Early diagnosis of Parkinson’s disease using cluster based Feature Selection

Machine Learning

22AIE213

V.Varshini(AIE22146)

Y.Bhagyasree(AIE22147)

D.Likitha Kamali(AIE22148)

Under the guidance of:

Dr.Madhusudana Rao Nalluri

Associate Professor   
School of Computing  
Department of Computer Science and Engineering

1. **INTRODUCTION**

Parkinson’s is a neurodegenerative brain disease that progresses gradually. Neurodegenerative refers to the loss of brain tissue. In the human brain, dopamine is normally produced by distinct brain cells in certain areas. These cells are localized in a specific region of the substratum nigra in the brain [1]. The substantia nigra and other brain areas that regulate body motions communicate with each other through the neurotransmitter dopamine. Dopamine enables individuals to move in a fluid and melodic manner.

Parkinson’s disease motor symptoms occur after 60-80% of dopamine-producing cells are destroyed. This is because insufficient dopamine is produced. Parkinson’s disease initially affects the lower brain stem, olfactory pathways, and enteric nervous system. From these areas, Parkinson’s disease extends to the brain’s upper regions, including the substantia nigra and the brain shell [2].

It is believed that the illness starts several years before the first symptoms, which include constipation, sleep difficulties, and a diminished or absent sense of smell. Shakiness and a slowdown in motion. Also, voice problems affect 90% of PD patients. In order to stop the disease’s progression, we are searching for techniques to identify these non-motor signs as soon as they arise during the illness.

In this paper, the proposed diagnosis method utilises feature selection and classification processes. The feature selection is done through Apriori and Clustering methods using correlation. LGBMclassifier and ensemble techniques of voting ensemble and stacking ensemble were used for the classification of Parkinsion’s patients.

1. **LITERATURE SURVEY**

In this paper [3] researchers used 12 datasets from the PPMI database to develop an AI system for Parkinson’s disease detection. They employed the majority voting algorithm for labeling and tested five machine learning algorithms, with SVM achieving 100% accuracy. Feature selection was conducted using four tree-based models, with Gradient Boosted Decision Tree identifying key features. An Artificial Neural Network achieved 91.41% accuracy using these features.

This study [4] utilized speech signals from 252 subjects to diagnose Parkinson’s disease (PD) using machine learning algorithms. Methods included testing for dysphonia in continuous speech and applying language signal features as input to the algorithms. Feature extraction based on clinical experience was also used to analyse the speech signals. The integrated classifiers achieved a diagnostic accuracy of up to 95%.

The study [5] used a dataset of 195 voice recordings from 31 patients to diagnose Parkinson’s disease, employing classification algorithms with feature selection based on Pearson correlation The Random Forest algorithm achieved an accuracy of 95.42%, balanced accuracy of 93.98%, and an F1 score of 0.98.

The study [6] compared machine learning models for early detection of Parkinson’s Disease using four main symptom modalities: tremor at rest, bradykinesia, rigidity, and voice impairment. Methods included Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-nearest neighbors (KNN), Stochastic Gradient Descent (SGD), and Gaussian Naive Bayes (GNB), as well as ensemble approaches like Random Forest (RF), Adaptive Boosting (AB), and Hard Voting (HV). The Random Forest algorithm applied to the Static Spiral Test for detecting tremor achieved the highest accuracy of 99.79%.

The study [7] employed feature selection methods, including Feature Importance and Recursive Feature Elimination (RFE), and classification algorithms such as Classification and Regression Trees (CART), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). The combination of SVM and RFE achieved the highest accuracy of 93.84% in diagnosing Parkinson’s disease using the least number of voice features.

The study [8] analyzed audio data from 30 individuals using four machine learning models: Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. The Random Forest classifier achieved the highest performance with a detection accuracy of 91.83% and a sensitivity of 0.95.

The study [9] used voice data from a Parkinson’s voice database available at UCI's machine learning repository and applied the XGBoost algorithm to diagnose Parkinson’s disease. The XGBoost algorithm was found to provide the highest accuracy with 85.71% among the tested algorithms.

