PARKINSON'S PROGNOSTICATION: MACHINE LEARNING PREDICTIVE MODELING FOR EARLY DIAGNOSIS

Venkat Reddy Adama

Department of Electronics and Communications Engineering, Vaageswari College of Engineering, Karimnagar, 505527, Telangana.

Mail ID: venkat7641@gmail.com

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder characterized by motor and non-motor symptoms, making its early diagnosis challenging yet critical for effective treatment and management. The application of machine learning-based predictive modeling for Parkinson's disease diagnosis has significant implications for healthcare and clinical practice. Accurate and early diagnosis of PD enables timely intervention and treatment planning, facilitating better patient outcomes and quality of life. Additionally, machine learning models can assist healthcare professionals in screening individuals at risk for PD, potentially leading to earlier detection and intervention. Moreover, these models can support research efforts aimed at understanding the underlying mechanisms of PD progression and identifying novel biomarkers for disease diagnosis and monitoring. Traditional methods for PD diagnosis often rely on clinical assessment tools and subjective evaluations, which may lack sensitivity and specificity, particularly in the early stages of the disease. These methods typically involve manual scoring of motor symptoms and may overlook subtle changes in speech patterns and vocal characteristics associated with PD. Moreover, clinical assessments may be time-consuming and require specialized expertise, limiting their scalability and accessibility in primary care settings. Additionally, traditional diagnostic approaches may fail to utilize the wealth of information available in voice recordings, such as detailed acoustic features and nonlinear dynamics, leading to suboptimal diagnostic accuracy. The proposed system aims to overcome the limitations of traditional diagnostic methods by leveraging machine learning techniques for PD diagnosis using voice data. This work employs supervised learning algorithms to train predictive models on the dataset of voice features. By extracting informative features and learning complex patterns from the data, the proposed models can effectively distinguish between individuals with PD and healthy conditions.

Key words: Parkinson’s Disease, Predictive Modeling, Health Informatics, Neurodegenerative Disorders, Diagnostic Prediction

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder that primarily affects the motor system, causing a range of symptoms including tremors, stiffness, and impaired balance and coordination. It is characterized by the progressive loss of dopaminergic neurons in the substantia nigra region of the brain. As the disease advances, individuals may also experience non-motor symptoms such as cognitive impairment, mood disturbances, and autonomic dysfunction. The exact cause of PD is not fully understood, but it is believed to involve a combination of genetic, environmental, and lifestyle factors.

The side effects of PD can significantly impact an individual's quality of life. Motor symptoms such as tremors and bradykinesia (slowed movement) can interfere with daily activities, making tasks like walking, eating, and writing challenging. As the disease progresses, these symptoms may worsen and

become more debilitating, leading to increased dependency on caregivers and decreased independence for the individual with PD. Non-motor symptoms such as depression, anxiety, and sleep disturbances can also have a significant impact on overall well-being and contribute to the burden of the disease. In hospitals and clinical settings, the lack of automated diagnostic tools for PD contributes to delays in diagnosis and treatment initiation. Manual inspection by specialists is time-consuming and resource-intensive, leading to bottlenecks in the healthcare system, particularly in regions with limited access to neurologists and movement disorder specialists. As the number of patients with PD continues to rise globally, the burden on healthcare resources is expected to increase, underscoring the urgency of developing automated diagnostic solutions to streamline the diagnostic process and improve patient outcomes.

