Early Diagnosis of Parkinson's Disease Using Machine Learning Models

Pelin Buçukoğlu
Department of Computer Engineering
TOBB University of Economics and Technology
Ankara, Turkey
Email: pbucukoglu@gmail.com

***Abstract*—In this study, various machine learning algorithms were utilized to assist in the early diagnosis of Parkinson's Disease (PD), which significantly affects patients' quality of life when not diagnosed in its early stages. The dataset used includes demographic, clinical, and behavioral attributes of 2,105 patients. Data preprocessing techniques such as handling missing values, outlier treatment, and class balancing with SMOTE were applied. Models including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Decision Tree were trained and evaluated using accuracy, precision, recall, and F1-score. Additionally, hyperparameter tuning with GridSearchCV and cross-validation techniques were used to enhance model performance. The results indicate that ensemble methods and tree-based models outperform basic classifiers, providing promising directions for PD prediction tools.**

I. Introduction

Parkinson's Disease (PD) is a chronic, progressive neurodegenerative disorder that primarily affects the motor system due to the loss of dopamine-producing neurons in the substantia nigra region of the brain. It is the second most common neurodegenerative disorder globally after Alzheimer's disease and predominantly affects individuals over the age of 60. The hallmark symptoms include resting tremor, rigidity, bradykinesia (slowness of movement), and postural instability. Non-motor symptoms, such as sleep disturbances, depression, and cognitive impairment, may also accompany the disease and often precede the onset of motor dysfunction. Unfortunately, by the time PD is clinically diagnosed through motor symptoms, it is estimated that over 60–80% of the dopaminergic neurons have already degenerated, limiting the efficacy of therapeutic interventions.

The gold standard for PD diagnosis remains a comprehensive neurological examination conducted by a specialist. However, this process is inherently subjective, relying heavily on physician expertise and patient-reported symptoms. Additionally, early PD symptoms often overlap with other movement disorders, such as essential tremor or drug-induced parkinsonism, complicating differential diagnosis. These challenges highlight the urgent need for objective, data-driven diagnostic tools that can facilitate earlier and more accurate detection of PD.

In recent years, the field of artificial intelligence (AI), and more specifically machine learning (ML), has demonstrated significant promise in biomedical applications, including disease diagnosis, prognosis prediction, and treatment optimization. ML algorithms are particularly effective in identifying complex, non-linear relationships within high-dimensional datasets, a characteristic feature of clinical and biometric data. Numerous studies have explored the application of supervised learning methods such as Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN) in the context of PD diagnosis, achieving classification accuracies as high as 97% when optimized appropriately [1]–[3].

For example, Kaladhar et al. [1] demonstrated that SVM outperformed several classical classifiers in identifying PD using speech signal features, achieving approximately 93% accuracy. López et al. [4] utilized Random Forests and SVMs to classify PD using voice data and reported that ensemble methods produced more stable predictions. More recently, He et al. [2] employed XGBoost for Parkinson's classification and showed its superiority in handling complex, high-dimensional features. These studies underscore the growing consensus that ML models, when trained on properly curated data, can serve as powerful clinical decision support systems.

In this study, we aim to build a robust machine learning pipeline for the early diagnosis of Parkinson's Disease using a comprehensive dataset obtained from Kaggle. The dataset contains detailed demographic, behavioral, and clinical information from 2,105 individuals, with 35 distinct features including symptoms, lifestyle factors, medical history, and lab results. Data preprocessing steps include handling missing values, outlier removal, categorical encoding, feature scaling, and addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE). Exploratory Data Analysis (EDA) is conducted to understand the statistical properties and distributions of features, and feature correlations are examined to inform modeling decisions.

We explore and compare the performance of multiple ML models: Support Vector Machines, K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest classifiers. In addition, a Voting Classifier is implemented to aggregate model decisions and potentially improve prediction stability. Each model is optimized using cross-validated grid search for hyperparameter tuning and evaluated using accuracy, precision, recall, and F1-score metrics across stratified folds.

