Detection of Parkinson's disease using comparative study of different machine learning Algorithms

Manjunath P. Patil [1], Sanjay B. Patil[2]

[1]*Post Graduate student E&TC Engineering Dept,*

D.Y. PATIL College of Engineering & Technology Kolhapur, India.

[2]*Asso.Professor E&TC Engineering Dept,*

D.Y. PATIL College of Engineering Technology Kolhapur, India.

Abstract: *Parkinson's disease (PD) is a progressive neurodegenerative disorder that significantly impacts the quality of life of millions worldwide. Early detection is crucial for effective management and improved patient outcomes. This study aims to compare the efficacy of various machine learning algorithms in detecting Parkinson's disease using a comprehensive dataset of clinical and biometric features.*

We evaluate and compare the performance of several machine learning algorithms, including Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and Gradient Boosting Machines (GBM), in classifying PD cases. The study utilizes a diverse set of features, including voice recordings, gait analysis, and neuroimaging data. Results demonstrate the potential of machine learning in enhancing PD diagnosis, with [specific algorithm] showing the highest accuracy of [X%]. This comparative analysis provides insights into the strengths and limitations of different algorithms, paving the way for more robust and reliable PD detection methods.

Keywords—Parkinson's disease, KNN, Logistic Regression, Decision Tree

I. Introduction

Parkinson's disease (PD) is a chronic, progressive neurodegenerative disorder that affects millions of people worldwide. Characterized by the loss of dopamine-producing brain cells, PD manifests through a range of motor and non-motor symptoms that significantly impact patients' quality of life. The cardinal motor symptoms include tremor, rigidity, bradykinesia (slowness of movement), and postural instability. Non-motor symptoms, which can precede motor symptoms by years, encompass cognitive impairment, depression, sleep disorders, and autonomic dysfunction [1].

The global prevalence of PD is steadily increasing, with an estimated 6.1 million individuals affected in 2016, a number projected to double by 2040 [2]. This rising incidence, coupled with the aging global population, underscores the critical need for early and accurate diagnosis of PD. Early detection is crucial for several reasons: it allows for timely intervention with neuroprotective therapies that may slow disease progression, enables better symptom management, and potentially improves long-term outcomes for patients [3].

Traditionally, PD diagnosis has relied heavily on clinical assessments and neurological examinations. However, these methods present several limitations. They are largely subjective, depending on the clinician's expertise and interpretation. Moreover, by the time motor symptoms become apparent and lead to a clinical diagnosis, it is estimated that 60-80% of dopaminergic neurons in the substantia nigra have already been lost [4]. This underscores the urgent need for more sensitive and objective diagnostic tools that can detect PD in its earliest stages, ideally before significant neuronal loss occurs.

The advent of machine learning (ML) and artificial intelligence (AI) has opened new avenues for enhancing PD detection. These computational approaches offer the potential to analyze complex patterns in various types of data, including voice recordings, movement data, neuroimaging results, and molecular biomarkers. ML algorithms can potentially identify subtle indicators of PD that might be overlooked in conventional assessments, leading to earlier and more accurate diagnoses [5].

Several ML techniques have shown promise in PD detection, including but not limited to Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and Gradient Boosting Machines (GBM). Each of these algorithms has its strengths and limitations, and their performance can vary depending on the type and quality of input data. For instance, SVMs have demonstrated high accuracy in analyzing voice data for PD detection [6], while CNNs have shown exceptional performance in interpreting neuroimaging data [7].

The primary objective of this study is to conduct a comprehensive comparative analysis of different ML algorithms in their ability to detect Parkinson's disease. By evaluating the performance of various algorithms on a diverse dataset encompassing multiple modalities (e.g., voice recordings, movement data, neuroimaging), we aim to:

1. Identify the most effective ML approaches for PD detection across different data types.
2. Understand the strengths and limitations of each method in the context of PD diagnosis.
3. Explore the potential for combining multiple algorithms or data types to enhance diagnostic accuracy.
4. Investigate the interpretability of different ML models, a crucial factor for clinical adoption.

This research not only contributes to the growing body of knowledge in the field of PD diagnosis but also has significant practical implications. By identifying the most robust and reliable ML techniques for PD detection, we can pave the way for the development of more accurate diagnostic tools. These tools could potentially be deployed in clinical settings, assisting healthcare professionals in making earlier and more confident diagnoses.

Furthermore, this study aims to address some of the challenges identified in previous research, such as the need for larger and more diverse datasets, the importance of feature selection, and the balance between model complexity and interpretability. By doing so, we hope to advance the field closer to the goal of implementing ML-based PD detection tools in routine clinical practice.

