**Machine Learning Algorithms for Diagnosis of Parkinson's Disease Based on Voice Characteristics**

Polina Sai Manoj1, Kondreddy Sumanth Kumar Reddy2, Sambath M3, J. Thangakumar4

Department of Computer Science and Engineering,

Hindustan Institute of Technology and Science, Chennai, India.

[119113101@student.hindustanuniv.ac.in](mailto:119113101@student.hindustanuniv.ac.in)

**Abstract**

Bradykinesia, tremors, rigidity, and postural instability are symptoms of a degenerative neurological condition called Parkinson's disease (PD) affecting the motor movements. Deep brain stimulation, medication, and other therapies can all be used to treat the Parkinson's disease symptoms, yet as of today, there is no effective treatment. Parkinson's disease must be recognized as early and precisely as feasible for effective disease treatment and the development of new therapies. The goal of this work is to create a model that can detect Parkinson's disease based on relevant clinical and demographic variables. It employs a number of machine learning techniques, including Logistic Regression, Support Vector Machine (SVM), Gradient Boosting Classifier, K-Nearest Neighbors (KNN), Random Forest Classifier, and ensemble method, which is a voting classifier. A significant dataset that contains data on changes in speaking patterns was used to train the model. Using measures for accuracy, precision, recall, and F1-score, the machine learning model's performance is assessed and contrasted with that of the most widely used techniques for Parkinson's disease diagnosis.

**Keywords:** Gradient Boosting Classifier, Random Forest Classifier, SVM, KNN, Voting Classifier, Logistic Regression.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Introduction**

A degenerative neurological condition which affects the motor functions is called PD. Tremors, bradykinesia, stiffness, and postural instability are its defining features. Despite the significant consequences Parkinson's disease has on those who have it and their families, there is yet no treatment. The discovery of new treatments and the best possible disease care for Parkinson's disease depend on the earliest and most accurate diagnosis.

The PD motor abnormalities are brought on by a decline in dopamine levels in the brain as a result of this deterioration. Traditionally, Parkinson's disease has been diagnosed using clinical assessments, such as the examination of motor symptoms, medical history, and family history. Yet, these methods are arbitrary and susceptible to error. In recent years, machine learning algorithms have demonstrated potential as a method of Parkinson's disease detection. The objective of this research is to design a machine-learning model which can diagnose Parkinson's disease. After being trained on a pertinent dataset including data on changes in speaking patterns, the model will merge Gradient Boosting Classifier and Random Forest Classifier utilizing techniques like Logistic Regression, SVM, KNN, and ensemble approaches. Using evaluation metrics, the performance is assessed. The program has the potential to speed up and improve Parkinson's disease has been identified, and provide helpful data for the development of new, more effective therapies.

1. **Related Work**

Andrea Sabo et al. [1] advised evaluating the concomitant validity of two gait measurement techniques in Parkinson's disease patients. A Zeno-instrumented walkway, which analyses gait characteristics such as stride length, stride time, and cadence, was utilised in the first method. The second technique was video-based gait analysis, which includes gathering gait features from video recordings of people walking. The study discovered a significant connection between the video-based gait analysis and the gait parameters recorded using the Zeno instrumented walkway, indicating that both techniques are viable and reliable ways to quantify gait in people with Parkinson's disease.

Rui Guo et al. [2] manifest offered research that suggests a novel machine learning-based technique for evaluating gait in PD patients. The technique makes use of a deep learning system known as a multi-scale sparse graph convolutional network (MS-SGCN), which can evaluate gait data at various sizes and extract key properties. According to the study, the MS-SGCN performed better than conventional machine learning algorithms in terms of determining Parkinsonian gait. The findings of the study imply that the MS-SGCN may be an effective instrument to assess gait in PD patients since it can provide more accurate and detailed information about these people's gait patterns.

Robbin Romijnders et al. [3] proposed a survey that examined elderly folks and those with Parkinson's disease in order to test the precision of gait event detection utilizing inertial measurement unit (IMU) sensors. IMUs are tiny sensors that may be fastened to the body and used to monitor acceleration and movement. In both older persons and those with Parkinson's disease, the study demonstrated that IMU-based gait event recognition was reliable for identifying gait events including heel strikes and toe-offs during curved walking and turning. The study also discovered that Parkinson's disease-related alterations in gait patterns, such as shorter steps and more step variability could be precisely detected by the IMU sensors.