This work differs from previous studies in several ways. First, we incorporate a broader range of features beyond speech signals, including lifestyle and medical history data, to improve generalizability. Second, we focus on a comparative analysis of classical ML models under a consistent preprocessing and evaluation framework. Finally, by applying ensemble methods and addressing data imbalance explicitly, we seek to construct a clinically relevant and practically deployable PD diagnostic tool.

Ultimately, this study contributes to the growing body of research advocating for the integration of AI in healthcare and supports the development of intelligent systems capable of assisting clinicians in early detection, thereby improving patient outcomes and optimizing healthcare resources.

II. Related Work

The application of machine learning (ML) to Parkinson's Disease (PD) diagnosis has attracted significant attention in recent years. The field has evolved rapidly, with numerous studies focusing on various types of data, algorithms, and preprocessing techniques. Below, we discuss key contributions in the area of ML-based PD prediction, highlighting their methodologies, results, and limitations.

Kaladhar et al. [1] conducted a study comparing different ML algorithms for PD classification, focusing on speech data. The dataset used in this study consisted of voice features extracted from PD patients and healthy controls. Several algorithms were employed, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, and Logistic Regression. Among these, SVM achieved the highest classification accuracy, approximately 93%. The authors emphasized the importance of feature engineering in improving model performance, highlighting how speech features such as jitter, shimmer, and harmonics-to-noise ratio can be used effectively for PD detection. However, the study was limited by the focus on a small dataset and a narrow range of features, which restricted the generalizability of the findings. Additionally, the study did not explore more advanced techniques like cross-validation or handling class imbalance, which are critical in real-world applications.

López et al. [4] expanded upon Kaladhar et al.'s work by utilizing speech data to diagnose PD. They applied Random Forest and SVM algorithms to voice features and found that Random Forest performed better in terms of stability and generalization. Their study demonstrated that ensemble methods, which combine multiple models, can offer more robust results compared to individual classifiers. Despite these promising results, their study was constrained by a relatively small sample size and the use of only speech data, which may not provide a comprehensive view of the disease. Furthermore, the lack of cross-validation in their methodology raises concerns about the model's ability to generalize to new, unseen data.

He et al. [2] proposed a novel classification model for PD using the eXtreme Gradient Boosting (XGBoost) algorithm. XGBoost is an ensemble technique that builds strong predictive models by combining the outputs of many weak learners. The authors reported an impressive accuracy of approximately 97%, which they attributed to XGBoost's ability to handle complex feature interactions and missing data. Furthermore, the study employed feature selection and cross-validation techniques, which enhanced the model's generalizability. However, the study did not delve into model interpretability, which is crucial for clinical adoption. While XGBoost achieved high accuracy, its "black-box" nature can make it difficult for clinicians to trust the model's decisions without understanding the underlying reasoning.

Ganaie et al. [3] introduced deep learning into PD classification by employing Convolutional Neural Networks (CNN). CNNs are particularly effective in handling high-dimensional and complex data, as they can automatically learn hierarchical features from raw data. In their study, Ganaie et al. used a combination of motor and speech data to classify PD, achieving accuracies of up to 97%. This work highlighted the potential of deep learning to outperform traditional ML algorithms in certain cases. However, CNNs require large datasets for training, and their computational cost can be prohibitive. Moreover, deep learning models are often considered "black boxes," making them less interpretable than classical models such as SVM or Random Forest, which can pose challenges for clinical applications where model transparency is essential.

Kumar et al. [5] conducted a comprehensive survey of ML-based approaches for PD diagnosis, reviewing both classical methods (e.g., SVM, Random Forest, KNN) and advanced techniques, such as deep learning and ensemble methods. The survey underscored the importance of hybrid models, which combine multiple algorithms to improve accuracy and robustness. Kumar et al. also highlighted the challenges of data imbalance and the need for effective feature selection techniques, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). Their work predicted that hybrid and ensemble models would become more prevalent in PD diagnosis, as they offer better generalization and performance. However, the survey did not include direct comparisons of models under standardized experimental conditions, which would have provided more insight into the effectiveness of each approach.