The study [10] used an Artificial Neural Network (ANN) and nineteen other ML algorithms to predict Parkinson’s disease using two different sets of voice data. Cross-validation was used instead of the train-test split approach, along with Optimal Hyperparameters Tuning. The best-performing algorithms were combined into two ensemble voting classifiers, achieving accuracies of 96.41% and 97.35%

In the study [11] focused on vocal impairments in Parkinson’s disease (PD), researchers evaluated decision tree ensemble methods including AdaBoost, random forest, and decision tree methods using datasets with imbalanced classes. They compared performance metrics such as precision, recall, F1-score, AUROC, and geometric mean with and without techniques to balance the dataset (e.g., RUSBoost). Feature selection methods like lasso and information gain identified the top 10 features for classification. AdaBoost combined with information gain achieved the highest F1-score of 0.903 among the tested methods.

1. **Literature Survey Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| s.no. | name | method | proposed algorithm | Accuracy |
| [5] | Early diagnosis of Parkinson’s disease using machine learning algorithms | SVM, ANN, CART | SVM | 93.84% |
| [6] | Early detection of Parkinson’s disease using machine learning | Support Vector Machine, (SVM), Random Forest, K-Nearest Neighbors (KNN) and Logistic Regression models | RandomForest classifier | 91.83% |
| [7] | EARLY DETECTION OF PARKINSONS USING MACHINE LEARNING | Logistic regression, SVM, Random Forest Regressor and K nearest neighbors | Logistic regression | 85.70% |
| [3] | A Comparison of Machine Learning Algorithms for Parkinson’s Disease Detection | Decision Tres, Random Forest, Logistic Regression, SVM, Naïve BAyes | RandomForest classifier | 95.42% |
| [2] | Parkinson disease prediction using machine learning-based features from speech signal | Random Forest,Logistic regression, KNN, DNN, Vote | Vote | 95.00% |
| [10] | Diagnostic classification of Parkinson’s disease based on non-motor manifestations and machine learning strategies | AdaBoost, Bagging, DT, HNN, MLP, NB, RF, RIPPER, SVM | SVM | 86.30% |
| [13] | Early diagnosis of Parkinson disease using Machine Learning Technique | L1-Norm SVM, Averaged Perceptron, Bayes PointMachine, Boosted DecisionTree, Decision Forests,Neural Networks, K-NN,SVM, Naïve Bayes, Metaclassifier, Decision tree,Bagging, AdaBoost,Gradient Boost, RF | L1-Norm SVM | 99% |
| [14] | A Comparative Study of Early Detection of Parkinson’s Disease using Machine Learning Techniques | LR, SVM, KNN, DT, SGD, RF, AB, HV | Random Forest classifier | 99.79% |
| [12] | A hybrid system for Parkinson’s disease diagnosis using machine learning techniques | naive bayes, k-nearest-neighbors, and random forest | genetic algorithm | 95.58% |

