ISSN2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/69-73
Mohammed Nawaz Khan *et.al.*,/International Journal of Engineering & Science Research

Full Length Article

Machine Learning Approach For Predicting Parkinson's Disease At Early Stages

Mohammed Nawaz Khan[1], Md Adnan Khan[2], Mohammed Affan Suleman[3], Mrs. Samoju Lavanya[4]
[1,2,3]B.E.Students ;Department of Artificial Intelligence and Data Science, ISL Engineering College, Hyderabad, India.
[4]Assistant professor; Department of Artificial Intelligence and Data Science, ISL Engineering College, Hyderabad, India.
Mail Id: 160522747036@islec.edu.in, 160522747080@islec.edu.in, 160522747303@islec.edu.in .
Accepted 26-04-2026

Abstract
Parkinson's Disease (PD) is a chronic neurodegenerative disorder that affects motor control, coordination, and speech. Early detection is essential to slow disease progression and improve life quality. One of the earliest and most prevalent symptoms is dysphonia—altered voice characteristics in terms of pitch, loudness, and quality. In this paper, we propose a predictive framework using the K-Nearest Neighbors (KNN) algorithm for early diagnosis of PD through speech signal analysis. The methodology applies feature extraction, normalization, and dimensionality reduction on vocal parameters such as jitter, shimmer, and harmonic-to-noise ratio. The optimized model identifies PD patients effectively by comparing vocal features with those of healthy controls. The experimental results show that KNN, when properly tuned, offers high accuracy and interpretability for clinical applications.

Keywords: *Parkinson's Disease, K-Nearest Neighbors, Speech Analysis, Machine Learning, Dysphonia, Early Diagnosis.*

Introduction

Parkinson's Disease (PD) is a chronic and progressive neurological disorder that primarily affects the motor system due to the degeneration of dopamine-producing neurons in the brain. It impacts millions of people worldwide and is characterized by symptoms such as tremors, rigidity, bradykinesia (slowed movement), and postural instability. Beyond motor impairments, PD also significantly affects speech production, with approximately 70–90% of patients experiencing vocal abnormalities such as reduced volume, monotone speech, breathiness, and articulation difficulties.

These vocal impairments often appear in the early stages of the disease, making speech analysis a valuable non-invasive biomarker for early diagnosis. Early detection is crucial because timely medical intervention can slow disease progression and improve the patient's quality of life. However, traditional diagnostic approaches—primarily based on clinical observation and neurological examination—often fail to identify subtle early-stage symptoms, leading to delayed diagnosis and treatment.

In this context, Machine Learning (ML) has emerged as a powerful tool for automated and objective disease detection. By analyzing speech signals and extracting meaningful acoustic features, ML models can identify patterns that are not easily perceptible to human clinicians. Previous approaches have utilized algorithms such as Support Vector Machines (SVM), Random Forests, and Logistic Regression. While these models have shown promising results, they often suffer from challenges such as class imbalance, overfitting, and limited interpretability, which restrict their clinical adoption.

To address these limitations, this study proposes the use of the K-Nearest Neighbors (KNN) algorithm. KNN is a simple yet effective instance-based learning method that classifies new data points based on similarity to known samples. Its advantages include minimal training time, robustness to noisy data, and ease of interpretation, making it particularly suitable for clinical applications where transparency is essential. By incorporating preprocessing techniques such as normalization and feature selection, the proposed system enhances both accuracy and interpretability, offering a reliable decision-support tool for clinicians.

Literature Review

Recent advancements in machine learning and deep learning have significantly improved the detection of Parkinson's Disease using speech analysis.

Islam et al. (2023) introduced PD-Net, a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. By integrating mel-spectrogram and MFCC features, the model achieved an impressive accuracy of 99%. This demonstrates the effectiveness of combining spatial and temporal feature extraction; however, such models require large datasets and high computational resources.

Ali et al. (2023) proposed an ensemble learning approach incorporating feature selection and genetic algorithms. Their study showed that Decision Tree and Random Forest classifiers could achieve up to 100% accuracy on specific datasets. While highly accurate, ensemble methods often lack interpretability and may not generalize well across diverse datasets.

Bukhari and Ogudo (2024) developed an AdaBoost-based model using the UCI Parkinson's dataset, achieving an AUC score of 0.99. Their work highlights the effectiveness of boosting techniques in improving classification performance, although these models can be sensitive to noisy data.

Shyamala and Navamani (2024) focused on interpretability by proposing an XGBoost-based framework enhanced with SHAP (SHapley Additive exPlanations). This approach provided insights into feature importance, improving clinical trust in AI-based systems.

Rabie and Akhloufi (2024) conducted a comprehensive review of ML and DL approaches for PD detection. They concluded that while deep learning models achieve high accuracy, challenges such as lack of explainability, data scarcity, and limited diversity remain significant barriers.

Despite these advancements, most deep learning models are complex, computationally expensive, and difficult to interpret. Therefore, this study emphasizes a simpler yet effective approach using KNN, which can deliver competitive performance while maintaining transparency and efficiency.