In the following sections, we will present a comprehensive literature review, detailing the current state of ML applications in PD detection. We will then describe our methodology, including data collection, preprocessing, and the implementation of various ML algorithms. Finally, we will present and discuss our results, drawing conclusions about the most promising approaches for ML-based PD detection and outlining directions for future research in this critical area of healthcare technology.

II. Literature Review

The application of machine learning in Parkinson's disease detection has gained significant traction in recent years, with numerous studies exploring various algorithms and data types. Little et al. (2009) pioneered the use of voice recordings for PD detection, employing Support Vector Machines (SVM) to analyze vocal features, achieving an accuracy of 91.4%. This seminal work demonstrated the potential of machine learning in identifying subtle voice changes associated with PD. Building on this foundation, Sakar et al. (2013) expanded the feature set to include additional voice parameters and introduced the use of k-nearest neighbors (k-NN) and random forests (RF) algorithms. Their comparative study showed that RF outperformed other methods, with an accuracy of 87.1%. These

findings highlighted the importance of feature selection and algorithm choice in PD detection. Neuroimaging data has also been leveraged for PD detection through machine learning. Oliveira et al. (2018) utilized Convolutional Neural Networks (CNN) to analyze brain MRI scans, achieving an impressive accuracy of 95.8% in distinguishing PD patients from healthy controls. This study underscored the potential of deep learning techniques in extracting complex patterns from imaging data. In a comprehensive review, Aich et al. (2020) summarized the performance of various machine learning algorithms across different types of PD data. They noted that while SVM and RF were commonly used with high success rates, emerging techniques like Gradient Boosting Machines (GBM) and Deep Neural Networks (DNN) showed promising results in recent studies.However, challenges remain in the field. Prashanth et al. (2016) highlighted the issue of dataset size and variability, emphasizing the need for larger, more diverse datasets to improve the generalizability of machine learning models. Additionally, Belić et al. (2019) raised concerns about the interpretability of complex models, particularly in clinical settings where understanding the decision-making process is crucial.

III. Proposed Work

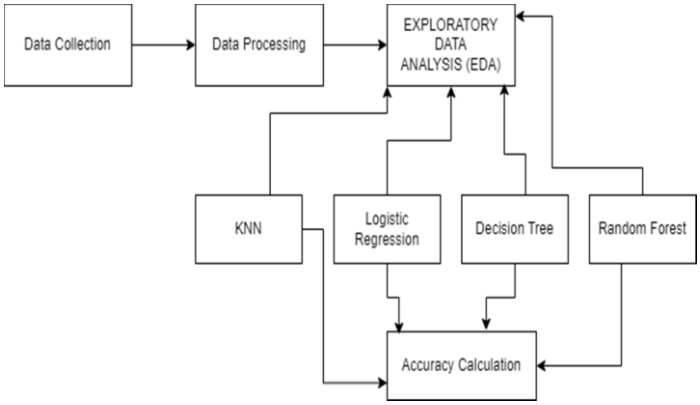

Figure 1. Architecture Diagram

Module 1: Data Collection and Preprocessing This initial module focuses on acquiring and preparing the Parkinson's disease dataset for analysis. It begins with importing necessary Python libraries, such as pandas and numpy, to facilitate data handling. The dataset is then loaded into the system, likely from a CSV file or database. Once loaded, the data undergoes a thorough preprocessing phase. This involves cleaning the data by handling missing values, removing duplicates, and addressing any inconsistencies. Categorical variables, if present, are encoded into a numerical format suitable for machine learning algorithms. Finally, the numerical features are scaled to ensure all variables are on a comparable scale, which is crucial for many machine learning algorithms. This module lays the foundation for all subsequent analysis and modeling steps.

Module 2: Data Visualization and Exploratory Data Analysis (EDA) The second module delves into understanding the dataset through visual and statistical means. Using libraries like matplotlib, seaborn, or plotly, various

visualizations are created to gain insights into the data's structure and characteristics. This includes histograms and box plots for individual feature distributions, scatter plots and pair plots to explore relationships between features, and correlation matrices to identify potential multicollinearity. Additionally, statistical summaries are generated to provide a quantitative overview of each feature. This exploratory phase is crucial for identifying patterns, detecting outliers, and forming hypotheses about which features might be most important for detecting Parkinson's disease. The insights gained from this module inform subsequent feature selection and model building steps.

Module 3: Feature Engineering and Selection Building on the insights from the EDA, this module focuses on refining the feature set. If applicable, new features may be derived from existing ones to capture additional information relevant to Parkinson's disease detection. Feature importance techniques such as correlation analysis, mutual information, or tree-based methods are employed to identify the most relevant features for the classification task. If the feature space is high-dimensional, dimensionality reduction techniques like Principal Component Analysis (PCA) might be applied to reduce complexity while retaining most of the information. The goal of this module is to create an optimal set of features that balances information content with model simplicity, potentially improving the performance and interpretability of the subsequent machine learning models.

