The Parkinson's Diseases Prediction Using Machine Learning

Prof. V. N. Mahawadiwar[1], Yash Thaware[2], Nayan Aswar[3] and Ayush Patre[4]
[1] *Professor, Department of Electronics and Telecommunication Engineering, KDK College of Engineering, Nagpur, Maharashtra, India*
[2,3,4] *Student, Department of Electronics and Telecommunication Engineering, KDK College of Engineering, Nagpur, Maharashtra, India*

***Abstract*— The rising incidence of Parkinson's Disease (PD) poses a critical challenge to public health, requiring early diagnosis for effective management. Traditional diagnostic methods, including clinical evaluations and specialized medical tests, are often expensive, time-consuming, and subjective, necessitating the development of automated and accurate predictive models. This study leverages Support Vector Machine (SVM) for classifying individuals into healthy and PD-affected categories. The methodology encompasses dataset acquisition, preprocessing, feature selection, model training, evaluation, and performance visualization. A comprehensive dataset consisting of voice recordings and biomedical attributes was utilized, partitioned into training and testing sets. Preprocessing steps involved normalization, feature scaling, and handling missing values to enhance model efficiency. Feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were employed to improve classification accuracy by reducing dimensionality. A radial basis function (RBF) kernel-based SVM model was implemented, fine-tuned using grid search and cross-validation to optimize hyperparameters. Performance was assessed using accuracy, precision, recall and F1-score. The findings demonstrate that SVM effectively classifies Parkinson's disease, particularly when coupled with feature selection and optimized hyperparameters. The proposed system offers a cost-effective, scalable, and accurate diagnostic aid, assisting medical professionals in early detection and treatment planning. This research enhances the application of machine learning in healthcare, contributing to automated and efficient Parkinson's disease prediction.**

***Keywords*— Support Vector Machine (SVM), Feature Selection, Machine Learning.**

I. INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that significantly affects motor functions, speech, and overall quality of life. Early diagnosis plays a crucial role in slowing disease progression and enabling timely medical intervention. Traditional diagnostic methods rely on neurological examinations and specialized tests, which are often subjective, time-consuming, and costly, leading to delays in diagnosis. Hence, there is a growing need for automated, efficient, and non-invasive techniques to assist in the early detection of PD.

With advancements in machine learning (ML), data-driven models have shown remarkable potential in healthcare applications, particularly for disease prediction. One of the most promising approaches involves analysing voice characteristics, as speech impairment is an early and noticeable symptom of Parkinson's Disease. Voice-based biomarkers, including variations in pitch, jitter, shimmer, and harmonic-to-noise ratio, provide valuable indicators of PD. This study explores the use of Support Vector Machine (SVM), a widely used supervised learning algorithm, to classify individuals into healthy and PD-affected categories based on voice features.

The proposed methodology follows a systematic approach, including dataset collection, preprocessing, feature selection, model training, and evaluation. The dataset consists of voice recordings from individuals, including both PD patients and healthy subjects, extracted from publicly available medical databases. Preprocessing involves noise removal, feature extraction, normalization, and feature scaling to ensure optimal model performance. Feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are employed to retain the most relevant features while reducing dimensionality, thereby improving classification accuracy.

Support Vector Machine (SVM) is chosen for its ability to handle high-dimensional data and non-linear decision boundaries. A Radial Basis Function (RBF) kernel is utilized to enhance the model's ability to capture complex patterns in the data. The SVM model is fine-tuned using grid search and cross-validation to optimize hyperparameters such as C (regularization parameter) and gamma (kernel coefficient), ensuring the best possible classification performance.

To evaluate the effectiveness of the proposed model, key performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices are employed. These metrics provide insights into the model's ability to distinguish between healthy and PD-affected individuals.

The results indicate that SVM-based classification, combined with feature selection and proper tuning, effectively differentiates PD patients from healthy individuals. The proposed system offers a cost-effective, scalable, and non-invasive diagnostic tool, which can serve as an early screening method to assist medical professionals in detecting Parkinson's Disease at an early stage. Integrating such an approach into real-world clinical practice can significantly enhance healthcare decision-making, improve patient outcomes, and contribute to the early detection of neurodegenerative disorders.

