International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2511142

Parkinson's Disease Detection Using Machine Learning

CH. Lalitha Haripriya[1], CH. Sirisha[1], A. Gnaneswar[1], K. Nagendra[1], Dr. R. Rajender[2]

[1]B. Tech Students, Department of Computer Science and System Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India

[2]Professor & Hod, Department of Computer Science and System Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted : 21 Feb 2025

Published: 23 Feb 2025

Publication Issue

Volume 11, Issue 1

January-February-2025

Page Number

3150-3154

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting movement and motor control. Early detection is crucial for effective treatment and management. This paper presents a machine learning-based approach to detect Parkinson's disease using speech and biomedical data. The proposed model utilizes various machine learning algorithms, including Support Vector Machines (SVM), Random Forest, and Deep Learning techniques, to classify PD and non-PD subjects. The model is trained on a publicly available dataset and achieves significant accuracy in classification. Multi-modal analysis enhances diagnostic accuracy, offering a non-invasive, cost-effective solution. Future work will focus on real-time monitoring, expanding datasets, integrating wearable technology, and improving model interpretability for clinical applications.

Keywords: Random Forest, Support Vector Machines, Feature Extraction, Clinical Data, Early Detection.

Introduction

The Parkinson's disease (PD) is a common neurological disorder impacting muscle movement in the body. It affects mobility, speech and posture leading to tremors, muscle rigidity and bradykinesia. It occurs due the death of neurons, resulting in a decrease of dopamine levels in the brain. Low levels of dopamine hamper communication between synapses, causing ineffective motor functions. While the progress of symptoms may vary from patient to patient, balance problems and tremors are the most prevalent side-effects of dopaminergic neuron death.

Parkinson's disease progresses in five stages, with 90% of patients exhibiting vocal cord injuries as an early symptom. Vocal impairment is not only easy to measure, but also falls under the category of telemedicine or remote medicine. Patients do not need to visit a doctor physically; instead, they can record their voice using a phone and perform a simple test at home. Common voice modulation symptoms include dysphonia and dysarthria. Patients can be asked to hold a single vowel's pitch for as long as possible, also known as sustained phonation or running speech tests can be, as administered, realistic test of impairment.

Following early detection, doctors can cater therapeutic solutions or deep brain simulation to reactivate the dopamine producing neurons in the brain, thereby slowing the progress of PD. Owing to its complex nature, there is no cure for Parkinson's till date. However, early identification followed by right medication can reduce the tremors and imbalance symptoms in patients, enabling them to lead a normal life.

This paper focuses on early detection through audio recordings of PWP using ML techniques. This novel approach emphasizes the relevance of audio as a non-invasive biomarker to detect PD. Our preliminary results show that Random Forest classifier model has an accuracy of 91.83% when trained on 22 attributes of MDVP audio data, compared to KNN, SVM and Logistic regression models. PWP suffer from mobility issues and are unable to travel for health check-ups. The proposed remote detection technique will provide a new lease of life to patients, as it classifies the severity of PD using speech data, that can be recorded on mobile phones.

Voice frequency plays a crucial role in Parkinson's detection. The fundamental frequency (F0), represented by MDVP:Fo(Hz) in the dataset, indicates the baseline pitch of a person's voice. Individuals with Parkinson's disease tend to have a lower and more monotonous F0, reflecting reduced vocal variability. This reduction in pitch modulation contributes to speech that sounds flatter and less dynamic. The dataset helps analyze the significance of F0 variations in distinguishing PD patients from healthy individuals, highlighting its relevance as a key biomarker in Parkinson's classification.

Another critical aspect of voice analysis is fundamental frequency stability, which provides insights into the overall quality and consistency of speech. Variations in MDVP:Fo(Hz) can indicate irregularities in vocal control, as seen in Parkinson's disease. A lower and more stable F0 often corresponds to reduced pitch modulation, leading to flatter speech. Analyzing these frequency-based attributes helps assess vocal instability and differentiate between Parkinson's patients and healthy individuals.

