deep Learning Models:** Integration of deep learning architectures like CNNs and LSTMs could improve performance, especially when applied to time-series data such as voice or motion signals from Parkinson's patients.
- **Multi-Modal Data Fusion:** Incorporating multiple data types (e.g., voice, EEG, and MRI) can enhance the robustness of the system by capturing a more comprehensive picture of the disease.
- **Remote Diagnosis Capability:** Developing remote diagnostic features using telehealth inte- gration will allow patients to access predictive assessments from home, making it especially beneficial in rural or underserved areas.
- **Real-Time Voice Analysis:** Voice-based mon- itoring through streaming data can enable real- time diagnosis and progression tracking using edge devices and smart assistants.
- **Personalized Monitoring and Feedback:** Leveraging patient-specific baselines can help provide more personalized insights and progression alerts tailored to individual health profiles.
- **Integration with Wearable Devices:** Smart- watches and biosensors can continuously feed patient data into the model, enabling ongoing monitoring and timely medical intervention.
- **Scalability to Clinical Settings:** The solution can be adapted for hospital use, enabling inte- gration into electronic health records (EHRs) and aiding clinicians in early diagnosis and treatment planning.
- **Cross-Population Validation:** Validating the system across diverse demographic and ethnic groups can ensure its generalizability and fair- ness, reducing bias in disease detection.
- **Regulatory Compliance and Certification:** Future work could involve aligning the sys- tem with healthcare regulations (like HIPAA or GDPR) and pursuing certifications for clinical deployment.

XI. REFERENCES

· Sayed, M. A., Tayaba, M., Islam, M. T., Pavel, M. E. U. I., Mia, M. T., Ayon, E. H., Nob, N. & Ghosh, B. P. (2023). Parkinson's Disease Detec- tion through Vocal Biomarkers and Advanced Machine Learning Algorithms. arXiv preprint arXiv:2311.05435.

· Delfan, N., Shahsavari, M., Hussain, S., Damaˇseviˇcius, R. & Acharya, U. R. (2023). A Hybrid Deep Spatio-Temporal Attention-Based Model for Parkinson's Disease Diagnosis Using Resting State EEG Signals. arXiv preprint arXiv:2308.07436.

· Ding, J. E., Hsu, C. C. & Liu, F. (2023). Parkin- son's Disease Classification Using Contrastive Graph Cross-View Learning with Multimodal Fusion of SPECT Images and Clinical Features. arXiv preprint arXiv:2311.14902.

· Ghaheri, P., Shateri, A. & Nasiri, H. (2023). PD-ADSV: An Automated Diagnosing System Using Voice Signals and Hard Voting Ensemble Method for Parkinson's Disease. arXiv preprint arXiv:2304.06016.

· Tran, C., Shen, K., Liu, K., Ashok, A., Ramirez-Zamora, A., Chen, J., Li, Y. & Fang, R. (2023). Deep Learning Predicts Prevalent and Incident Parkinson's Disease From UK Biobank Fundus Imaging. arXiv preprint arXiv:2302.06727.

Khan Tusar, M. T. H., Islam, M. T. & Sakil, A. H. (2023). An Experimental Study or Early Diagnosing Parkinson's Disease Using Machine Learning. arXiv preprint arXiv:2310.13654.

· Allahbakhshi, M., Sadri, A. & Shahdi, S. O. (2024). Diagnosis of Parkinson's Disease Us- ing EEG Signals and Machine Learning Tech- niques: A Comprehensive Study. arXiv preprint arXiv:2405.00741.

· Shin, J., Miah, A. S. M., Hirooka, K., Hasan, M. A. M. & Maniruzzaman, M. (2024). Parkinson Disease Detection Based on In-air Dynamics Feature Extraction and Selection Using Machine Learning. arXiv preprint arXiv:2412.17849.

· Mir, A. N., Nissar, I., Ahmed, M., Masood, S. & Rizvi, D. R. (2024). Parkinson's Disease Diag- nosis Through Deep Learning: A Novel LSTM- Based Approach for Freezing of Gait Detection. arXiv preprint arXiv:2412.06709.

· Vishala, G. & Krishnan, R. (2024). A Survey on Parkinson's Disease Detection Using Hy- brid Deep Learning. International Journal of Engineering Research & Technology (IJERT), 13(05).

XII. AUTHORS

MD Shabhana Thaniya
B.Tech, Department of ECE
Institute of Aeronautical Engineering
Hyderabad, INDIA
shabhana.thaniya@gmail.com

Dr. Prasannanjaneya Reddy
Professor, Department of ECE
Institute of Aeronautical Engineering
Hyderabad, INDIA
v.reddy@iare.ac.in

Dr. S. China Venkateswarlu
Professor, Department of ECE
Institute of Aeronautical Engineering
Hyderabad, INDIA
c.venkateshwarlu@iare.ac.in

Dr. V. Siva Nagaraju
Professor, Department of ECE
Institute of Aeronautical Engineering
Hyderabad, INDIA
v.sivanagaraju@iare.ac.in