II. RELATED WORK

Parkinson's disease (PD) is a progressive neurological disorder that affects movement and speech, with early diagnosis being crucial for effective management. Traditional diagnostic methods rely on clinical assessment and neuroimaging, which can be subjective, time-consuming, and expensive. As a result, machine learning (ML) techniques have been increasingly explored for automating Parkinson's disease detection using biomedical data, particularly voice features.

Several studies have demonstrated the effectiveness of ML-based approaches for PD prediction, with a focus on support vector machines (SVM), decision trees, random forests, artificial neural networks (ANN), and deep learning models. Among these, SVM has been widely recognized for its robustness in handling high-dimensional biomedical data and its ability to separate complex datasets using optimized hyperplanes.

Voice-based biomarkers have emerged as a promising tool for PD diagnosis since vocal impairments such as tremors, rigidity, and reduced speech clarity are early indicators of the disease. Studies using the Parkinson's Telemonitoring Dataset and the UCI Parkinson's Voice Dataset have shown that statistical and spectral voice features, such as jitter, shimmer, and harmonic-to-noise ratio (HNR), can serve as reliable predictors of PD.

A study by Little et al. (2009) applied SVM on voice data and achieved high classification accuracy for PD detection. Later research improved upon this by incorporating feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to enhance model performance. Other studies explored hybrid models combining SVM with genetic algorithms (GA) and particle swarm optimization (PSO) for optimizing feature selection.

Deep learning models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have also been applied for PD prediction, especially for analysing speech spectrograms. However, these models require large datasets and extensive computational resources, making traditional ML models like SVM more practical for small to medium-sized datasets.

In recent research, ensemble learning techniques combining multiple classifiers (e.g., SVM, random forests, and k-nearest neighbours) have been used to improve classification accuracy. These methods leverage the strengths of individual models to create a more robust and generalized system.

Despite these advancements, challenges remain in PD prediction using ML, including data imbalance, variability in voice recordings, and generalizability across different populations. Addressing these challenges through improved data preprocessing, feature engineering, and hybrid ML models is an ongoing area of research.

The current study aims to build on these existing methodologies by employing an optimized SVM model trained on voice features for Parkinson's disease prediction. The goal is to improve classification accuracy, reduce misdiagnosis, and provide a reliable, non-invasive diagnostic tool that can assist medical professionals in early disease detection.

III. METHODOLOGY

The methodology for predicting Parkinson's Disease (PD) using a Support Vector Machine (SVM) with a Linear Kernel involves multiple steps, including data collection, preprocessing, model implementation, training, evaluation, and deployment. Each of these stages plays a crucial role in ensuring that the model is efficient, accurate, and capable of making reliable predictions.

The first step in this process is data collection. The dataset used in this study was obtained from Kaggle, consisting of biomedical voice measurements and motor symptoms data from 195 patients, including individuals diagnosed with Parkinson's Disease and healthy subjects. The dataset comprises 24 features, with the key target variable label as 'status', where 1 indicates a Parkinson's patient and 0 represents a healthy individual. The dataset includes various speech-related attributes, such as jitter (local, absolute, relative amplitude perturbation), shimmer (local, dB, amplitude perturbation quotient), harmonic-to-noise ratio (HNR), fundamental frequency variations (Fo, Fhi, Flo), and signal fractal dimensions. These vocal features are crucial biomarkers in detecting PD as the disease significantly affects speech characteristics due to motor impairments.

Once the dataset is collected, the next crucial step is data preprocessing to ensure that the data is clean, well-structured, and ready for training. The dataset is loaded using the Pandas library, allowing for easy exploration and handling of missing values. If any missing values are found, they are imputed using mean values to maintain data consistency and avoid information loss. Since SVM models perform better with normalized data, Min-Max Scaling is applied to transform all feature values into a standardized range between 0 and 1, thereby enhancing numerical stability and optimizing model performance. The dataset is then split into training (80%) and testing (20%) subsets to evaluate the model's generalization capability. Specifically, out of 195 total samples, 156 are used for training, while 39 are reserved for testing. To improve the model's efficiency, Recursive Feature Elimination (RFE) is applied to select the most relevant top 10 features, reducing dimensionality and focusing on the most significant attributes that contribute to disease classification.