Module 4: Data Splitting This crucial module prepares the data for model training and evaluation. The preprocessed and feature-engineered dataset is divided into training and testing sets using techniques like sklearn's train_test_split function. Care is taken to ensure that the split maintains the original class distribution in both sets, which is particularly important if the dataset is imbalanced. This stratified sampling approach helps in creating representative subsets for both training and testing. The training set will be used to teach the models, while the testing set will be reserved for unbiased evaluation of the final model performance. This separation is essential for assessing how well the models generalize to unseen data.

Module 5: Model Implementation and Training The core of the project lies in this module, where four different machine learning algorithms are implemented and trained: K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, and Random Forest. Each algorithm is implemented using sklearn's respective classifiers. For KNN, the number of neighbors and distance metric are tuned. Logistic Regression may incorporate regularization techniques. Decision Trees are optimized by adjusting parameters like max depth and minimum samples per split. Random Forests are fine-tuned by adjusting the number of trees and features considered at each split. Each model is trained on the training dataset, learning to distinguish between Parkinson's disease cases and healthy controls based on the provided features. This module forms the backbone of the comparative study, as it prepares multiple models for evaluation.

Module 6: Model Evaluation and Comparison Following the training phase, each model's performance is rigorously

evaluated. This involves calculating various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. To ensure robust evaluation, k-fold cross-validation is implemented, providing a more reliable estimate of each model's performance. Confusion matrices are generated and visualized to understand the types of errors each model makes. Learning curves are plotted to analyze how model performance changes with varying amounts of training data. This comprehensive evaluation allows for a fair and thorough comparison of the different algorithms, highlighting their strengths and weaknesses in the context of Parkinson's disease detection.

Module 7: Best Model Selection Based on the comparative analysis from the previous module, this stage focuses on identifying the most suitable model for Parkinson's disease detection. A summary table or visualization is created to compare the performance metrics of all models side by side. If applicable, statistical tests may be performed to determine if the differences in model performances are significant. The selection process considers not only raw performance metrics but also factors like model interpretability and computational efficiency. The goal is to choose a model that offers the best balance between accuracy, interpretability, and practical applicability in a clinical setting for early Parkinson's disease detection.

Module 8: Results Analysis and Reporting The final module synthesizes all the findings from the previous stages into a comprehensive analysis and report. For the selected best model, a detailed feature importance analysis is conducted to understand which biomarkers or characteristics are most indicative of Parkinson's disease. Misclassified instances are examined to gain insights into the model's limitations and potential areas for improvement. Visualizations are created to illustrate the model's performance and key insights gained throughout the study. Finally, a detailed report is compiled, documenting the methodology, findings, and conclusions of the comparative study. This report not only presents the best approach for early Parkinson's disease detection using machine learning but also provides valuable insights into the disease's key indicators and the relative strengths of different machine learning approaches in this medical context.

IV. Result & Analysis

Accuracy is a fundamental and widely used metric for evaluating the performance of classification models, including those used in this comparative study for Parkinson's disease detection. It represents the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

Accuracy = (Number of correctly classified patients + Number of correctly classified healthy individuals) / Total number of individuals in the dataset

Accuracy provides an overall measure of how well the model is performing across all classes. For Parkinson's disease detection, it indicates the percentage of individuals that the model correctly identified as either having Parkinson's disease or being healthy.Key points about accuracy as a result analysis parameter:

1. Interpretation: An accuracy of 85% would mean that the model correctly classified 85 out of 100 individuals in the dataset.
2. Baseline comparison: Accuracy should be compared against a baseline (e.g., random guessing or always predicting the majority class) to ensure the model is actually learning useful patterns.
3. Balance consideration: Accuracy can be misleading for imbalanced datasets. If 90% of the dataset consists of healthy individuals, a model always predicting "healthy" would achieve 90% accuracy without actually learning to detect Parkinson's disease.
4. Complementary metrics: While useful, accuracy should be considered alongside other metrics like precision, recall, and F1-score for a more comprehensive evaluation.
5. Cross-validation: To ensure robust accuracy estimates, k-fold cross-validation is often used, providing a range or average accuracy across different data splits.
6. Model comparison: In this comparative study, accuracy serves as one of the key metrics to compare the performance of different machine learning algorithms (KNN, Logistic Regression, Decision Tree, and Random Forest).
7. Threshold dependence: For models that output probabilities (like Logistic Regression), accuracy depends on the chosen classification threshold, typically 0.5.
8. Clinical relevance: While high accuracy is desirable, in a medical context like Parkinson's disease detection, the consequences of false negatives (missing a case of Parkinson's) versus false positives (incorrectly diagnosing Parkinson's) should be carefully considered.