In Parkinson's patients, variations in fundamental frequency (MDVP:Fo(Hz)) reflect speech impairments such as reduced pitch variation and monotony. A lower F0 is associated with weakened vocal control, making speech sound more monotonous and less dynamic. Additionally, Parkinson's patients may experience instability in vocal frequency, leading to inconsistent speech patterns. These irregularities contribute to overall speech degradation, making voice analysis a valuable tool for PD detection.

Proposed Methodology

The dataset, collected from PPMI and UCI, includes attributes such as age, eye blinking rate, handshaking presence, and fundamental frequency (MDVP:Fo). It is preprocessed, analyzed, and visualized for attribute significance.Four machine learning models—Logistic Regression, SVM, Random Forest, and K-Nearest Neighbors—are trained on 75% of the data to classify individuals as Parkinson's patients or healthy based on these features. The study also examines the impact of class imbalance and the relevance of selected attributes in Parkinson's disease classification.

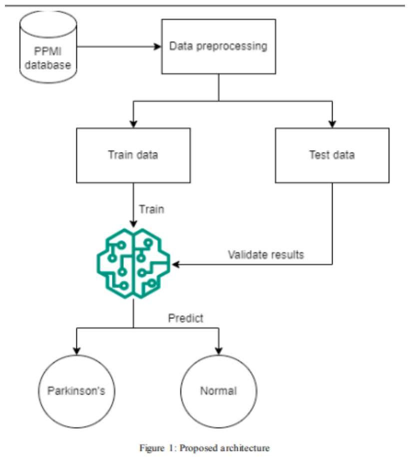

Figure 1 : Architecture of Parkinson's Disease Detection

Important The given diagram represents the proposed architecture for Parkinson's disease detection using a machine learning approach. The process begins with data acquisition from the PPMI (Parkinson’s Progression Markers Initiative) database, followed by a data preprocessing stage to prepare the dataset. The preprocessed data is then split into training and testing datasets. The training dataset is used to train a machine learning model, while the test dataset is used to validate the model's performance. Once trained, the model predicts whether a given sample corresponds to a person with Parkinson’s disease or a normal individual. The overall workflow ensures systematic learning and evaluation to improve the model’s accuracy in detecting Parkinson’s disease.

The research paper aims to identify the most relevant attributes in classification of PD and impact of imbalance in medical data in classification.

Dataset

Biomedical The dataset provided contains 171 records with six attributes related to Parkinson’s disease detection. It includes the age of individuals, eye blinking frequency, handshaking (tremors), and MDVP: Fo (Hz), which represents the fundamental frequency of voice—a key parameter in diagnosing Parkinson’s disease. The result column is a binary indicator where '1' signifies the presence of Parkinson’s disease and '0' indicates a healthy individual. The name column is entirely empty and can be disregarded. This dataset appears to be structured for classification purposes, likely aimed at identifying Parkinson’s disease based on a combination of vocal and physical symptoms. Further analysis could involve data visualization, feature importance evaluation, or even training a machine learning model to predict Parkinson’s disease based on these attributes.

The table in the image presents a list of attributes and their corresponding purposes, specifically related to Parkinson's disease detection. It include various vocal measurements derived from the MDVP (Multi-Dimensional Voice Program) analysis which are used to assess the characteristics of speech affected by Parkinson’s disease.

	A	B	C	D	E	F
1	age	eye_blinki	handshaki	MDVP:Fo(I	result	name
2	48	1	1	119.992	1	
3	46	1	0	122.4	0	
4	50	1	0	120.552	0	
5	49	1	1	95.73	1	
6	45	0	0	95.056	0	
7	45	0	1	91.904	0	
8	46	1	1	139.173	1	
9	50	1	0	152.845	0	
10	45	1	0	144.188	0	
11	48	0	0	168.778	0	
12	47	0	0	153.046	0	
13	46	0	1	156.405	0	
14	50	1	0	153.848	0	
15	47	1	1	153.88	1	
16	47	1	0	167.93	0	
17	45	1	0	104.4	0	
18	45	0	0	146.845	0	
19	45	0	0	162.568	0	
20	47	0	1	197.076	0	
21	49	0	1	199.228	0	
22	49	1	0	203.184	0	
23	47	1	1	201.464	1	
24	48	0	1	176.17	0	
25	45	0	1	180.198	0	

Figure 2 : Parkinson's Disease Patient Data.