Kimberley-Dale Ng et al [4] recommended using computer vision techniques to evaluate older persons with dementia's mobility and fall risk. The study discovered that utilizing computer vision techniques, gait metrics including speed, step length, and stride duration could be precisely quantified. The study also discovered that these gait characteristics were related to fall risk in dementia-affected older persons, more prone to falling were individuals with slower gait rates, shorter step lengths, and longer stride durations. Computer vision algorithms can provide objective and non-invasive gait measures to track changes in mobility over time and identify those who are more likely to fall.

Shallu Sehgal et al. [5] explained the fundamental ideas and procedures of optimal grasshopper algorithms, as well as examples of their uses in diverse industries. analysing the outcomes and constraints of earlier research that employed enhanced grasshopper algorithms in order to spot PD evaluating the performance of the improved grasshopper algorithm in comparison to other techniques, such as conventional clinical evaluations and other machine learning algorithms, for the determination of PD. The accuracy, sensitivity, and non-invasiveness of the improved grasshopper algorithm were discussed, along with some of its weaknesses and restrictions on how effectively it may be applied to other illnesses or populations.

Armando de Jesús Plasencia Salgueiro et al. [6] formulated using a smartphone app to track people with Parkinson's disease gait and gauge their adherence to treatment a possible study subject. To assess the gait data and categorize the user's state, the app employs deep reinforcement learning algorithms. During the smartphone-based gait assessment, data on the subject's walking gait is gathered using the smartphone's accelerometer and gyroscope. In order to train the model, the algorithms employ a reward system in which successful results (like a steady gait) are rewarded and unsuccessful outcomes (like a shuffling gait) are penalized. As a result, the algorithm can grow and develop over time.

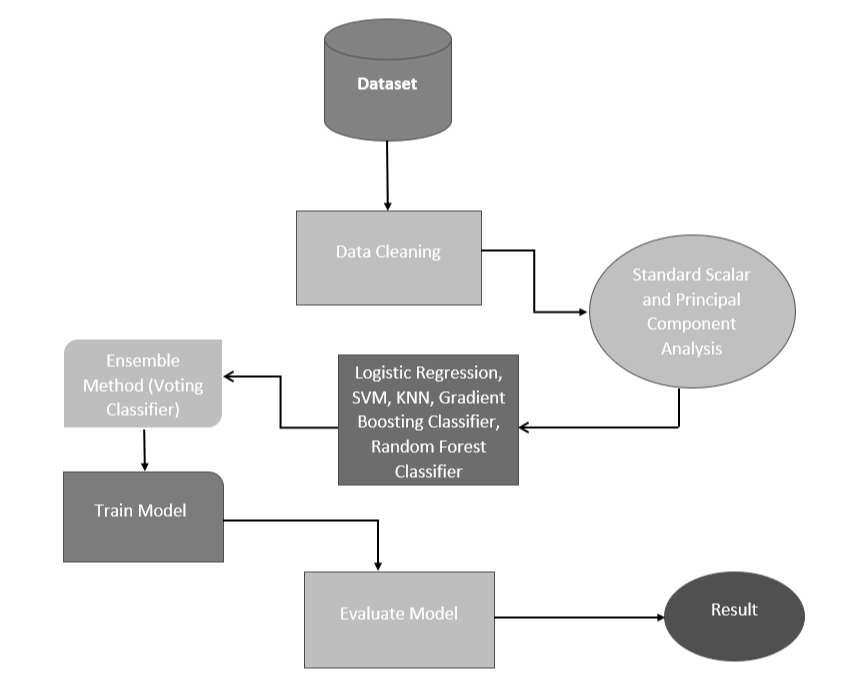
Ashena Gorgan Mohammadi et al. [7] an artificial neural network that can learn to encode and decode data is called an autoencoder. When Parkinson's illness is identified, autoencoders are trained on the Parkinson's disease sufferers' speech patterns and healthy individuals in order to identify patterns or characteristics that distinguish the two groups. Pitch, amplitude, and frequency modulation are some of the numerous approaches that may be used to study vocal features. These vocal traits may be represented by autoencoders as a group of latent variables, which can subsequently be utilised to provide a diagnosis.

1. **Methodology**

The needed dataset for the proposed system is gathered from multiple sources and databases. The dataset that is being gathered is in.csv file format.

**Dataset description**

Patients with Parkinson's disease donated the voice feature dataset that was utilised in this investigation. The Kaggle platform is where the dataset was gathered. The dataset has 756 cases and 23 characteristics, including jitter, shimmer, and harmonics-to-noise ratio (HNR). The dataset was pre-processed using standard scalar normalization and principal component analysis [8]. There were no missing values or outliers present in the dataset.