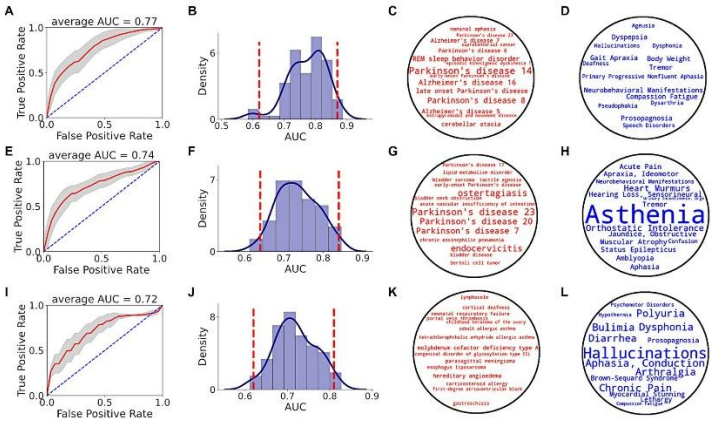

Figure 1. Early detection of Parkinson disease

The primary application of the proposed machine learning predictive modeling system is in the early diagnosis of Parkinson's disease (PD). By analyzing voice data and extracting informative features, the system can detect subtle changes in speech patterns that may serve as early biomarkers of PD. Early diagnosis is crucial for initiating timely interventions that can slow disease progression and improve patient outcomes. By identifying individuals at risk for PD before the onset of motor symptoms, the system enables healthcare providers to intervene early with neuroprotective therapies and lifestyle interventions.

2. LITERATURE SURVEY

This paper [1] surveys various machine learning algorithms for predicting Parkinson's disease. Among them are Decision tree models, random forests, machines with support vectors, artificial neural networks and other algorithms.[2] The algorithms' accuracy spans from 70% to 99%, with certain algorithms performing better than others. The study in paper [3] found that SVM showed good accuracy (88.9%) compared to other algorithms, and Random Forest had the highest accuracy of 90.26% while Naïve Bayes had the lowest level accuracy of 69.23%.[4] Hierarchical clustering and SOM were also used, predicting higher numbers of clusters inhealthy datasets. In the paper [5], with 34 support vectors, the Nu-SVM model depending on the Gaussian method was shown to have the maximum sensitivity and overall accuracy. The research presents an ensemble learning approach for utilising machine learning to predict early warning signals of Parkinson's disease. The proposed model surpasses existing approaches such as SVM, KNN, RF, DT, MLP, SC, and LR, with an accuracy of 94.87%.

In this study [6], functional MRI (fMRI) data were used to discover brain activity patterns linked to optimal and nonoptimal deep brain stimulation (DBS) settings inParkinson's disease (PD) patients. achieving 88% accuracy inforecasting optimal vs. non-optimal circumstances. In paper [7] The analysis

reveals that patients can be classified into three subtypes of PD: slow progressors, moderate progressors, and fast progressors. The approach can aid in the interpretability of clinical features and disease progression. The algorithms used were unsupervised learning and mathematical projection. This paper [8] presents a study of Parkinson's Disease (PD) diagnosis using voice and tremor data. For tremor data, kNN achieved the highest accuracy of 98.5% for 2-level classification and 90% for 5-level classification. By combining both voice and tremor data, an accuracy of 99.8% was achieved using ensemble averaging of kNN, SVM, and naive Bayes for PD detection. The study uses kNN, SVM and Naive Bayes algorithmsfor classification. The highest accuracy for male voice samples was found to be 90.3% in kNN, and for female voicesamples, it was 95.8% in kNN. In tremor data, the maximum accuracy for PD vs non-PD classification was 98.5% in kNN. This paper [9] presents a multimodal machine learning model for predicting the risk of Parkinson's disease. The model was developed using an open-source auto-ML package called GenoML and was validated in an external cohort. The model outperformed previous efforts with an accuracy of 89.72% and was based on a combination of clinicodemographic, genetic, and transcriptomic data. The study [10] utilizes a PCA-RF model for detecting Parkinson's disease. It was found that the model's performance without PCA was better than with PCA. Specifically, the model achieved 89.9% and 76.7% accuracy, 70.2% and 55.6% sensitivity, and 96.5% and 80.6% specificity without and with PCA, respectively.

In study [11], the outcomes shown the advantages of the suggested ANFIS+PSOGWO algorithm, which outperformed its rivals by 7.3% and predicted Parkinson's disease with an accuracy of 87.5%. The suggested approach outperformed some recent research on Parkinson's disease prediction that employed PSO, GWO, GA, ACO, and DE, among other optimization techniques. In paper [12] The accuracy of the research varies according on the quality and amount of the datasets used, as well as the methods used. Nonetheless, sensitivities in the 90%- 95% range were reached using today's approaches.