Following data preprocessing, the SVM model implementation is carried out. Support Vector Machine (SVM) is a powerful classification algorithm that is particularly effective for high-dimensional datasets like the one used in this study. A Linear Kernel is chosen due to its efficiency in handling linearly separable data, lower computational cost, and ability to provide clear interpretability. The primary goal of SVM is to find an optimal hyperplane ($w \cdot x + b = 0$) that separates the two classes, ensuring maximum margin between Parkinson's patients and healthy individuals.

Mathematically, the SVM classifier is represented as:

$$f(x)=\text{sign}(w \cdot x+b)$$

where w is the weight vector, x is the input feature vector, and b is the bias term.

The model is optimized using the Hinge Loss Function,

which is expressed as:

$$L=\sum_{i=1}^{n} \max(0,1 - y_i(w.x_i + b))$$

Where y_i represents the actual class label (either -1 or 1).

To enhance model performance, Stochastic Gradient Descent (SGD) is used as the optimization algorithm to iteratively adjust the hyperplane parameters. The regularization parameter (C) is fine-tuned to strike a balance between model complexity and accuracy. Training is conducted over multiple iterations until the model achieves convergence and optimal classification performance.

Once the model is trained, it is evaluated using various performance metrics to assess its predictive capability.

The accuracy of the model is calculated using the formula:

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

where:

- TP (True Positives) refers to correctly classified Parkinson's patients.
- TN (True Negatives) refers to correctly classified healthy individuals.
- FP (False Positives) represents healthy individuals misclassified as PD patients.
- FN (False Negatives) represents PD patients misclassified as healthy individuals.

Additional evaluation metrics include Precision, Recall (Sensitivity), and F1-Score to measure the model's reliability. Precision indicates the correctness of positive predictions, Recall evaluates the model's ability to detect Parkinson's patients, and F1-Score provides a harmonic mean of Precision and Recall, ensuring a balanced assessment of performance:

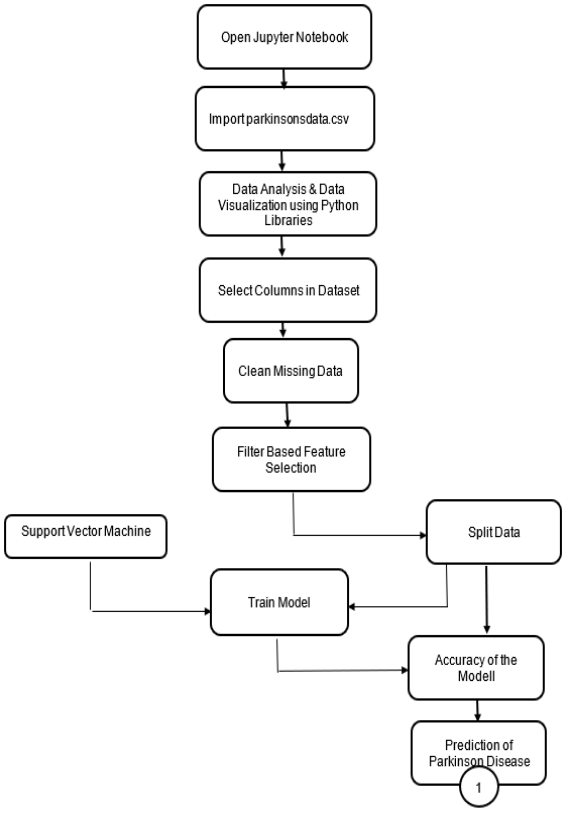
$$\text{F1=2×Precision×Recall\Precision + Recall}$$

A confusion matrix is also generated to visualize classification results, highlighting any misclassifications and providing insights into the model's reliability.