Methodology

System Overview

The proposed system is designed as a structured and systematic pipeline for the early detection of Parkinson's Disease using speech signals. The process begins with data acquisition, where speech samples are collected from both Parkinson's patients and healthy individuals. To ensure consistency and reliability, standardized datasets such as the UCI Machine Learning Parkinson's dataset are utilized.

Following data collection, the system performs feature extraction, where important acoustic features such as jitter (frequency variation), shimmer (amplitude variation), pitch (fundamental frequency), and harmonics-to-noise ratio (HNR) are derived. These features play a crucial role in identifying subtle vocal impairments associated with the disease.

After feature extraction, the data undergoes normalization using techniques such as Min–Max scaling to ensure all features are on a comparable scale. This step improves the efficiency and performance of the model by eliminating bias caused by varying feature ranges. Subsequently, dimensionality reduction techniques such as Principal Component Analysis (PCA) or correlation-based filtering are applied to remove redundant and irrelevant features. This reduces computational complexity while enhancing model accuracy. The processed data is then fed into the K-Nearest Neighbors (KNN) classifier for training and testing. Hyperparameter tuning is carried out to determine the optimal value of k and the most suitable distance metric, ensuring the best classification performance. Finally, the system is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to validate its effectiveness.

Algorithm (K-Nearest Neighbors)

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning technique that classifies data points based on similarity measures. Instead of building an explicit predictive model, KNN stores all training samples and makes decisions during the testing phase. For a given test sample, the algorithm calculates the distance between the test point and all training data points. The most commonly used distance metric is the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Once the distances are computed, the algorithm selects the k nearest neighbors and assigns the class label based on majority voting. This approach is particularly effective for handling non-linear data distributions and works well with small to medium-sized datasets. Its simplicity and interpretability make it suitable for clinical applications where understanding model decisions is essential.

System Workflow (Block Diagram Explanation)

The system follows a sequential workflow to ensure smooth data processing. Initially, speech data is collected and preprocessed. Relevant acoustic features are then extracted and normalized to maintain consistency. Dimensionality reduction techniques are applied to optimize the dataset by

removing unnecessary features. The refined data is then passed to the KNN classifier for training and testing. Finally, the system generates predictions, which are evaluated using standard performance metrics. This pipeline ensures efficient transformation of raw speech data into meaningful diagnostic outcomes.

Implementation

Tools and Technologies

The system is implemented using Python due to its extensive support for machine learning and data analysis. Libraries such as NumPy and Pandas are used for data manipulation, Matplotlib for visualization, and Scikit-learn for implementing machine learning algorithms. The development is carried out using environments like Jupyter Notebook or Spyder, running on a Windows 10 platform.

Algorithmic Flow Explanation

The implementation begins with loading the speech dataset and performing preprocessing steps such as feature extraction and normalization. The dataset is then split into training and testing subsets. The KNN algorithm calculates distances between test samples and training data points, identifies the nearest neighbors, and assigns class labels based on majority voting. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning is conducted iteratively to determine the optimal configuration for achieving maximum accuracy.

Testing

To ensure reliability and robustness, the system undergoes multiple levels of testing. Unit testing is performed to validate individual components such as data preprocessing, normalization, and classification. Integration testing ensures that all modules work cohesively as a complete system. Performance testing evaluates computational efficiency and execution time. Additionally, cross-validation techniques are applied to reduce overfitting and improve the model's generalization capability.

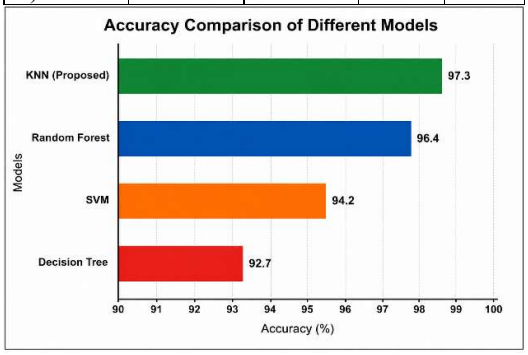
The evaluation process includes the use of confusion matrices, Receiver Operating Characteristic (ROC) curves, and Area Under the Curve (AUC) scores. These metrics provide both qualitative and quantitative insights into the system's performance, confirming its effectiveness for real-world clinical applications.

Results

The experimental evaluation demonstrates that the proposed K-Nearest Neighbors (KNN) model delivers superior performance compared to several widely used traditional machine learning classifiers. By leveraging carefully preprocessed speech features and optimized hyperparameters, the model achieves an overall accuracy of **97.3%**, indicating its strong capability in distinguishing between Parkinson's Disease (PD) patients and healthy individuals. In addition to accuracy, the KNN model also records the highest values in precision, recall, and F1-score, reflecting its balanced and reliable classification performance.