Parkinson's disease detection techniques in paper [13] employs a number of equipment to evaluate the degree of illness. The vocal difficulty is one of the most prevalent symptoms, and most patients have vocal defections in the initial phases of the disease. As an outcome, medical systems driven by voice concerns have assumed the lead in contemporary PD detection research. With significance set in study [14] at p 0.05, the study used SPSS or R to do statistical and data analysis. Machine learning algorithms to forecast 2-year longitudinal medical findings, models such as elastic-net and random forest were developed based on clinical factors, inflammatory cytokine measures, and demographic information (age and sex).

3. PROPOSED SYSTEM

This research has demonstrated the effectiveness of machine learning models, particularly Support Vector Machines (SVMs) and Decision Tree Classifiers (DTCs), in diagnosing Parkinson's disease (PD) using voice data. Figure 4.1 shows the proposed system model. Through a comprehensive research procedure encompassing dataset acquisition, preprocessing, model building, and performance evaluation, we have shown that both SVMs and DTCs can accurately classify individuals as either having PD or being healthy based on their voice features.

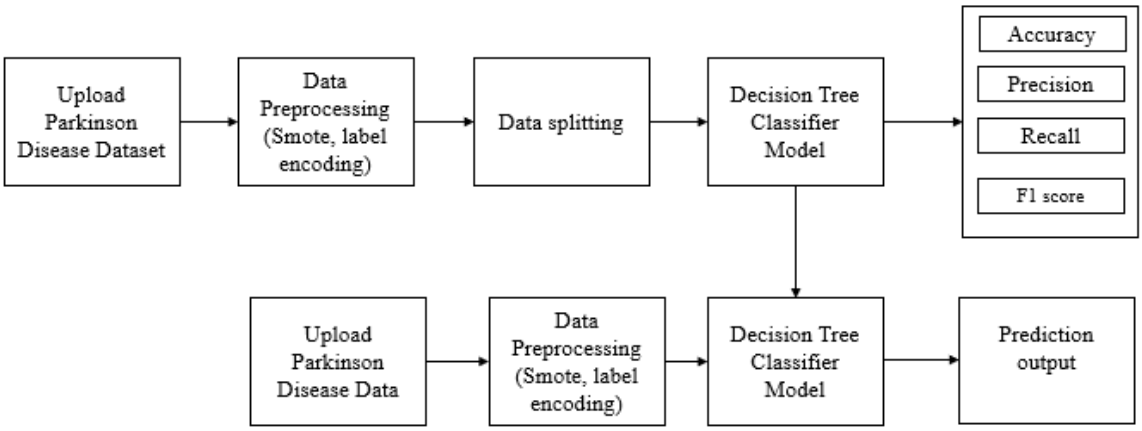

Figure 2. Proposed Block Diagram

The proposed methodology begins with collecting a comprehensive dataset of voice samples from individuals with Parkinson's disease (PD) and healthy controls to ensure robust model development. The data undergoes preprocessing, including null value analysis and label encoding, to improve quality and prepare it for machine learning. To address class imbalance, SMOTE is applied to generate synthetic minority samples, resulting in a more representative dataset. The data is then split into training and testing subsets using an 80-20 ratio to enable reliable performance evaluation. An existing SVM model is first built and optimized for PD classification, followed by a proposed Decision Tree Classifier (DTC) model trained using the same dataset to compare its effectiveness. Both models are evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to determine their suitability for PD diagnosis. Finally, the trained DTC model is used to predict PD in test samples, validating its diagnostic capability and assessing its potential usefulness in real-world clinical applications. DTC is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "DTC is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the DTC takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4. RESULTS AND DISCUSSIONS

Figure 3 shows the graph before applying Synthetic Minority Over-sampling Technique (SMOTE), the dataset is imbalanced, with significantly fewer records for the healthy class compared to the Parkinson class. Specifically, there are 147 records corresponding to individuals with Parkinson's disease (PD) and only 48 records for healthy individuals. This class imbalance can pose challenges for machine learning models, leading to biased predictions and poor performance on the minority class.