1. **Research Gap**

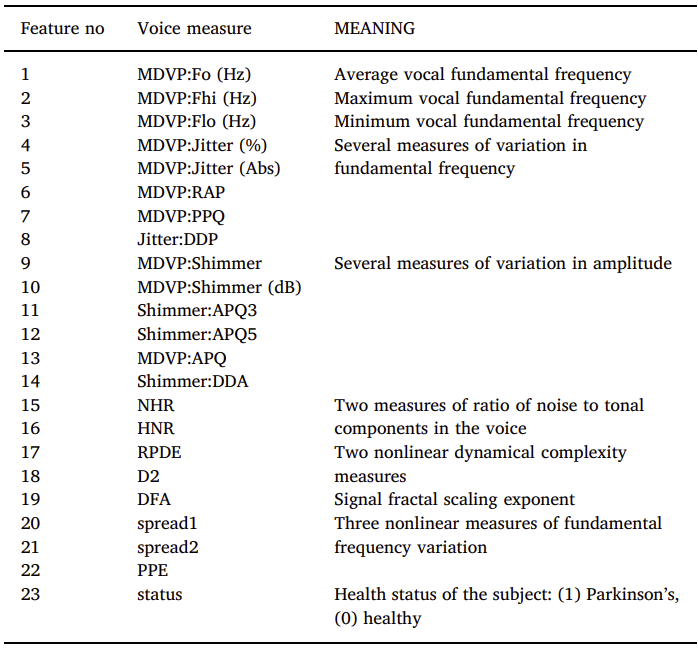
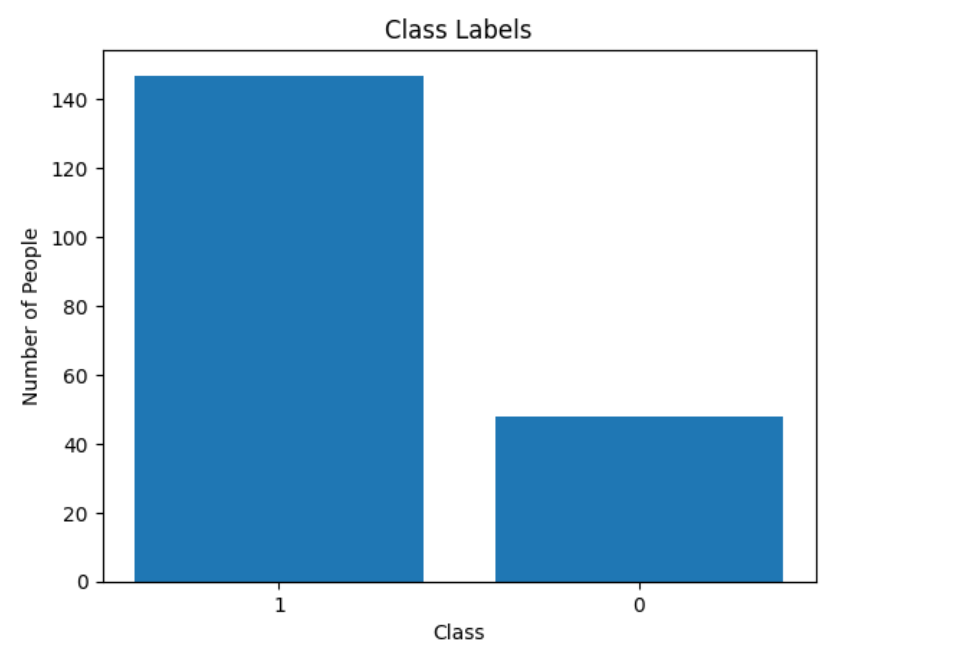
Although many models achieve good accuracies, using all features leads to high computational time and cost. Additionally, relying on MRI-based or motion-based features complicates the feature extraction process. These methods often require advanced equipment, specialized software, and intricate preprocessing steps, making them less practical for widespread use. In contrast, voice-based features offer a more accessible and efficient alternative. They are easier to extract, requiring less sophisticated equipment and simpler processing techniques. This ease of extraction, combined with their lower computational demands, makes voice-based features a more feasible option for many applications.

1. **Proposed Solution**

* Using Voice Based features which are easy to extract than MRI or motion based features
* Using SMOTE to balance the data to better improve the performance of the model
* Using Apriori based Feature Selection to select the most important features
* Using KMeans Clustering and SOM clustering for selecting the features
* Using Boosting and Ensemble techniques to better classify the data with minimum computational time and cost

1. **Materials**

The dataset used in the experiments of this study consists of the features obtained from the speech signals of 31 people at the National Centre for Voice and Speech, Denver, Colorado. The dataset was created by Max Little from University of Oxford and donated to UCI Machine Learning Repository [13]. 23 of the 31 people have PD and 8 of them are the control group. There are 195 biomedical voice measurements in the dataset. Status column in the database defines the class and gets 0 for healthy, 1 for PD. The “status” feature gives the health state of the subject; There are 147 people diagnosed and 48 are healthy. The table below gives detailed information about the features of the dataset.



