Advancements in Parkinson's Disease Detection using Machine Learning

Yatin Saluja[1], Ravindra B. V.[2,*], Narendra V. G.[1]

[1]*Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India 576104*
[2]*Manipal School of Information Sciences, Manipal Academy of Higher Education, Manipal, Karnataka, India 576104*
**Corresponding author: ravindra.bv@manipal.edu*

Abstract

This study presents a machine learning-based approach to enhance the detection of Parkinson's disease (PD) using classification models such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naïve Bayes, and Logistic Regression. Leveraging a publicly available dataset, the proposed framework addresses challenges in feature selection, data diversity, and model interpretability. Preprocessing techniques were employed to ensure data quality, including standard scaling and class distribution analysis. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. Among the tested models, KNN achieved the highest accuracy of 90% and F1-score of 0.93, indicating strong diagnostic capability. The system provides healthcare practitioners with interpretable results through visual outputs, enabling informed clinical decision-making. This work demonstrates the potential of accessible, data-driven methods for early PD detection and lays the groundwork for future improvements through longitudinal analysis and multi-modal data integration.

Keywords: Parkinson's disease, machine learning, SVM, KNN classifier, Naïve Bayes, Logistic Regression, feature selection, model evaluation.

River Journal, 22–31.

1 Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily affects motor functions and significantly impacts the quality of life. Early and accurate diagnosis remains a clinical challenge due to overlapping symptoms with other conditions and the absence of definitive biomarkers. Recent advances in machine learning (ML) have shown promise in aiding medical diagnostics by uncovering patterns in clinical and physiological data that may not be evident through traditional methods. This study investigates the application of ML algorithms—Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naïve Bayes, and Logistic Regression—for the early detection of Parkinson's disease. These models are trained on a publicly available dataset consisting of biomedical voice measurements known to be affected by PD. The aim is to evaluate and compare the performance of each model in distinguishing between healthy individuals and PD patients based on extracted features. By integrating statistical analysis and performance metrics such as precision, recall, and F1-score, the research highlights the diagnostic capabilities of each classifier. The work also emphasizes the need for interpretable models in clinical settings and the potential for machine learning to augment healthcare diagnostics. This research contributes to the development of accessible, data-driven tools for improving early Parkinson's detection.

2 Literature Survey

Parkinson's disease (PD) is difficult to diagnose in its early stages, leading to growing interest in machine learning (ML) approaches for its detection and progression analysis. Several studies have demonstrated the effectiveness of ML in identifying PD from biomedical data, particularly voice features. Nilashi et al. [1] evaluated ensemble methods for predicting Unified Parkinson's Disease Rating Scale (UPDRS) scores, noting the challenge of effective feature selection. Radha et al. [2] used MFCC, spectral centroid, and RMS features from speech samples, applying CNN, ANN, and HMM models. ANN achieved the highest accuracy (96.2%). Alalayah et al. [3] improved early PD detection through feature selection (RFE), data balancing (SMOTE), and dimensionality reduction (t-SNE, PCA). Their multilayer perceptron with PCA reached 98% accuracy. Similarly, Shraddha et al. [4] employed SVM and MFCC features on 195 voice samples, achieving 89% accuracy.

García-Ordás et al. [5] combined classification and regression deep learning models to assess PD severity using UPDRS scores, achieving 99.15% classification accuracy. Templeton et al. [6] used tablet-based digital biomarkers and decision trees, achieving 92.6% accuracy in distinguishing PD patients. XGBoost classifiers with MRMR and RFE-based feature selection also performed well in studies by Priyadharshini et al. [7] and Patil et al. [8], both reporting over 92% accuracy. Wroge et al. [9] used mPower voice recordings and decision tree ensembles, obtaining 86% accuracy and suggesting the integration of multiple data modalities. Other studies such as Govindua et al. [10] reported promising results using SVM, KNN, and Random Forests with PCA. Mei et al. [11] conducted a comprehensive review of 209 ML-based PD studies, highlighting limitations in reproducibility and clinical translation. Saeed et al. [12] combined wrapper-based feature selection with KNN to achieve 88.33% accuracy. Varghese et al. [13] showed that support vector regression outperformed other models in predicting PD severity from speech data. Overall, the literature highlights the potential of ML in PD detection but also reveals persistent gaps, such as limited data diversity, inconsistent validation practices, and challenges in model explainability. This study aims to address these through dataset augmentation, optimized feature engineering, interpretable model development, and robust evaluation.