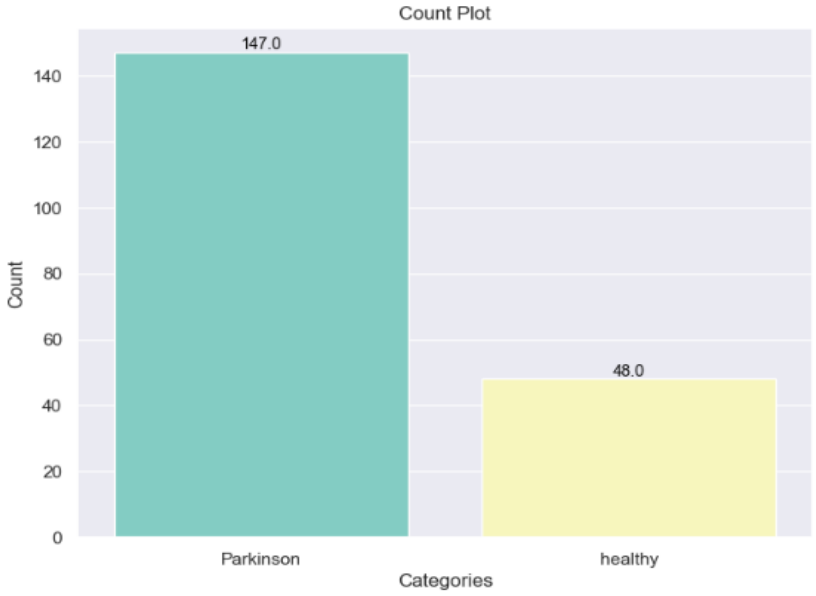

Figure 3. Count plot of PD target before applying SMOTE.

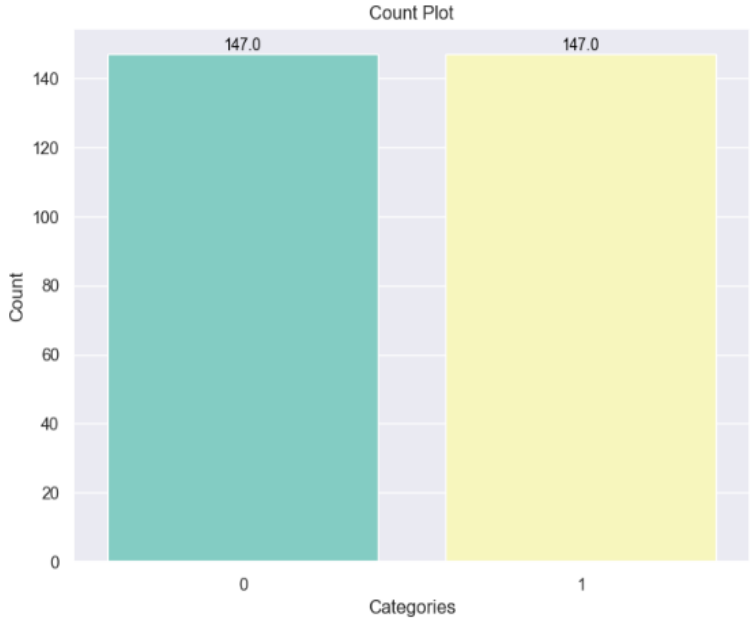

Figure 4. Count plot of PD target after applying SMOTE.

```
Model loaded successfully.
Support Vector Machine Classifier Accuracy    : 77.96610169491525
Support Vector Machine Classifier Precision   : 78.34101382488478
Support Vector Machine Classifier Recall      : 78.67132867132867
Support Vector Machine Classifier FSCORE      : 77.94075352315215

 Support Vector Machine Classifier classification report
              precision    recall  f1-score   support

   PARKINSON       0.73      0.86      0.79        28
     HEALTHY       0.85      0.71      0.77        31

    accuracy                           0.78        59
   macro avg       0.79      0.78      0.78        59
weighted avg       0.79      0.78      0.78        59
```

Figure 5. Existing SVM performance.