1. **Details of the Hardware and Software used:**

The proposed model is run on a system equipped with a 11th Gen Intel Core i5-11400H processor, 16 GB RAM, and x64-based processor. The software used is Jupyter Notebook version 3.11.5. Python libraries like sklearn, numpy, pandas, matplotlib, xgboost, lightgbm and minisom have also been used.

1. **Methods**

**Dataset Split-up:**

The dataset of the proposed model is split up in the 3:2 ratio i.e, 70% of the data is used for training the machine learning model while 30% is used for testing the performance of the model.

**Feature Selection Methods:**

**Apriori Algorithm:**

The apriori algorithm [16] helps in feature selection by identifying the most relevant and significant features based on their frequency and associations within the dataset. Once the frequent features are identified, the apriori algorithm generates association rules that describe how the presence of certain features implies the presence of others.

**Clustering:**

Clustering is used to group similar data points or instances into clusters. It can be powerful method for feature selection by identifying groups of similar features within the data. Two clustering methods are used in our proposed model for feature selection:

***K Means*:**

It is an unsupervised learning algorithm which groups the data into k different clusters. Using K-Means [17] for feature selection involves utilizing the clusters generated and calculating the mutual information between the features using correlation. This generates a certain percent of topmost important features based on a threshold.

**SOM (Self Organising Maps) :**

SOM is an Artificial Neural Network which can be used for clustering [18] and dimensionality reduction allowing us to reduce the complexity by following and unsupervised learning and training its network through a competitive learning algorithm. The important features are found by calculating the correlation of features of a cluster which are generated by the SOM algorithm.

**CLASSIFIERS:**

**VOTING ENSEMBLE:**

It is an ensemble learning method [19] that combines several machine learning models to make predictions. The concept is to combine each model’s prediction and use this aggregate value to classify a data point. The proposed model uses the voting ensemble of Random Forest (RF), XGBClassifier (XGB), SVM and KNN.

RF uses a collection of decision trees where each tree is constructed using a random subset of the dataset to find random subsets of features during partition. XGB (Extreme Gradient Boosting) uses a boosting technique to create an ensemble model by having each weak learner correct the mistakes of its predecessors. It minimises the cost function by repeatedly changing the model’s parameters based on the errors. SVM finds the optimal hyperplane that best separates the data. KNN finds the best K- nearest neighbours to a given data pints based on a distance metric.

**STACKING ENSEMBLE:**

Stacking ensemble combines predictions from multiple different models(base models) using another model(meta-model). First, each base is trained separately on the data to make predictions. These predictions become input to the meta-model, which learns how to best combine them to make a final prediction. This approach often improves accuracy compared to using just one-model, as it captures diverse perspectives and patterns in the data. The classifiers used in this ensemble technique are the same as the voting ensemble.

**LIME:**

LIME (Local Interpretable Model-agnostic Explanations) [20] is an explainable AI technique that provides insights into individual predictions made by machine learning models. It works by perturbing the input data around a specific instance and observing the resulting changes in the model's output. LIME highlights the features that most significantly influence the prediction for that instance. This helps users understand why a model made a particular decision, enhancing transparency and trust in complex machine learning systems.

**LGBMClassifier:**

Light Gradient Boosting Machine classifier [21] is a gradient boosting framework that uses tree-based learning algorithms. It works by sequentially combining multiple decision trees to improve predictive accuracy. It starts by initializing with an ensemble of decision trees. Each tree is built sequentially, where each subsequent tree focuses on reducing the errors of the previous trees. This approach allows LightGBM to iteratively improve its predictions, with each tree learning from the mistakes of the ensemble.

1. **Flowchart**

Speech Features Dataset

SMOTE

Feature Selection (Apriori/Clustering)

Classification

Performance Evaluation

LGBMClassifier

Accuracy

Voting Ensemble

Precision

Recall

Stacking Ensemble

F1 Score

1. **Results and Discussion:**

**Evaluation Metrics of previous paper results:**

The evaluation metrics used in this paper are Accuracy, Precision, Recall, F1 Score.