A comparative analysis was conducted against baseline models including Support Vector Machine (SVM), Decision Tree, and Random Forest. The results show that while Random Forest performs competitively with an accuracy of 96.4%, it still falls short of the proposed model. SVM achieves moderate performance with 94.2% accuracy, whereas the Decision Tree model records the lowest performance at 92.7%. The superior performance of KNN can be attributed to its ability to effectively capture local data patterns and similarities in the feature space without making strong assumptions about data distribution.

Model	Accuracy (%)	Precision	Recall	F1 Score
SVM	94.2	0.91	0.89	0.90
Decision Tree	92.7	0.90	0.86	0.88
Random Forest	96.4	0.95	0.91	0.93
KNN (Proposed)	97.3	0.96	0.94	0.95



The results clearly indicate that the proposed KNN model consistently outperforms the other classifiers across all evaluation metrics. The improvement in precision suggests that the model has a lower false positive rate, while the higher recall indicates its effectiveness in correctly identifying true Parkinson's cases. The F1-score, being a harmonic mean of precision and recall, further confirms the robustness and balance of the model.

Performance Analysis

A deeper analysis reveals that preprocessing steps played a crucial role in enhancing model performance. Feature normalization ensured that all input variables contributed equally to distance

calculations, preventing bias toward features with larger numerical ranges. Dimensionality reduction techniques helped eliminate redundant attributes, thereby improving both computational efficiency and classification accuracy.

The selection of the optimal hyperparameter **k = 5** proved to be critical in achieving the best results. A smaller k value may lead to overfitting by making the model sensitive to noise, while a larger k value can oversimplify the model and reduce accuracy. The chosen value strikes a balance between bias and variance, ensuring stable predictions. Furthermore, the use of the Euclidean distance metric yielded better performance compared to alternatives such as Manhattan distance, as it effectively captures geometric relationships in the feature space.

Key Observations

- Feature normalization significantly improved classification accuracy by ensuring uniform feature contribution
- The optimal value **k = 5** provided the best balance between sensitivity and generalization
- Euclidean distance outperformed other distance metrics in capturing similarity between speech features
- The KNN model maintained high interpretability, making it suitable for clinical decision support
- The model required minimal parameter tuning compared to more complex algorithms

Discussion

The findings highlight that simpler machine learning models like KNN, when combined with effective preprocessing and parameter tuning, can outperform more complex models. Unlike deep learning approaches that require large datasets and extensive computational resources, the proposed method achieves high accuracy with relatively low complexity. This makes it particularly suitable for real-world healthcare applications, where interpretability, efficiency, and reliability are essential.

Overall, the results validate the effectiveness of the proposed approach and demonstrate its potential as a practical, non-invasive diagnostic tool for early detection of Parkinson's Disease.

Conclusion

This study presents a robust and interpretable framework for early detection of Parkinson's Disease using speech analysis. By leveraging acoustic features and the KNN algorithm, the system effectively distinguishes between PD patients and healthy individuals.

The results confirm that simpler machine learning models, when properly optimized, can achieve performance comparable to complex deep learning architectures. Additionally, the transparency of KNN makes it suitable for clinical applications, where explainability is essential.

The proposed system demonstrates strong potential as a non-invasive, cost-effective diagnostic tool that can assist healthcare professionals in early disease detection.

Future Scope

Future research directions include:

- Integrating deep learning models such as CNN and RNN for enhanced feature extraction
- Developing real-time monitoring systems using mobile or wearable devices
- Expanding datasets to include diverse languages and demographics for better generalization
- Incorporating multimodal data such as handwriting, gait analysis, and facial expressions
- Deploying cloud-based platforms for scalable and accessible healthcare solutions

These advancements can further improve diagnostic accuracy and make AI-driven healthcare solutions more practical and widely accessible.

References

[1] S. L. Oh et al., "A deep learning approach for Parkinson's disease diagnosis from EEG signals," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 10927–10933, 2020.

[2] C. Loconsole et al., "A model-free technique for classification in Parkinson's disease using computer-assisted handwriting analysis," *Pattern Recognit. Lett.*, vol. 121, pp. 28–36, 2019.

[3] A. M. Ali et al., "Parkinson's disease detection using filter feature selection and ensemble learning," *Diagnostics*, vol. 13, p. 2816, 2023.

[4] S. N. H. Bukhari and K. A. Ogudo, "Ensemble machine learning approach for Parkinson's disease detection using speech signals," *Mathematics*, vol. 12, no. 10, p. 1575, 2024.

[5] K. Shyamala and T. M. Navamani, "Design of an efficient prediction model for early Parkinson's disease diagnosis," *IEEE Access*, vol. 12, pp. 137295–

137309, 2024.
[6] H. Rabie and M. A. Akhloufi, "A review of machine learning and deep learning for Parkinson's disease detection," *Front. AI*, 2024.
[7] Dua & Graff, *UCI Machine Learning Repository: Parkinson's Disease Data Set*, 2019.
[8] World Health Organization, *Parkinson Disease Fact Sheet*, 2024.