3 Methodology

The dataset used for this study was obtained from Kaggle in `.csv` format. Preliminary exploration involved assessing descriptive statistics and identifying missing values. The `name` column, being non-informative, was dropped. The target variable `status` was encoded as a binary value (0 or 1), and feature scaling was applied using `StandardScaler` to normalize the values—essential for models such as SVM, KNN, Logistic Regression, and Naïve Bayes. For data analysis, a count plot was used to assess class balance. A correlation heatmap was generated to explore feature interdependence (Figure 1), while box plots (Figure 2) highlighted variations in feature distributions between Parkinson-positive and negative cases. These visual tools helped identify patterns and validate the dataset's structure before modeling. Feature selection was not explicitly performed; all relevant features, aside from the dropped `name` column, were retained based on domain knowledge and exploratory data analysis. Four classifiers were trained and evaluated: Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Naïve Bayes, and Logistic Regression.

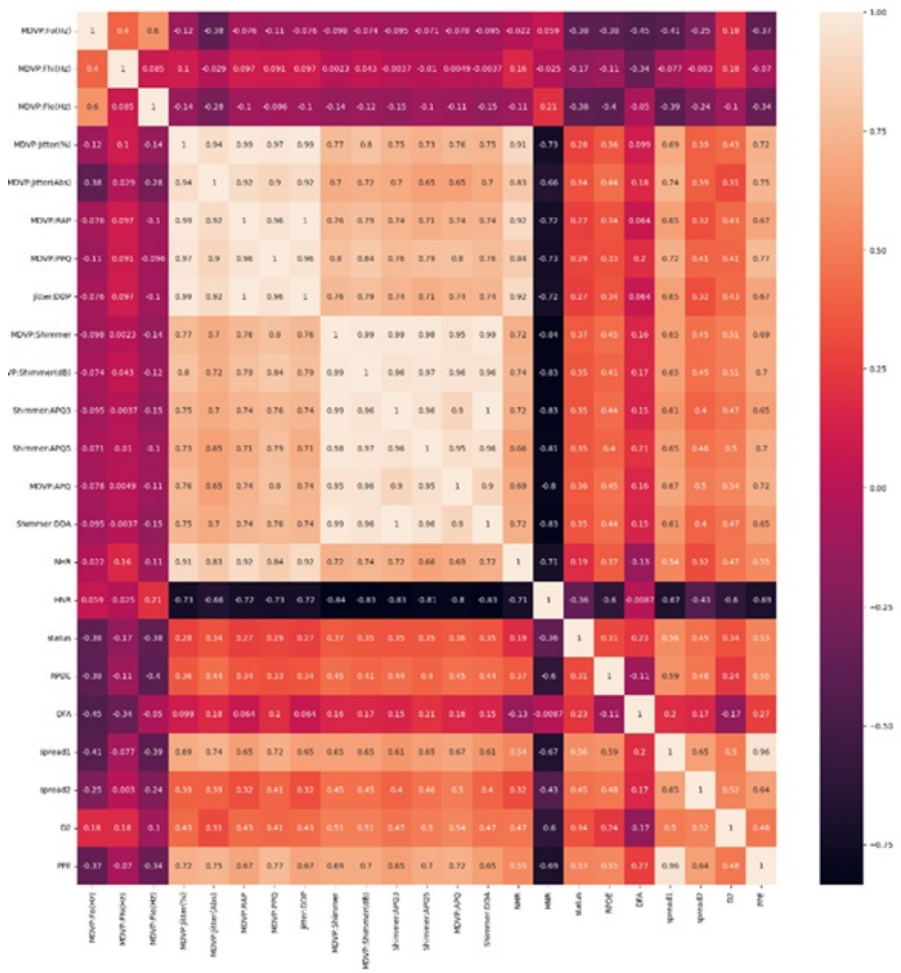


Figure 1 Correlation heatmap of features.