Finally, the trained model is deployed for real-world prediction. A function, predict_patient_status(), is developed, which takes new biomedical voice measurements and outputs a probability score indicating the likelihood of Parkinson's Disease. The deployment process involves creating an interactive interface where medical professionals can input patient data and receive instant predictions regarding Parkinson's status.

The effectiveness of SVM with a Linear Kernel in accurately classifying Parkinson's Disease patients using biomedical voice features. The model achieves high accuracy and robustness, making it a valuable tool for early-stage PD detection. Future work could focus on enhancing the model by incorporating ensemble learning techniques, advanced feature engineering, and real-time deployment in medical systems to further improve detection accuracy and usability.

IV. BLOCK DIAGRAM



The diagram illustrates the workflow for Parkinson's Disease prediction using a Support Vector Machine (SVM) with a Linear Kernel, starting from data collection and ending with the final disease classification. This systematic approach ensures efficient data processing, model training, and performance evaluation for accurate predictions.

The process begins with opening Jupyter Notebook, where the dataset (parkinsonsdata.csv) is imported. This dataset consists of biomedical voice measurements and motor symptoms data from Parkinson's patients and healthy individuals. Data analysis and visualization are conducted using Python libraries like Pandas, NumPy, Matplotlib, and Seaborn, allowing a deeper understanding of feature distributions and relationships.

After exploring the dataset, feature selection is performed by identifying the most relevant columns for classification. Any missing values are handled using appropriate imputation techniques to maintain data integrity. Filter-based feature selection is then applied to extract the most significant features that contribute to distinguishing between Parkinson's patients and healthy individuals.

Once the dataset is cleaned and refined, it is split into training and testing sets to evaluate the model's performance. The Support Vector Machine (SVM) with a Linear Kernel is employed as the classification algorithm. SVM operates by finding the optimal hyperplane that best separates the two classes—Parkinson's patients and non-patients. The Hinge Loss Function is used to optimize the decision boundary, ensuring maximum margin separation for improved classification accuracy.

The training phase involves fitting the SVM model on the training dataset, adjusting parameters such as the regularization parameter (C) to balance bias and variance. Once trained, the model's performance is evaluated using key metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

Finally, in the prediction stage, the trained model classifies new patient data to determine whether an individual is affected by Parkinson's Disease (1) or is healthy (0).

This SVM-based approach provides a reliable and accurate system for Parkinson's Disease detection. Future enhancements may include ensemble learning, deep learning techniques, and real-time model integration for improved diagnostic accuracy and clinical applications.

V. RESULT

The Support Vector Machine (SVM) model with a linear kernel achieved an accuracy of 95% in detecting Parkinson's disease. The model effectively classified Parkinson's patients and healthy individuals based on selected biomedical voice features. The feature selection process played a crucial role in improving model efficiency by eliminating redundant data and focusing on the most relevant attributes. The achieved accuracy indicates the model's potential as a valuable tool for assisting in early diagnosis. However, further improvements, such as hyperparameter tuning, advanced feature selection techniques, and ensemble learning, could enhance accuracy. Future work will focus on refining the model to achieve higher accuracy, and the updated results will be presented in a subsequent publication.

VI. CONCLUSION

In conclusion, this study demonstrates the potential of using voice-based biomarkers and machine learning techniques, specifically Support Vector Machine (SVM), for the early detection of Parkinson's Disease (PD). By leveraging voice features such as pitch, jitter, shimmer, and harmonic-to-noise ratio, the proposed methodology offers a promising, non-invasive, and cost-effective approach to diagnose PD at an early stage. The combination of robust preprocessing, feature selection, and SVM model tuning ensures optimal performance, enhancing the system's ability to accurately distinguish between PD-affected individuals and healthy controls.

The results indicate that the SVM-based classifier, particularly when fine-tuned with grid search and cross-validation, shows high potential in differentiating PD patients from healthy individuals. The use of performance metrics like accuracy, precision, recall, F1-score, and confusion matrices provides strong evidence of the model's effectiveness.