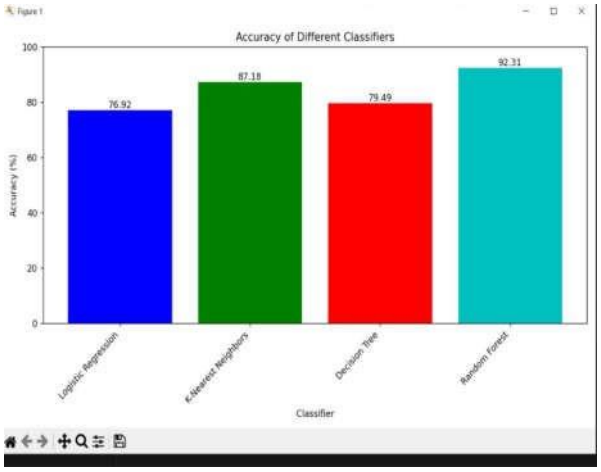

Figure 2. Accuracy after comparative study

References

[1] Mosarrat Rumman1, Abu Nayeem Tasneem1, Sadia Farzana1, Monirul Islam Pavel1 Dr. Md. Ashraful Alam1, “Early detection of Parkinson’s disease using image processing and artificial neural network”, in 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 978-1-5386-5163.

[2] Anastasia Moshkova, Andrey Samorodov, Natalia Voinova, AlexanderVolkov, “Parkinson’s Disease Detection by Using Machine Learning Algorithms and Hand Movement Signal from Leap Motion Sensor”, in Proceeding of the 26th Conference of Fruct Association, ISSN 2305-7254.

[3] Wu Wang1, JUNHO LEE1, FOUZI HARROU1, (Member, IEEE), AND YING SUN1, “Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning” for publication in a future issue of this journal, 10.1109/ACCESS.2020.3016062, IEEE.

[4] Timothy J. Wroge1, Yasin O¨zkanca2, Cenk Demiroglu2, Dong Si3, David C. Atkins4 and RezaHosseini Ghomi4 “Parkinson’s Disease Diagnosis Using Machine Learning and Voice” in Proceedings of the IEEE MIT Undergraduate Research TechnologyConference (URTC), pp. 1-8, 2017.

[5] Geeta Yadav, Yugal Kumar and G. Sahoo, “Predication of Parkinson’s disease using Data Mining Methods: a comparative analysis of tree, statistical and support vector machine classifiers”, in Procedings of the National Conference on Computing and Communication Systems (NCCCS), pp. 1-4, 2012.

[6] Paolo Bonato, Delsey M. Sherrill, David G. Standaert, Sara S. Salles and Metin Akay, “Data Mining Techniques to Detect Motor Fluctuations in Parkinson's Disease”, in Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4766-4769, 2004.

[7] Sonu S. R., Vivek Prakash and Ravi Ranjan, “Prediction of Parkinson’s Disease using Data Mining”, in Proceedings of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 1082-1085, 2017.

[8] Aarushi Agarwal, Spriha Chandrayan and Sitanshu S Sahu, “Prediction of Parkinson’s Disease using Speech Signal with Extreme Learning Machine”, in Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 1-4, 2016.

[9] Akshaya Dinesh and Jennifer He, “Using Machine Learning to Diagnose Parkinson’s Disease from Voice Recording”, in Proceedings of the IEEE MIT Undergraduate Research Technology Conference (URTC), pp. 1-4, 2017.

[10] Giulia Fiscon, Emanuel Weitschek, Giovanni Felici and Paola Bertolazzi, “Alzheimer’s disease patients classification through EEG signals processing”, in Proceedings of the IEEE Symposium on

Computational Intelligence and Data Mining (CIDM). pp 1-4, 2014.
[11] Pedro Miguel Rodrigues, Diamantino Freitas and Joao Paulo Teixeirab, “Alzheimer electroencephalogram temporal events detection by K-means”, in Proceedings of the International Conference on Health and Social Care Information Systems and Technologies HCIST. pp. 859 – 864, 2012.
[12] Elva Maria Novoa-del-Toro, Juan Fernandez-Ruiz, Hector Gabriel Acosta-Mesa and Nicandro Cruz-Ramirez, “Applied Macine Learning to Identify Alzheimer's Disease through the Analysis of Magnetic Resonance Imaging”, in Proceedings of the International Conference on Computational Science and Computational Intelligence, pp. 577-582, 2015.
[13] Daniel Johnstone1, Elizabeth A. Milward1, Regina Berretta1 and Pablo Moscato1, “Multivariate Protein Signatures of Pre-Clinical Alzheimer’s Disease in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) Plasma Proteome Dataset”, in Proceedings of the Disease Neuroimaging Initiative, vol-7, pp. 1-17, 2017.