For each model, the dataset was split using `train_test_split`, and models were trained on the standardized training set. The SVM classifier used a linear kernel, while KNN was configured with $k = 5$ neighbors. Naïve Bayes followed the Gaussian distribution assumption, and Logistic Regression was trained with default hyperparameters. Model predictions were evaluated using accuracy, precision, recall, and F1-score on the testing set. Confusion matrices were also generated to visualize classification performance. The evaluation metrics were computed as follows:

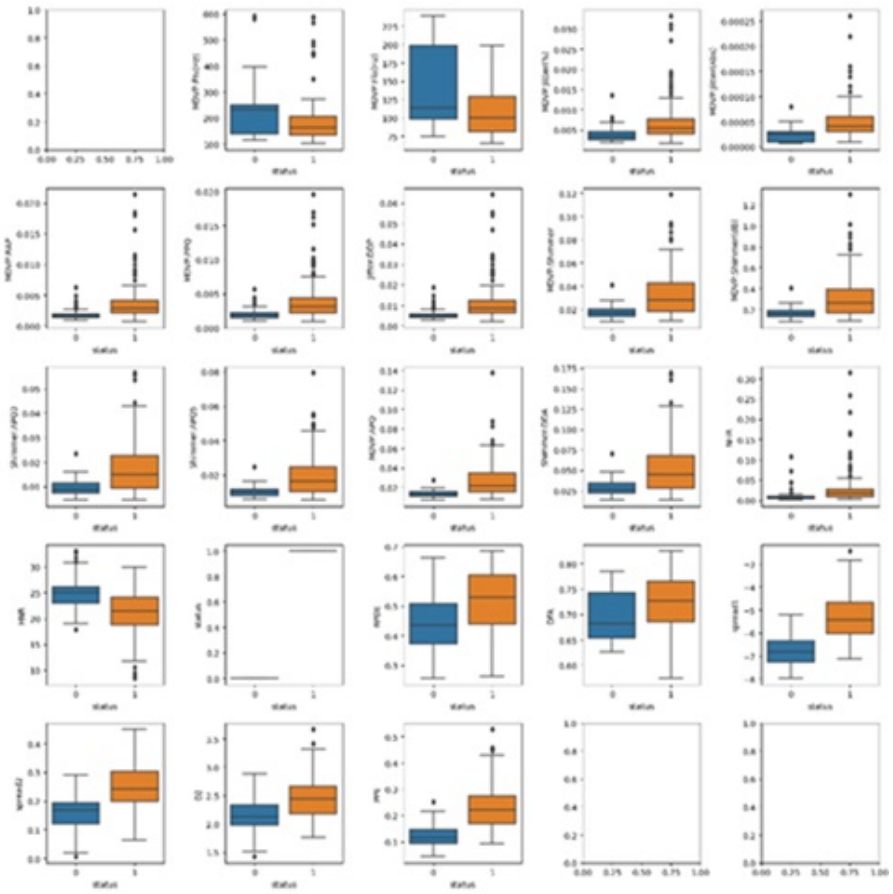


Figure 2 Box plots of selected features by class.

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

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

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

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP is true positives, TN true negatives, FP false positives, and FN false negatives. These metrics offered a comprehensive understanding of each model's ability to generalize and detect Parkinson's disease accurately.

4 Results and Discussion

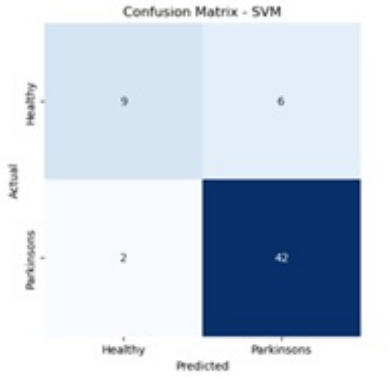
The performance of four machine learning models—SVM, KNN, Logistic Regression, and Naïve Bayes—was evaluated using standard classification metrics. The SVM and Logistic Regression models demonstrated similar performance, with high precision and recall, while the KNN model achieved the highest accuracy overall. Naïve Bayes performed comparatively lower, particularly in recall, indicating challenges in identifying positive cases. The evaluation was carried out on both training and testing datasets, and the results were visualized using confusion matrices (Figures 3(a) to 3(d)) and comparative performance plots (Figure 4).