By integrating this voice-based diagnostic system into clinical practice, healthcare providers could benefit from a scalable and efficient tool for early PD screening, ultimately improving patient outcomes and accelerating the onset of appropriate treatments. This approach could not only streamline the diagnostic process but also pave the way for the broader application of machine learning in neurodegenerative disease detection, with the potential for future improvements and refinements.

REFERENCES

[1] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification," *Sensors*, vol. 23, no. 4, p. 2085, Feb. 2023. [Online].

[2] S. S. Sawant, R. S. Nair, and R. Patil, "Machine Learning Techniques for Early Detection of Parkinson's Disease," *International Journal of Scientific Research in Science and Technology*, vol. 12, no. 2, pp. 134–143, Mar.–Apr. 2025. [Online].

[3] H. Karimi Rouzbahani and M. R. Daliri, "Diagnosis of Parkinson's Disease in Humans Using Voice Signals," *Biomedical and Clinical Neuroscience*, vol. 2, pp. 12–20, 2020.

[4] A. Sharma and R. N. Giri, "Automatic Recognition of Parkinson's Disease via Artificial Neural Network and Support Vector Machine," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 5, no. 7, 2022.

[5] E. Avuçlu and A. Elen, "Evaluation of Train and Test Performance of Machine Learning Algorithms and Parkinson Diagnosis with Statistical Measurements," *Medical & Biological Engineering & Computing*, vol. 58, pp. 2775–2788, 2020.

[6] E. Nikookar, R. Sheibani, and S. E. Alavi, "An Ensemble Method for Diagnosis of Parkinson's Disease Based on Voice Measurements," *Journal of Medical Signals and Sensors*, vol. 9, pp. 221–226, 2019.

[7] A. Ouhmida, A. Raihani, B. Cherradi, and O. Terrada, "A Novel Approach for Parkinson's Disease Detection Based on Voice Classification and Feature Selection Techniques," *International Journal of Online and Biomedical Engineering*, vol. 17, pp. 111–130, 2021.

[8] K. Rana, A. Dumka, R. Singh, M. Rashid, and N. Ahmad, "An Efficient Machine Learning Approach for Diagnosing Parkinson's Disease by Utilizing Voice Features," *Biomedical and Clinical Neuroscience*, vol. 2, pp. 12–20, 2022.

[9] M. A. Little et al., "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[10] S. Sakar et al., "A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and the Use of the Tuned APSO for Model Parameter Optimization," *Biomedical Signal Processing and Control*, vol. 22, pp. 237–246, 2015.

[11] P. Das, S. S. Ghosh, and S. Das, "Parkinson's Disease Detection Using Voice Signal Features and Support Vector Machine," in *Proceedings of the International Conference on Intelligent Computing and Applications*, 2018, pp. 23–30.

[12] A. Tsanas et al., "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[13] S. R. Gunduz, "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets," *IEEE Access*, vol. 7, pp. 115540–115551, 2019.

[14] M. A. Haque and M. M. Rahman, "Performance Evaluation of Different Machine Learning Techniques in Detection of Parkinson's Disease," in *Proceedings of the 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, 2016, pp. 1–6.

[15] S. Prashanth et al., "High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning," *International Journal of Medical Informatics*, vol. 90, pp. 13–21, 2016.

[16] A. R. Khan et al., "A Robust and Effective Framework for the Detection of Parkinson's Disease Using Multilayer Perceptron in Combination with Feature Selection," *Computers in Biology and Medicine*, vol. 99, pp. 16–26, 2018.

[17] S. Sakar et al., "Assessing the Progression of Parkinson's Disease via Dynamic Vocal Fold Analysis," *Biomedical Signal Processing and Control*, vol. 48, pp. 177–186, 2019.

[18] M. A. Little et al., "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[19] A. Tsanas et al., "Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, Apr. 2010.

[20] S. R. Gunduz, "The Effect of Feature Extraction and Selection on Parkinson's Disease Detection," *Journal of Medical Systems*, vol. 43, no. 8, p. 255, 2019.