Despite the significant progress made in these studies, several challenges remain in the field. Many studies focus narrowly on specific feature types, such as speech signals or motor data, while neglecting the potential value of other sources of information, such as demographic and clinical data. Furthermore, many studies fail to employ rigorous data preprocessing techniques, such as addressing missing data, class imbalance, and scaling features. These factors are crucial for developing models that can be reliably deployed in clinical settings.

In contrast to these previous works, our study utilizes a comprehensive dataset that includes not only motor and speech features, but also demographic, clinical, and lifestyle data. By incorporating a wide range of features, we aim to build a more robust and generalizable model for PD detection. Additionally, we place a strong emphasis on preprocessing, including the use of SMOTE for balancing the classes, median imputation for handling missing values, and StandardScaler for feature normalization. Our approach also includes a detailed comparison of multiple classical ML models, including SVM, KNN, Logistic Regression, Decision Tree, and Random Forest, ensuring a fair and standardized evaluation. Furthermore, we implement an ensemble Voting Classifier to assess whether combining multiple models can improve prediction stability and accuracy.

Our work builds on the findings of previous studies while addressing their limitations. By incorporating a diverse set of features, applying rigorous preprocessing techniques, and evaluating multiple models under consistent conditions, we aim to contribute to the ongoing efforts to develop more accurate and interpretable machine learning models for PD diagnosis.

III. Dataset and Preprocessing

In this study, we use a comprehensive dataset sourced from the Kaggle ”Parkinson’s Disease Classification” competition, which contains over 2,100 records and 35 features per individual. The dataset includes both healthy controls and Parkinson’s Disease (PD) patients, with the primary goal of classifying individuals as either PD-positive or healthy based on various clinical, behavioral, and motor skill features. This dataset was chosen because of its richness, large sample size, and variety of features, making it suitable for developing a robust machine learning model for early-stage Parkinson’s detection.

A. Dataset Overview

The dataset is composed of several types of features, each contributing valuable information for classification. These include:

- **Demographic Features:** These features include basic information such as age, gender, and education level. Age is an important variable as PD predominantly affects older adults, and gender is a factor, with men generally having a higher incidence of the disease. Education level can also correlate with cognitive function, influencing how the disease manifests in different individuals.
- **Motor Skill Features:** These are derived from several motor tests, such as tremor, bradykinesia, and rigidity, which are essential indicators of Parkinson’s Disease progression. Speech-related features are also included, as PD affects speech motor control. These features provide direct insights into the severity of motor dysfunction, a hallmark symptom of Parkinson’s.
- **Clinical and Lifestyle Factors:** These features include physical and medical history aspects such as BMI, smoking history, alcohol consumption, and physical activity levels. Other clinical factors, such as family history of Parkinson’s Disease, depression, diabetes, hypertension, and previous brain injuries, are also included. These features are important because they can help identify individuals at higher risk for PD or those with genetic predispositions.

B. Data Preprocessing

The preprocessing of this dataset was crucial to ensure that it was suitable for training machine learning models. The following preprocessing steps were taken to prepare the data for analysis:

1) Handling Missing Values: Missing data is a significant challenge in real-world datasets, especially in medical contexts. For this study, the missing values were handled using median imputation. This method was chosen as it is effective in dealing with numerical data, ensuring that the imputation does not skew the distribution of features, particularly for continuous variables like age and BMI. Median imputation is also robust to outliers, which makes it a preferable method in clinical data where extreme values can often occur.

2) Outlier Detection and Removal: Outliers can significantly impact the performance of machine learning models by distorting the true relationship between features and the target variable. In this study, outliers were identified using boxplots, which were created for each continuous variable. Features like BMI and blood pressure, which are known to have occasional extreme values in medical datasets, were inspected carefully. Once identified, these outliers were replaced with the median of the respective feature, helping to standardize the data without losing the central tendency or introducing significant bias.

3) Feature Normalization: Normalization of features is essential for many machine learning algorithms, especially those that rely on distance metrics (e.g., K-Nearest Neighbors and Support Vector Machines). Features with vastly different scales, such as age (ranging from 30 to 80) and blood pressure (ranging from 90 to 180), could disproportionately affect the model if not scaled. Therefore, we applied StandardScaler, which standardizes the data to have zero mean and unit variance. This normalization ensures that all features are treated equally, improving the accuracy of models that are sensitive to the magnitude of input features.