The classification performance of all four models was visualized using confusion matrices, as shown in Figure 3. These matrices help highlight the distribution of true positives, true negatives, false positives, and false negatives, providing insights into each model's strengths and weaknesses. KNN and SVM models exhibit high true positive rates, while Naïve Bayes shows a higher false negative rate, indicating its limitations in correctly identifying PD-positive cases. The comparative metrics for all models—precision, recall, accuracy, and F1-score—are summarized in Table 1 and visualized in Figure 4.

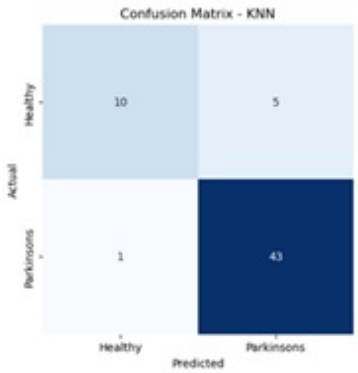
Table 1 Evaluation metrics of all the models

Model	**Precision**	**Recall**	**Accuracy**	**F1 Score**
SVM	0.8750	0.9545	0.8644	0.9130
KNN	0.8958	0.9773	0.9000	0.9348
NB	0.9167	0.7500	0.7627	0.8250
LR	0.8750	0.9545	0.8644	0.9130

The KNN model achieved the best overall performance with 90% accuracy and the highest F1-score (0.9348), indicating a balanced capability to minimize both false positives and false negatives. SVM and Logistic Regression followed closely, showing identical performance with strong recall (0.9545) and precision (0.8750). In contrast, Naïve Bayes, despite having the highest precision (0.9167), had a lower recall (0.7500), suggesting a higher rate of false negatives.

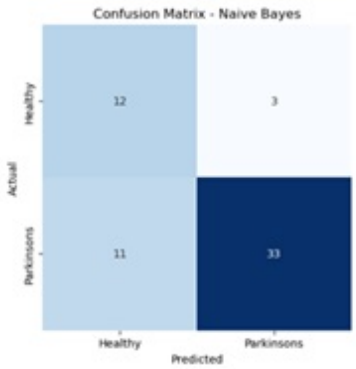


(a) SVM Model



(b) KNN Model

(c) Logistic Regression Model



(d) Naïve Bayes Model

Figure 3 Confusion matrices of all four classification models: (a) SVM, (b) KNN, (c) Logistic Regression, and (d) Naïve Bayes. Each matrix visualizes model performance across predicted and actual classes.

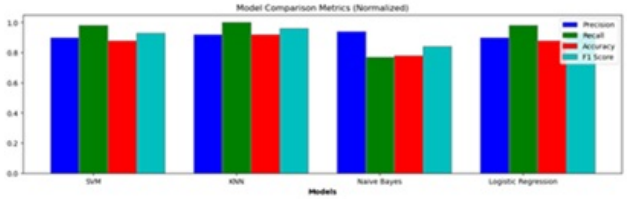


Figure 4 Comparison of evaluation metrics across all four models.

This indicates that while Naïve Bayes is more cautious in predicting positives, it may miss a larger proportion of true PD cases. These findings reaffirm the efficacy of non-linear distance-based models like KNN in handling subtle feature distributions in biomedical data, provided scaling and preprocessing are handled correctly.

5 Conclusion

This study explored the application of machine learning techniques for the early detection of Parkinson's disease (PD), focusing on four classifiers: Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Logistic Regression (LR), and Naïve Bayes (NB). Among these, the KNN model outperformed others with the highest accuracy (90%) and F1-score (0.9348), indicating strong predictive capability and balance between precision and recall. Both SVM and Logistic Regression achieved comparable results, demonstrating high sensitivity and consistent precision. In contrast, Naïve Bayes, although yielding the highest precision (0.9167), suffered from lower recall (0.7500), suggesting limitations in correctly identifying PD-positive cases. These findings validate the use of supervised learning for reliable PD screening and support the development of data-driven tools in clinical diagnostics. To enhance future implementations, the study recommends the integration of longitudinal datasets, multi-modal feature fusion (e.g., voice, motor, handwriting, EEG), and transfer learning techniques. Additionally, efforts should be directed toward real-time deployment and clinical collaboration to improve usability and validation in healthcare settings. By addressing these directions, machine learning-driven PD detection systems can be made more accurate, interpretable, and impactful for early intervention and patient care.