MODEL TRAINING:

This research paper studies Logistic Regression, Random Forest classifier, Support Vector classifier and K nearest neighbors’ models in 3 approaches.

- Complete dataset of 171 records and 6 attributes.
- Dataset with 171 records and 6 attributes after Principal Component Analysis (PCA).

Result and Discussion

After training the models, we achieved the following results.

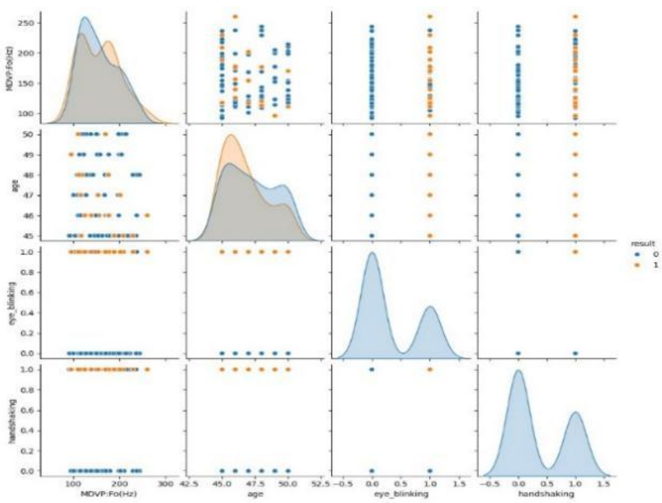

Figure 3 : The image shows the Pair Plot graph

The image represents a pair plot visualization used for exploratory data analysis (EDA) in Parkinson's disease detection. It shows the relationships between different numerical features in the dataset, such as MDVP: Fo(Hz) (vocal frequency), age, eye blinking, and handshaking. The diagonal plots display the distribution of individual features, while the off-diagonal scatter plots show pairwise relationships between features. The color coding indicates two classes: blue (healthy individuals) and orange (individuals with Parkinson's disease). The visualization helps identify patterns, such as how handshaking and eye blinking may be key indicators of the disease, while other features like MDVP: Fo(Hz) and age show some overlap between the two classes. This analysis aids in selecting relevant features for machine learning models to improve disease classification accuracy.

4.1. SUPPORT VECTOR MACHINE (SVM):

The confusion matrix for the SVM model shows perfect classification with 17 true positives and 17 true negatives, indicating 100% accuracy with no misclassifications. This suggests that the model is highly effective in distinguishing between the two classes.

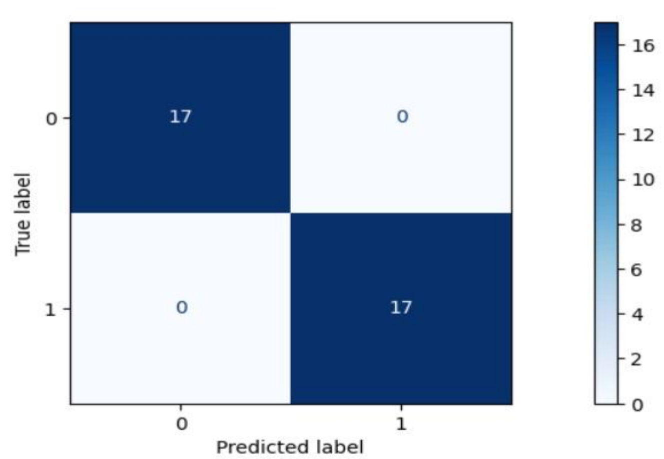

Figure 4: Confusion matrix for SVM

4.2. RANDOM FOREST:

The confusion matrix above illustrates the performance of the Random Forest model in detecting Parkinson's Disease. The model correctly classified all 17 healthy individuals (class 0) and all 17 Parkinson's patients (class 1), resulting in zero misclassifications.