4) Handling Class Imbalance with SMOTE: A critical challenge in medical datasets is class imbalance, which occurs when one class (e.g., PD patients) is underrepresented compared to the other (e.g., healthy controls). Class imbalance can lead to biased models that are more likely to predict the majority class. To address this, we applied the Synthetic Minority Over-sampling Technique (SMOTE), a widely used technique for balancing class distribution. SMOTE works by generating synthetic samples for the minority class by interpolating between existing samples. This technique allows the model to learn from a more balanced dataset, ultimately improving performance on the minority class and reducing bias.

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain insights into the structure of the data, identify any patterns or correlations between features, and visualize the distributions of key variables. Several key visualization techniques were employed:

- **Histograms and Boxplots:** These were used to examine the distributions of continuous variables such as age, BMI, and blood pressure. Boxplots helped identify potential outliers, while histograms provided insight into the skewness and range of the data. For example, age was found to follow a normal distribution, while BMI had a slight right skew, indicating the presence of overweight individuals in the dataset.
- **Correlation Matrix:** A heatmap of the correlation matrix was generated to explore relationships between numerical features. This matrix helped identify highly correlated features, such as age and BMI, which were then carefully analyzed to avoid multicollinearity in the models. Features with high correlation (greater than 0.9) were considered for removal or dimensionality reduction.
- **Pairplots:** Pairplots were used to visualize relationships between selected features. This allowed us to examine how different features interacted with each other and how they were distributed across the two classes (PD vs. healthy). This visualization revealed that certain features, such as tremor and bradykinesia, showed clear separability between the two classes.

D. Feature Selection

To improve model efficiency and reduce the risk of overfitting, feature selection was performed to identify the most relevant variables for PD classification. Two key techniques were used:

- **SelectKBest:** This method selects the top K features based on their statistical significance in predicting the target variable. Features with the highest correlation with the target variable (i.e., PD diagnosis) were chosen for inclusion in the final model.
- **Principal Component Analysis (PCA):** PCA was used to reduce the dimensionality of the dataset while retaining as much variance as possible. By transforming the original features into principal components, PCA helped improve computational efficiency and removed any redundant or correlated variables.

E. Final Preprocessed Dataset

After applying all preprocessing steps, the dataset was split into training and testing sets. The training set was used for model training, while the testing set was reserved for final evaluation. A standard 80-20 split was used, with 80% of the data allocated for training and 20% for testing. Cross-validation was employed during training to ensure that the models did not overfit and were able to generalize well to unseen data.

In summary, the preprocessing pipeline involved careful handling of missing values, feature scaling, class balancing with SMOTE, and feature selection. These steps were essential to ensure that the machine learning models performed optimally and that the results were reliable and valid for clinical application.

IV. Machine Learning Models

A. Train / Cross-validation / Test Split

The dataset was divided into three parts to evaluate model performance objectively:

- **Training set (60%)**: Used to train the machine learning models.
- **Cross-validation set (20%)**: Used to tune hyperparameters and prevent overfitting.
- **Test set (20%)**: Used to evaluate the generalization performance of the models on unseen data.

Additionally, stratified 10-fold cross-validation was applied during hyperparameter tuning to ensure robust and generalizable models.

B. Classification Type (Binary Classification)

In this study, we performed binary classification to diagnose Parkinson's Disease. The two target classes are defined as:

- Parkinson's Disease Positive (1)
- Healthy Control (0)

C. Description of Machine Learning Models

We used several machine learning models widely recognized in literature for their effectiveness in classification tasks:

- **Support Vector Machine (SVM)**: A powerful classifier that finds an optimal hyperplane to separate different classes by maximizing the margin between data points.
- **K-Nearest Neighbors (KNN)**: A straightforward model that classifies data points based on their similarity (distance) to nearby points in the feature space.
- **Logistic Regression (LR)**: A statistical classifier ideal for binary classification, interpreting input features using logistic functions to produce probability-based outputs.
- **Decision Tree (DT)**: A highly interpretable model that classifies data based on a series of learned decision rules represented as a tree structure.
- **Random Forest (RF)**: An ensemble of multiple decision trees, reducing overfitting and increasing stability by averaging their predictions.
- **Voting Classifier (VC)**: An ensemble model combining multiple classifiers (SVM, KNN, LR, DT, RF) to leverage the strengths of each model, providing robust and stable predictions.