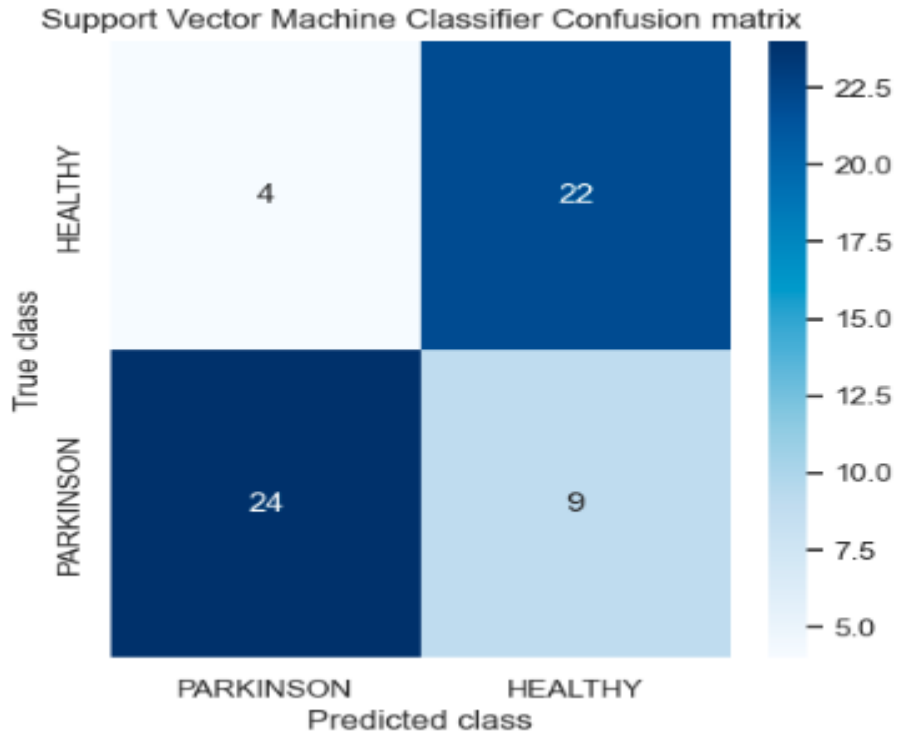

Figure 6. Existing SVM confusion matrix.

```
Model loaded successfully.
DecisionTreesClassifier Accuracy    : 98.30508474576271
DecisionTreesClassifier Precision   : 98.14814814814814
DecisionTreesClassifier Recall      : 98.48484848484848
DecisionTreesClassifier FSCORE      : 98.28737300435414

 DecisionTreesClassifier classification report
              precision    recall  f1-score   support

   PARKINSON       0.97      1.00      0.98        32
     HEALTHY       1.00      0.96      0.98        27

    accuracy                           0.98        59
   macro avg       0.98      0.98      0.98        59
weighted avg       0.98      0.98      0.98        59
```