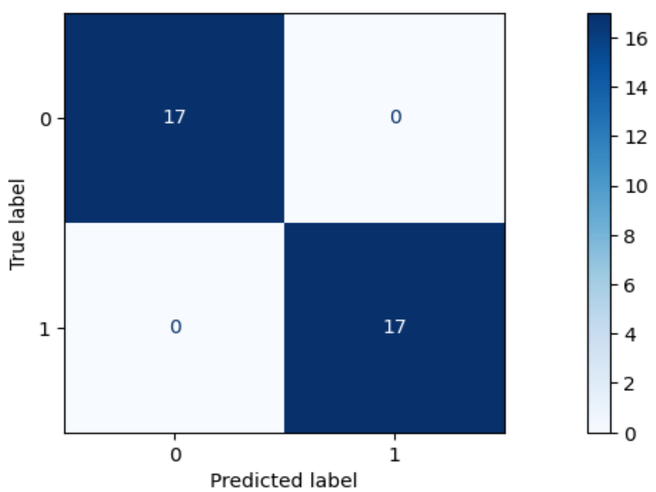

Figure 5: Confusion matrix for Random Forest Model

This indicates 100% accuracy, with no false positives (wrongly predicting Parkinson's for a healthy person) and no false negatives (failing to detect Parkinson's in an actual patient)

4.3. ACCURACY:

The model's precision, recall, f1-score, and accuracy were calculated:

	Metric	DT	RF	LR	SVM	NB	KNN
0	Accuracy	0.864407	0.949153	0.830508	0.983051	0.762712	0.949153
1	F1-Score	0.826087	0.941176	0.782609	0.980392	0.650000	0.938776
2	Recall	0.730769	0.923077	0.692308	0.961538	0.500000	0.884615
3	Precision	0.950000	0.960000	0.900000	1.000000	0.928571	1.000000
4	R2-Score	0.449883	0.793706	0.312354	0.931235	0.037296	0.793706

Figure 6 : Accuracy Table

The table compares machine learning models (DT, RF, LR, SVM, NB, KNN) based on accuracy, F1-score, recall, precision, and R^2-score. SVM performs best with 98.3% accuracy, 100% precision, and the highest R^2-score (0.931), making it the most reliable model. RF and KNN also show strong performance (94.9% accuracy), while NB performs the worst (76.2% accuracy, 50% recall). Logistic Regression has the lowest recall (69.2%), making it less effective for this dataset. Overall, SVM is the best choice, providing the highest predictive reliability.

Conclusion

In this research, we developed and evaluated machine learning models for the accurate detection of Parkinson's disease. The results demonstrate that Support Vector Machine (SVM) outperforms other models with the highest accuracy, precision, and recall, making it the most suitable choice for early detection. Random Forest (RF) and K-Nearest Neighbors (KNN) also showed competitive performance, whereas Naïve Bayes (NB) exhibited the weakest results due to lower recall. The study highlights the importance of feature selection and data preprocessing in improving classification accuracy. Future work can explore deep learning techniques, larger datasets, and real-time implementation to enhance diagnostic efficiency and clinical applicability.

References

[1]. D. Trivedi H. Jaeger and M. Stadtschnitzer. (2019). "Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. Available: https://doi.org/10.5281/zenodo.2867216

[2]. Prabhavathi, K., Patil, S. (2022). A Review "Tremors and Bradykinesia. In: Arjunan, S.P., Kumar, D.K. (eds) Techniques for Assessment of Parkinsonism for Diagnosis and Rehabilitation", Series in Bio Engineering. Available: https://doi.org/10.1007/978-981-16-3056-9_9

[3]. Braak, H., Braak, E. (2000) "Pathoanatomy of Parkinson's disease" J Neurol 247, II3–II10 Available: https://doi.org/10.1007/PL00007758

[4]. Ali, L., Chakraborty, C., He, Z. et al. (2022), " "A novel sample and feature dependent ensemble approach for Parkinson's disease detection". Available: https://doi.org/10.1007/s00521-022-07046-2

[5]. Aghav-Palwe S, Mishra D (2020)" "A "Statistical tree-based feature vector for content-based image retrieval." Int J Computer Sci Eng 2020". Available: https://doi.org/10.1504/IJCSE.2020.106868

[6]. D. V. Rao, Y. Sucharitha, D. Venkatesh, K. Mahamthy and S. M. Yasin (2022), "Diagnosis of Parkinson's Disease using Principal Component Analysis and Machine Learning algorithms with Vocal Features,".2021 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 200-206, doi: 10.1109/ICSCDS53736.2022.9760962