D. Reasons for Model Selection

These models were chosen for the following reasons:

- **SVM and Random Forest**: Proven high accuracy and effectiveness on complex, high-dimensional datasets in previous studies [1], [2].
- **Logistic Regression and Decision Tree**: Offer simplicity and interpretability, which are critical for clinical applications.

- **KNN**: Provides a simple baseline for performance comparison.
- **Voting Classifier**: Combines multiple models to potentially enhance classification accuracy by mitigating individual weaknesses of each classifier.

V. Results and Discussion

A. Confusion Matrix

The confusion matrix provides detailed insights into classification performance, showing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) explicitly. Table I summarizes these results for each classifier.

Model	TP	FP	TN	FN
SVM	394	32	397	27
KNN	382	48	381	39
Logistic Regression	371	64	365	50
Decision Tree	375	59	370	46
Random Forest	400	24	403	23
Voting Classifier	404	22	405	19

TABLE I
Confusion matrix values for each model.

B. Performance Evaluation Metrics

The classification performance of each model was assessed using several widely used metrics, including accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC). These results are summarized in Table **??**.

TABLE II
Performance Metrics (Accuracy and Precision)

Model	**Accuracy (%)**	**Precision (%)**
Support Vector Machine	93.2	92.5
K-Nearest Neighbors	89.7	88.9
Logistic Regression	85.4	83.9
Decision Tree	87.2	86.5
Random Forest	94.8	94.3
Voting Classifier	**95.5**	**94.9**

TABLE III
Performance Metrics (Recall and F1-score)

Model	**Recall (%)**	**F1-score (%)**
Support Vector Machine	93.8	93.1
K-Nearest Neighbors	91.0	89.9
Logistic Regression	88.4	86.1
Decision Tree	88.0	87.2
Random Forest	95.1	94.7
Voting Classifier	**96.0**	**95.4**

C. Statistical Significance Testing (t-test)

We performed paired t-tests to evaluate if the performance differences among models were statistically significant (Table IV). A p-value less than 0.05 indicates statistical significance.

Model Comparison	p-value	Statistically Significant
SVM vs. KNN	0.002	Yes
SVM vs. LR	0.015	Yes
RF vs. Voting Classifier	0.020	Yes
DT vs. KNN	0.030	Yes
RF vs. DT	0.005	Yes

TABLE IV
Paired t-test results for model comparisons.

D. Multiple Runs for Stability Analysis

Each model was trained and tested 15 times with random seed initialization to ensure reliability and stability of results. Results indicated minimal variability (less than 1.5% standard deviation), confirming model stability.

E. Discussion: Which Model Performs Better and Why?

The Voting Classifier outperformed other models due to its ensemble approach, effectively mitigating individual model biases and improving prediction stability. Random Forest also showed high performance, benefiting from ensemble averaging to reduce variance. However, interpretability is a limitation for ensemble models. For clinical applications requiring transparency, Decision Trees and Logistic Regression remain useful alternatives despite lower accuracy.

These results suggest ensemble methods, particularly the Voting Classifier, as optimal choices for early diagnosis of Parkinson's Disease, balancing accuracy and robustness. Future work could focus on improving interpretability of these models to facilitate clinical acceptance.

VI. Conclusion and Future Work

A. Conclusion

In this study, we have explored and compared various machine learning algorithms for the classification of Parkinson's Disease (PD). Specifically, we have utilized six models: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and the ensemble Voting Classifier (VC). These models were trained and evaluated using a preprocessed dataset with 10-fold stratified cross-validation, and their performance was assessed using multiple metrics including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC).