**Fig.1** Proposed architecture

**Normalisation**

Two pre-processing techniques are employed, they are standard scalar and principal component analysis (PCA), to pre-process the data prior to training our machine learning model [9]. Standard scalar is a commonly used technique that involves scaling the characteristics should have a unit variance and a zero mean. To make certain that all features are on a similar scale and to prevent features with larger values from dominating the model. In addition, we used PCA to make the data's dimensions smaller. PCA is a technique that involves transforming the original features into a new set of features that are linearly uncorrelated and capture the most important variables in the data. This is done to reduce the number of features and to remove any redundant or irrelevant features that may negatively affect the performance of the model.

Specifically, we applied a standard scalar to each feature of the dataset to scale the values and ensure that all features are on a similar scale [10]. To make the data less dimensional, we then used PCA to the scaled data. By reducing the number of features, we were able to simplify the model and improve its performance. Together, standard scalar and PCA allowed us to pre-process the data in a way that improved the performance of our machine-learning model. These techniques are commonly used in machine learning to normalize the data and reduce its dimensionality, which can lead to better performance and more efficient computation.

**Machine learning algorithms**

The key components of Logistic Regression include the input variables, output probability, and parameters that are learned during training. The algorithm is particularly useful when the output variable is binary and the input variables are continuous or categorical. We used L1 regularization to select important features and tuned the hyperparameters using cross-validation.

The K-Nearest Neighbors (KNN) works by finding the KNN of a new input sample and assigning it to the class that the majority of the K neighbors belong to. The key components of the KNN algorithm include the input samples, output classes, K value, and distance metric used to calculate the similarity between samples. The algorithm is particularly useful when the decision boundary is nonlinear or when the dataset has a small number of features. We utilized Euclidean distance as the distance metric and 5-fold cross-validation to choose the best K value.

The Support Vector Machine (SVM) method divides the input data into two classes with the greatest possible margin by locating the ideal hyperplane. Grid search was performed to fine-tune the hyperparameters while employing a Radial Basis Function (RBF) kernel [11].

The Gradient Boosting Classifier algorithm works by iteratively adding new weak learners to the model to improve its accuracy over time. The key components of the Gradient Boosting Classifier algorithm include the weak learner, the loss function, and the boosting process used to improve the model's accuracy. The algorithm is particularly useful when dealing with complex, nonlinear relationships between input variables. We used a combination of hyperparameter tuning and early stopping to prevent the overfitting of the model.

The Random Forest algorithm uses bagging and random feature selection to improve the model's accuracy and reduce overfitting. The key components of the Random Forest algorithm include the decision tree, bagging, and random feature selection. The algorithm is particularly useful when dealing with complex, nonlinear relationships between input variables. We used a combination of hyperparameter tuning and cross-validation to avoid the model from being overfitted [12].

**Ensemble method**

In order to enhance how well our machine learning model performed, we investigated ensemble approaches in this study. A group of machine learning approaches known as ensemble methods combine several models to provide predictions. Specifically, we used two different algorithms, SVM and Random Forest Classifier, and combined them using the voting classifier.

The voting classifier is a simple ensemble method that aggregates the predictions of multiple models by taking the majority vote. In our case, we combined the predictions of the SVM and Random Forest Classifier to make the final prediction.

**Evaluation Methods**

In this work, we applied two common evaluation methods, cross-validation with 5-fold validation and grid search. Together, cross-validation and grid search allowed us to evaluate how well our machine-learning model is performing on multiple subsets of the data and to identify the best hyperparameters for the model [13]. This approach enabled us to obtain reliable estimates of the model's performance and to optimize its performance for our specific dataset.

**Evaluation Metrics**

We used several evaluation metrics which allow us to measure the accuracy, precision, recall, and overall performance of the model. Additionally, we used receiver operating characteristic (ROC) curves and confusion matrices, also known as error matrices, to illustrate the classification performance of the suggested machine learning methods. Additionally, the ROC curve's area under the curve was utilised to determine effectiveness of a model. Higher AUC values are a sign of better categorization results. Using the values in the confusion matrix, we can calculate metrics such as

It provides a clear and concise summary of predictions made by the model and can be used to calculate various metrics to assess the model's accuracy [14], [15].