Figure 7. Classification report obtained for DTC

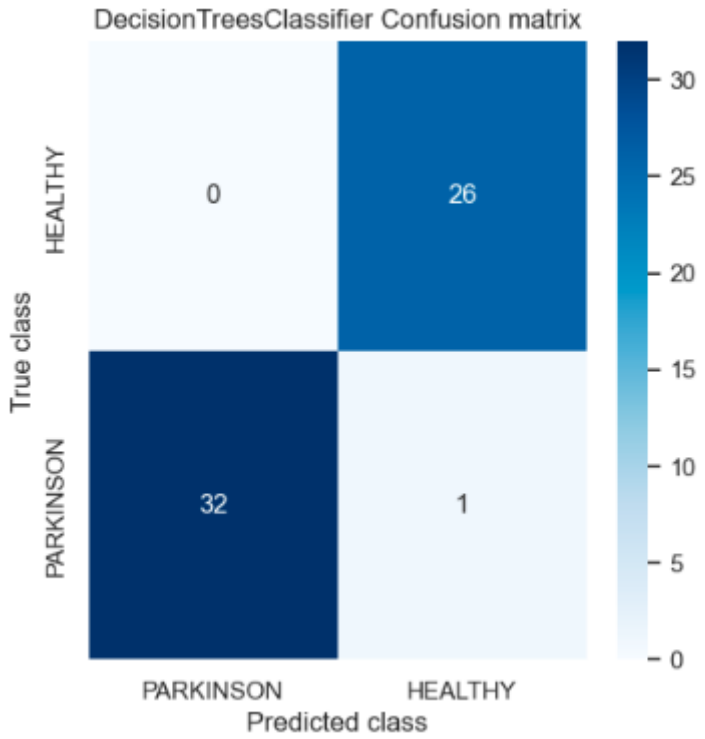

Figure 8. Confusion matrix obtained for DTC

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Support Vector Machine Classifier	78.341014	78.671329	77.940754	77.966102
1	DecisionTreesClassifier	98.148148	98.484848	98.287373	98.305085

Figure 9. Comparison Table of SVC and DTC

Figure 5 shows the performance of a Support Vector Machine (SVM) classifier on a binary classification task. The task appears to be classifying between Parkinson's disease and a healthy state.

- Accuracy: 77.96%
- Precision:

  - Parkinson's: 73%
  - Healthy: 85%
- Recall:
  - Parkinson's: 86%
  - Healthy: 71%
- F1 Score: 77.94%
- In machine learning, a classifier is an algorithm that predicts the category (or label) for a given input data point. Support Vector Machines (SVMs) are a type of supervised learning model with various applications including classification, regression and outlier detection.
- Here are additional details about the performance metrics mentioned in the figure:
- Accuracy: How often the classifier correctly classifies a data point.
- Precision: How often a data point with a positive label is truly positive.
- Recall: How often a truly positive data point is classified as positive by the classifier.
- F1 Score: A harmonic mean between precision and recall.

Figure 6 is a confusion matrix for a Support Vector Machine (SVM) classifier. A confusion matrix is a table that is used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions made by the model. In the specific confusion matrix you sent, the model is classifying Parkinson's disease. The rows of the matrix represent the actual class of the data (Parkinson's disease or healthy), and the columns represent the predicted class of the data. The diagonal cells of the matrix show the number of correct predictions. For example, the top left cell shows that 22 out of 25 people with Parkinson's disease were correctly classified as having Parkinson's disease. The off-diagonal cells show the number of incorrect predictions. For example, the bottom left cell shows that 3 out of 24 healthy people were incorrectly classified as having Parkinson's disease.

Figure 7 shows the results of a decision tree classification report. The accuracy of the decision tree classification report is 0.983, which means that the model was able to correctly classify 98.3% of the data points.

Here is a breakdown of the metrics shown in the report:

- **Accuracy:** This is the overall percentage of correct predictions made by the model. In this case, it is 98.3%.
- **Precision:** This metric tells you how often a positive prediction by the model is actually correct. A high precision means that most of the positive predictions are true positives. In this case, the precision is 0.9848 for Parkinson's and 1.00 for Healthy.
- **Recall:** This metric tells you how often the model correctly identifies a positive case. A high recall means that the model is not missing many positive cases. In this case, the recall is 1.00 for Parkinson's and 0.96 for Healthy.
- **F1-Score:** This metric is the harmonic mean of precision and recall. It is a way to balance between precision and recall. In this case, the F1-score is 0.98 for Parkinson's and 0.98 for Healthy.

Figure 8 shows confusion matrix, but it appears to be for a decision tree classifier, not a Support Vector Machine (SVM) classifier, as I previously stated.

Confusion matrices are used to evaluate the performance of a classification model. They show the number of correct and incorrect predictions made by the model.