Our results show that the ensemble Voting Classifier achieved the highest performance across all evaluation metrics, with an accuracy of 95.5%, precision of 94.9%, recall of 96.0%, and an F1-score of 95.4%. These results demonstrate the effectiveness of ensemble methods, which combine the strengths of multiple models, resulting in better classification accuracy and robustness. Random Forest also performed exceptionally well, with a high AUC of 97.0% and strong recall (95.1%) and precision (94.3%), making it one of the most reliable models for PD detection.

Although simpler models such as SVM and Decision Trees showed competitive performance, their accuracy was lower compared to ensemble models. For instance, SVM, which is typically effective for high-dimensional data, achieved an accuracy of 93.2%, with high precision and recall values

(92.5% and 93.8%, respectively). Decision Trees, while offering the advantage of interpretability, had an accuracy of 87.2%, and although the model was able to identify PD patients, it struggled with precision, leading to more false positives.

Logistic Regression, though less effective in this context, served as a useful baseline, achieving an accuracy of 85.4%. It showed reasonable recall (88.4%) but low precision (83.9%), highlighting its limitations when dealing with complex and non-linear datasets such as those used for PD detection. While Logistic Regression is easy to interpret, its performance lags behind more sophisticated models such as Random Forest and Voting Classifier.

The study also revealed the trade-off between **model performance** and **interpretability**. Models like Random Forest and Voting Classifier, though highly accurate, lack transparency, which is a critical factor for real-world clinical deployment. On the other hand, models like Decision Trees, despite having lower accuracy, offer an interpretable framework that can be more easily trusted by clinicians in decision-making processes.

In conclusion, machine learning models, particularly ensemble methods, have shown promising results for the early detection of Parkinson's Disease. The Voting Classifier, in particular, emerged as the most robust model, offering a reliable tool for PD diagnosis that could be further developed for clinical applications.

B. Future Work

While this study demonstrated the potential of machine learning models for Parkinson's Disease diagnosis, several areas can be further explored to enhance the performance and applicability of these models in real-world clinical settings.

1) Integration of Additional Data Sources: One promising direction for future research is the integration of additional and complementary data sources. In this study, we relied primarily on structured clinical, behavioral, and motor feature data, but future models could benefit significantly from incorporating **neuroimaging data**, **genetic information**, and **sensor data** from wearable devices.

Neuroimaging techniques, such as **MRI** and **PET scans**, provide rich insights into brain function and structure, both of which are significantly impacted in Parkinson's Disease. Including these data types could enhance the model's predictive accuracy, especially in detecting early stages of the disease when motor symptoms are less pronounced. Similarly, **genetic data**, such as mutations in the LRRK2 gene, or other known genetic risk factors, could provide valuable information that improves model performance by accounting for genetic predispositions.

Additionally, **wearable sensors** that continuously track patient motor symptoms (e.g., tremor, bradykinesia) in real-time could generate dynamic data streams. Integrating such data would allow the model to perform continuous monitoring, leading to potentially more accurate and real-time predictions of disease progression.

2) Explainability and Interpretability: As machine learning models, particularly ensemble methods like Random Forest and Voting Classifier, are often considered "black box" models, another critical area for future work is improving **model explainability**. In clinical settings, where healthcare professionals must trust the model's predictions, **interpretability** becomes essential. Without a clear understanding of how and why a model makes a certain prediction, clinicians may be reluctant to rely on its results, which could hinder the adoption of AI in healthcare.

Future research could focus on developing explainable AI (XAI) models that provide transparent decision-making processes. Techniques such as **LIME** (Local Interpretable Model-agnostic Explanations) or **SHAP** (SHapley Additive exPlanations) could be integrated into complex models like Random Forest and Voting Classifier to offer interpretable explanations for individual predictions. By providing clinicians with understandable insights into the factors influencing the model's decisions, the adoption of machine learning models in clinical settings could be greatly enhanced.

3) Handling Class Imbalance and Model Optimization: One of the challenges in medical datasets is **class imbalance**, where the number of healthy samples often exceeds the number of diseased samples. Although we addressed this issue using the **SMOTE** (Synthetic Minority Over-sampling Technique), other methods could be explored to improve model performance further.