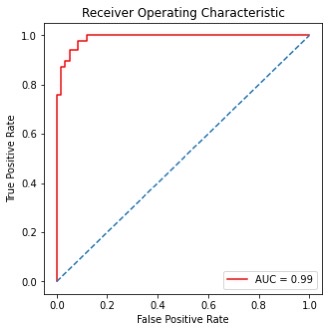
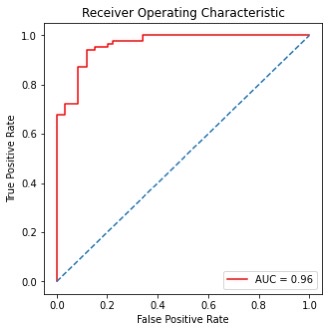
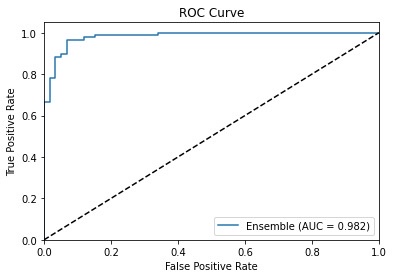
1. **Results and Discussion**

Five machine learning algorithms and the ensemble technique were used. Their categorization outcomes for the two classification tasks in the 5-fold cross-validation procedure is displayed in **Table I**. The SVM and voting classifier did the best and achieved a 0.93 accuracy in the two-class classification task, followed by the other classifiers that are the highest precision was obtained by Random Forest (0.89), followed by KNN (0.84), logistic regression (0.84), and gradient boosting (0.75).

**TABLE I**

Classification Outcomes Using the Ensemble Technique and All Five Machine Learning Algorithms

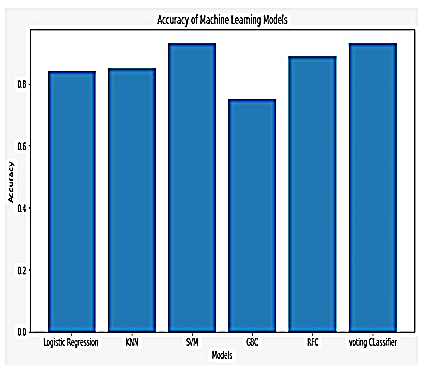
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifiers** | **Accuracy** | **Recall** | **Precision** | **F1-Score** | **AUC** |
| **Logistic Regression** | 0.84 | 0.84 | 0.85 | 0.84 | 0.88 |
| **KNN** | 0.84 | 0.84 | 0.85 | 0.85 | 0.92 |
| **SVM** | 0.94 | 0.94 | 0.94 | 0.94 | 0.99 |
| **Gradient Boosting Classifier** | 0.75 | 0.75 | 0.81 | 0.72 | 0.90 |
| **Random Forest Classifier** | 0.90 | 0.90 | 0.90 | 0.90 | 0.96 |
| **Voting Classifier** | 0.94 | 0.94 | 0.94 | 0.94 | 0.98 |

********

**(a) (b) (c)**

**Fig 3:** Roc curve for the five-machine learning algorithm and ensemble method. (a) Support vector Machine. (b) Random Forest and (c) Voting classifier.

**Fig [3]** (a) shows a ROC curve for an SVM model and AUC for this model is 0.99. **Fig [3]** (b) shows a ROC curve for a random forest classifier and AUC for this model is 0.96. **Fig [3]** (c) shows a ROC curve for a voting classifier and AUC for this model is 0.98 indicating that it SVM and voting classifier performs well in distinguishing between cases and non-cases.



**Fig: 4** Accuracies of Models

Within the task of two-class categorization, the findings showed that the SVM and voting classifier had the highest accuracy rates of 0.93. While the other classifiers achieved slightly lower accuracy rates. The results also indicated that the highest precision was obtained by Random Forest (0.89), followed by KNN (0.84), logistic regression (0.84), and gradient boosting (0.75). These findings suggest that these algorithms can also be useful in diagnosing PD, although their performance is slightly lower than that of SVM a voting classifier. Therefore, these algorithms ca be used as alternative or complementary methods to SVM and voting classifiers in PD diagnosis based on vocal features.

The ROC curve analysis also revealed that SVM and voting classifier had the highest AUC values, indicating that they were better at distinguishing between PD cases and non-cases. The AUC values for SVM and voting classifier were 0.99 and 0.98, respectively. These findings suggest that SVM and voting classifier are more accurate than the other algorithms in identifying PD cases based on vocal features. Nonetheless, it's crucial to keep in mind that these algorithms' performance may change based on the data source and the features used for diagnosis.