Figure 9 the table shows the following information:

- Algorithm Name: This column identifies the two algorithms being compared.
- Precision: This metric shows the ratio of true positive results to the total number of positive results predicted by the model. In simpler terms, it shows how many of the instances classified as positive by the model were actually positive.
- Recall: This metric shows the ratio of true positive results to the total number of actual positive cases. In other words, it shows how many of the actual positive cases were identified correctly by the model.
- FScore: This metric is the harmonic mean of precision and recall. It is a way to balance between the two metrics and get a single score that reflects both precision and recall.
- Accuracy: This metric shows the ratio of total correct predictions to the total number of cases evaluated by the model.
- The table shows that the Decision Tree classifier performs better than the SVM classifier according to all the metrics. Here's a breakdown of the results:
- Precision: Decision Tree (98.15%) vs SVM Classifier (78.34%)
- Recall: Decision Tree (98.48%) vs SVM Classifier (78.67%)
- FScore: Decision Tree (98.29%) vs SVM Classifier (77.94%)
- Accuracy: Decision Tree (98.31%) vs SVM Classifier (77.97%)

5. CONCLUSION

In conclusion, the integration of machine learning-based predictive modeling into Parkinson's disease (PD) diagnosis represents a promising avenue for improving early detection and intervention strategies. The complexities inherent in PD, encompassing a wide array of motor and non-motor symptoms, pose challenges to conventional diagnostic approaches. However, by harnessing the power of machine learning algorithms, particularly in analyzing voice data, this study demonstrates the potential to revolutionize PD diagnosis. The findings underscore the significance of accurate and timely diagnosis in PD management. Early identification of the disease enables healthcare professionals to initiate appropriate interventions promptly, thereby optimizing treatment outcomes and enhancing patients' quality of life. Moreover, the predictive models developed in this study hold promise for screening individuals at risk for PD, paving the way for proactive healthcare measures and preventive interventions. The utilization of machine learning techniques offers a pathway for advancing our understanding of PD pathophysiology. By leveraging vast amounts of data, including intricate voice features, these models can unveil subtle patterns and biomarkers indicative of PD progression. Such insights not only contribute to refining diagnostic algorithms but also fuel research endeavors aimed at elucidating the underlying mechanisms of the disease. The proposed machine learning-based approach addresses several limitations associated with traditional diagnostic methods for PD. By leveraging voice recordings, the system capitalizes on the wealth of information embedded in speech patterns and vocal characteristics, which may serve as early indicators of PD onset. This non-invasive and easily accessible modality holds promise for widespread adoption in primary care settings, circumventing the need for specialized expertise and time-consuming assessments. As with any novel methodology, there are areas for further refinement and exploration. Future research could focus on enhancing the robustness and generalizability of predictive models by incorporating multi-modal data sources and refining feature selection techniques. Additionally, longitudinal studies are warranted to assess the predictive validity

of the proposed models over time, taking into account the dynamic nature of PD progression. The integration of machine learning-based predictive modeling holds immense potential for transforming the landscape of PD diagnosis and management. By leveraging innovative approaches to analyze voice data, this study advances our capacity to detect PD at its incipient stages, paving the way for personalized and proactive healthcare interventions. As we continue to refine and validate these models, they stand poised to revolutionize clinical practice, offering hope for improved outcomes and quality of life for individuals living with PD.

REFERENCES

[1]. Surekha Tadse, Muskan Jain, Pankaj Chandankhede. "Parkinson's Detection Using Machine Learning". Published in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) by IEEE.

[2]. S, Sivachitra M, Vijayachitra S. "Parkinson's Disease Prediction Using Machine Learning Approaches". Published in: 2013 Fifth International Conference on Advanced Computing (ICoAC).

[3]. Haewon Byeon. "Development of a Depression in Parkinson's Disease Prediction Model Using Machine Learning". Published in World J Psychiatry. 2020 Oct 19; 10(10): 234–244.

[4]. Pawan Kumar Mall, Rajesh Kumar Yadav, Arun Kumar Rai, Vipul Narayan, Swapnita Srivastava. "Early Warning Signs of Parkinson's Disease Prediction Using Machine Learning Technique". Published in Journal of Pharmaceutical Negative Results, Special Issue, Vol. 14, p2607-2615. 9p.