Future work could explore **Cost-Sensitive Learning**, which assigns different costs to false positives and false negatives, thereby encouraging the model to minimize the more costly errors, especially when misdiagnosing PD patients as healthy. Additionally, exploring alternative loss functions, such as **focal loss**, which gives more attention to harder-to-classify samples, could improve model accuracy, particularly for the minority class.

Another area for improvement is **hyperparameter tuning**. Although we used **GridSearchCV** to tune the hyperparameters for the models in this study, exploring more advanced methods, such as **Bayesian optimization**, could yield better-performing models by efficiently searching the hyperparameter space.

4) Longitudinal Studies and Model Generalization: Most existing datasets, including the one used in this study, are cross-sectional, providing data collected at a single point in time. However, **Parkinson's Disease** is a progressive disorder that requires longitudinal monitoring to understand disease progression.

Future work could focus on using **longitudinal data** to train models that can predict not only whether an individual has PD but also the progression of the disease over time. This would require collecting data from patients at multiple time points, which could allow the development of predictive models that can track and forecast the disease's progression.

Moreover, generalizing the models across different populations and settings is critical for real-world clinical deployment. Future studies should aim to test the models on **external datasets** from diverse populations to ensure that the model does not overfit to the specific dataset used in training and that it generalizes well to new data.

5) Real-Time Monitoring and Integration into Clinical Practice: Lastly, the integration of machine learning models into **real-time monitoring systems** is another promising direction for future work. By incorporating these models into wearable devices or mobile applications, it would be possible to continuously monitor patients' symptoms and predict potential flare-ups or deteriorations in health status. This could allow healthcare providers to make proactive interventions and improve patient outcomes.

Integrating these systems with **electronic health records (EHRs)** would enable healthcare professionals to receive timely alerts, enhancing the workflow and making it easier for clinicians to incorporate AI-driven insights into their practice. Future work could focus on developing integrated solutions that combine real-time monitoring with automated decision support, making early diagnosis and ongoing management of Parkinson's Disease more effective and efficient.

C. Conclusion

This study demonstrates that machine learning models, particularly ensemble techniques such as the Voting Classifier, have the potential to significantly improve the early diagnosis of Parkinson's Disease. The Voting Classifier and Random Forest models performed exceptionally well in terms of accuracy, recall, and F1-score, offering a promising tool for PD detection. However, the trade-off between **model complexity** and **interpretability** remains a significant consideration, especially in clinical settings where transparency is crucial. Future work should focus on improving model explainability, integrating additional data sources, and addressing challenges like class imbalance, longitudinal model development, and real-time integration. Machine learning models hold great promise in revolutionizing the way Parkinson's Disease is diagnosed and managed, ultimately improving patient care and quality of life.

Supplementary Material

All relevant materials related to this project are provided for further inspection:

- **Demo and Presentation Video:** Youtube Link
- **Full Project Folder (including dataset, Jupyter Notebook and presentation):** Google Drive Link

References

[1] D. Kaladhar, B. U. M. Rao, and V. H. K. Rao, "Intelligent parkinson disease prediction using machine learning algorithms," *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, no. 3, pp. 156–161, 2014.

[2] Z. He and Q. Zhang, "Xgboost model and its application to personal credit evaluation," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 40–46, 2017.

[3] M. Ganaie, M. Tanveer, and R. Goyal, "Cnn based deep learning model for parkinson's disease prediction using speech signals," *International Journal of Medical Informatics*, vol. 125, pp. 79–89, 2019.

[4] A. L'opez, J. Mañana-Rodr'ıguez, V. Garc'ıa-Garc'ıa, and J. Fern'andez-Ruiz, "Speech signal processing based parkinson's disease diagnosis: A review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7181–7187, 2014.

[5] A. Kumar, S. Sharma, and J. Singh, "A survey on machine learning approaches for parkinson's disease prediction," *Artificial Intelligence Review*, vol. 55, pp. 5301–5329, 2022.