1. **Conclusion**

This study tested how well multiple machine learning algorithms could predict Parkinson's illness using speech data. Support Vector Machines (SVM) and Voting Classifiers produced the best accuracy rates, with an overall accuracy of 93%, according to the results. These discoveries have significant repercussions in order to detect Parkinson's disease early and begin therapy, which can be essential for effective therapeutic outcomes. To ascertain the efficacy of these algorithms with larger and more varied datasets, additional study is necessary. Future research may also examine the use of different features, such as movement data or brain imaging.

**References**

[1] Sabo, A., Gorodetsky, C., Fasano, A., Iaboni, A., & Taati, B. (2022). Concurrent Validity of Zeno Instrumented Walkway and Video-Based Gait Features in Adults With Parkinson’s Disease. *IEEE Journal of Translational Engineering in Health and Medicine*, *10*, 1-11.

[2] Guo, R., Shao, X., Zhang, C., & Qian, X. (2021). Multi-scale sparse graph convolutional network for the assessment of Parkinsonian gait. *IEEE Transactions on Multimedia*, *24*, 1583-1594.

[3] Romijnders, R., Warmerdam, E., Hansen, C., Welzel, J., Schmidt, G., & Maetzler, W. (2021). Validation of IMU-based gait event detection during curved walking and turning in older adults and Parkinson’s Disease patients. *Journal of neuroengineering and rehabilitation*, *18*(1), 28.

[4] Ng, K. D., Mehdizadeh, S., Iaboni, A., Mansfield, A., Flint, A., & Taati, B. (2020). Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. *IEEE journal of translational engineering in health and medicine*, *8*, 1-9.

[5] Sehgal, S., Agarwal, M., Gupta, D., Sundaram, S., & Bashambu, A. (2020). Optimized grass hopper algorithm for diagnosis of Parkinson’s disease. *SN Applied Sciences*, *2*, 1-18.

[6] Salgueiro, A. D. J. P., Shichkina, Y., García, A. G., & Rodríguez, L. G. (2021). Parkinson’s disease classification and medication adherence monitoring using smartphone-based gait assessment and deep reinforcement learning algorithm. *Procedia Computer Science*, *186*, 546-554.

[7] Mohammadi, A. G., Mehralian, P., Naseri, A., & Sajedi, H. (2021). Parkinson's disease diagnosis: The effect of autoencoders on extracting features from vocal characteristics. *Array*, *11*, 100079.

[8] Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić-Jovičić, M., & Kostić, V. S. (2019). Artificial intelligence for assisting diagnostics and assessment of Parkinson’s disease—A review. *Clinical neurology and neurosurgery*, *184*, 105442.

[9] Gunduz, H. (2021). An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomedical Signal Processing and Control*, *66*, 102452.

[10] Mohammed, M. A., Elhoseny, M., Abdulkareem, K. H., Mostafa, S. A., & Maashi, M. S. (2021). A multi-agent feature selection and hybrid classification model for Parkinson's disease diagnosis. *ACM Transactions on Multimidia Computing Communications and Applications*, *17*(2s), 1-22.

[11] Haq, A. U., Li, J. P., Memon, M. H., Malik, A., Ahmad, T., Ali, A., ... & Shahid, M. (2019). Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson’s disease using voice recordings. *IEEE access*, *7*, 37718-37734.

[12] Sathiya, T., Reenadevi, R., & Sathiyabhama, B. (2021). Random Forest Classifier based detection of Parkinson’s disease. *Annals of the Romanian Society for Cell Biology*, 2980-2987.

[13] Masud, M., Singh, P., Gaba, G. S., Kaur, A., Alroobaea, R., Alrashoud, M., & Alqahtani, S. A. (2021). CROWD: crow search and deep learning-based feature extractor for classification of Parkinson’s disease. *ACM Transactions on Internet Technology (TOIT)*, *21*(3), 1-18.

[14] Tsai, C. C., Chen, Y. L., Lu, C. S., Cheng, J. S., Weng, Y. H., Lin, S. H., ... & Wang, J. J. (2022). Diffusion tensor imaging for the differential diagnosis of Parkinsonism by machine learning. *Biomedical Journal*.

[15] Mittal, V., & Sharma, R. K. (2021). Machine learning approach for classification of Parkinson disease using acoustic features. *Journal of Reliable Intelligent Environments*, *7*(3), 233.