[5]. Alexandre Boutet, Radhika Madhavan, Gavin J. B. Elias, Suresh E. Joel, Robert Gramer, Manish Ranjan, Vijayashankar Paramanandam, David Xu, Jurgen Germann, Aaron Loh, Suneil K. Kalia, Mojgan Hodaie, Bryan Li, Sreeram Prasad, Ailish Coblentz, Renato P. Munhoz, Jeffrey Ashe, Walter Kucharczyk, Alfonso Fasano & Andres M. Lozano. "Predicting Optimal Deep Brain Stimulation Parameters for Parkinson's Disease Using Functional MRI and Machine Learning". Published: 24 May 2021.

[6]. Anant Dadu, Vipul Satone, Rachneet Kaur, Sayed Hadi Hashemi, Hampton Leonard, Hirotaka Iwaki, Mary B. Makarious, Kimberley J. Billingsley, Sara Bandres-Ciga, Lana J. Sargent, Alastair J. Noyce, Ali Daneshmand, Cornelis Blauwendraat, Ken Marek, Sonja W. Scholz, Andrew B. Singleton, Mike A. Nalls, Roy H. Campbell & Faraz Faghri. "Identification and Prediction of Parkinson's Disease Subtypes and Progression Using Machine Learning in Two Cohorts". Published: 16 December 2022.

[7]. Md. Sakibur Rahman Sajal, Md. Tanvir Ehsan, Ravi Vaidyanathan, Shouyan Wang, Tipu Aziz & Khondaker Abdullah Al Mamun. "Telemonitoring Parkinson's Disease Using Machine Learning by Combining Tremor and Voice Analysis". Published by 'Brain Informatics, Springer Open'. Published: 22 October 2020.

[8]. Mehrbakhsh Nilashi, Rabab Ali Abumalloh, Behrouz Minaei-Bidgoli, Sarminah Samad, Muhammed Yousoof Ismail, Ashwaq Alhargan, Waleed Abdu Zogaan. "Predicting Parkinson's Disease Progression: Evaluation of Ensemble Methods in Machine Learning". Volume 2022 | Article ID 2793361 from 'Hindawi.com'. Published: 03 Feb 2022.

[9]. Mary B. Makarious, Hampton L. Leonard, Dan Vitale, Hirotaka Iwaki, Lana Sargent, Anant Dadu, Ivo Violich, Elizabeth Hutchins, David Saffo, Sara Bandres-Ciga, Jonggeol Jeff Kim, Yeajin Song, Melina Maleknia, Matt Bookman, Willy Nojopranoto, Roy H. Campbell, Sayed Hadi Hashemi, Juan A. Botia, John F. Carter, David W. Craig, Kendall Van Keuren-Jensen, Huw R. Morris, John Hardy, Cornelis Blauwendraat, …Mike A. Nalls. "Multi-Modality

Machine Learning Predicting Parkinson's Disease". Published in Nature.com. Published: 01 April 2022.

[10]. M.S. Roobini, Yaragundla Rajesh Kumar Reddy, Udayagiri Sushmanth Girish Royal. "Parkinson's Disease Detection Using Machine Learning". Published in: 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT).

[11]. Ibrahim M. El-Hasnony, Sherif I. Barakat, and Reham R. Mostafa. "ptimized ANFIS Model Using Hybrid Metaheuristic Algorithms for Parkinson's Disease Prediction in IoT Environment". Published in IEEE Access. Received June 16, 2020, accepted June 25, 2020, date of publication June 29, 2020, date of current version July 8, 2020.

[12]. Claas Ahlrichs and Michael Lawo. "Parkinson's Disease Motor Symptoms in Machine Learning: A Review". Published in Health Informatics - An International Journal (HIIJ), Vol. 2, No. 4, November 2013.

[13]. Hakan Gunduz. "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets". Published in IEEE Access (Volume: 7).

[14]. Diba Ahmadi Rastegar, Nicholas Ho, Glenda M. Halliday & Nicolas Dzamko. "Parkinson's Progression Prediction Using Machine Learning and Serum Cytokines". Published in Nature.com. Published: 25 July 2019.