References

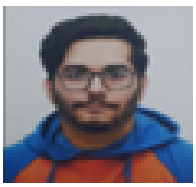
[1] M. Nilashi, R.A. Abumalloh, B. Minaei-Bidgoli, S. Samad, M.Y. Ismail, A. Alhargan, and W.A. Zogaan. Predicting Parkinson's disease progression: Evaluation of ensemble methods in machine learning. *Journal of Healthcare Engineering*, 2022:1–17, 2021.

[2] N. Radha, S.M. R.M., and S. Holy. Parkinson's disease detection using machine learning techniques. *Revista Argentina de Clínica Psicológica*, 30(2):543–552, 2021.

[3] K.M. Alalayah, E.M. Senan, H.F. Atlam, I.A. Ahmed, and H.S.A. Shatnawi. Automatic and early detection of Parkinson's disease by analyzing acoustic signals using classification algorithms based on recursive feature elimination method. *Diagnostics*, 13(11):1924, 2023.

[4] S.V. Bante, S.M. Barhate, and S.J. Sharma. Parkinson's disease detection using machine learning. *Journal of Emerging Technologies and Innovative Research*, 9(6):21–23, 2022.
[5] M.T. García-Ordás, J.A. Benítez-Andrades, J. Aveleira-Mata, J.M. Alija-Pérez, and C. Benavides. Determining the severity of Parkinson's disease in patients using a multi-task neural network. *Multimedia Tools and Applications*, 83:6077–6092, 2024.
[6] J.M. Templeton, C. Poellabauer, and S. Schneider. Classification of Parkinson's disease and its stages using machine learning. *Scientific Reports*, 12:14036, 2022.
[7] G. Priyadharshini, T. Gowtham, M.H. Bhoopathi, M. Reshma, V. Tamilarasi, and P. Nandhini. Detection of Parkinson disease using machine learning. *Engineering and Technology Journal for Research and Innovation*, 4(1):32–38, 2022.
[8] S. Patil, S. Jaybhaye, S. Bokariya, P. Jain, S. Phapale, and T. Hande. Parkinson's disease prediction system in machine learning. In *Proceedings of the First International Conference on Data Science and Advanced Computing (ICDSAC 2023)*, *ITM Web of Conferences*, 56:05002, 2023.
[9] T.J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D.C. Atkins, and R.H. Ghomi. Parkinson's disease diagnosis using machine learning and voice. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7, 2018.
[10] A. Govindua and S. Palwe. Early detection of Parkinson's disease using machine learning. *Procedia Computer Science*, 218:249–261, 2023.
[11] J. Mei, C. Desrosiers, and J. Frasnelli. Machine learning for the diagnosis of Parkinson's disease: A review of literature. *Frontiers in Aging Neuroscience*, 13:633752, 2021.
[12] F. Saeed, M. Al-Sarem, M. Al-Mohaimeed, A. Emara, W. Boulila, M. Alasli, and F. Ghabban. Enhancing Parkinson's disease prediction using machine learning and feature selection methods. *Computers, Materials & Continua*, 71(3):5639–5658, 2022.
[13] B.K. Varghese, G.B.A.D. Amali, and U.D.K.S. Devi. Prediction of Parkinson's disease using machine learning techniques on speech dataset. *Research Journal of Pharmacy and Technology*, 12(2):644–648, 2019.

Biography



Yatin Saluja is an undergraduate student in Computer Science and Engineering (AI-ML) at Manipal Institute of Technology, MAHE. His interests include machine learning, deep learning, and social network analysis.

Dr. Narendra V. G. is a Professor in the Department of Computer Science and Engineering at MIT, MAHE. With over 25 years of academic and research experience, his expertise includes image processing, soft computing, and smart agriculture. He leads a major project on Indian Language Machine Translation.

Dr. Ravindra B. V. is an Assistant Professor (Selection Grade) at the Manipal School of Information Sciences, MAHE. He has 15 years of industry experience and over a decade in academia. His research spans machine learning, NLP, and medical data mining.