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision = TP/(TP + FP)

Recall = TP/(TP + FN)

F1-Score = (2\*Precision\*Recall)/(Precision + Recall)

*True Positive (TP)* is the correctly predicted positive cases.

*False Positive (FP)* gives the incorrectly predicted positive cases.

*True Negative (TN)* gives the correctly predicted negative cases.

*False Negative (FN)* gives the incorrectly predicted negative cases.

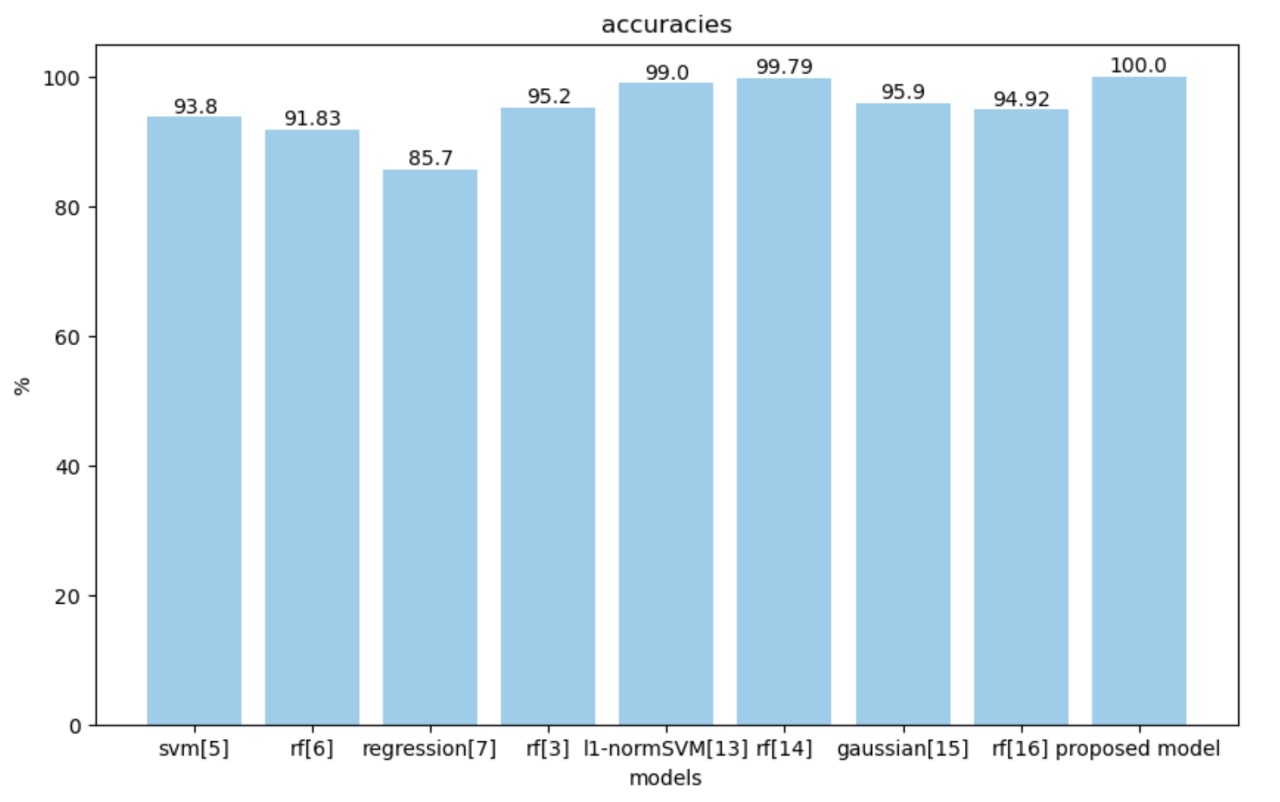
*Accuracy* is the proportion of correctly predicted cases out of the total cases. Precision: Proportion of correctly predicted positive cases out of all predicted positive cases.

*Recall (Sensitivity)* is the proportion of correctly predicted positive cases out of all actual positive cases.

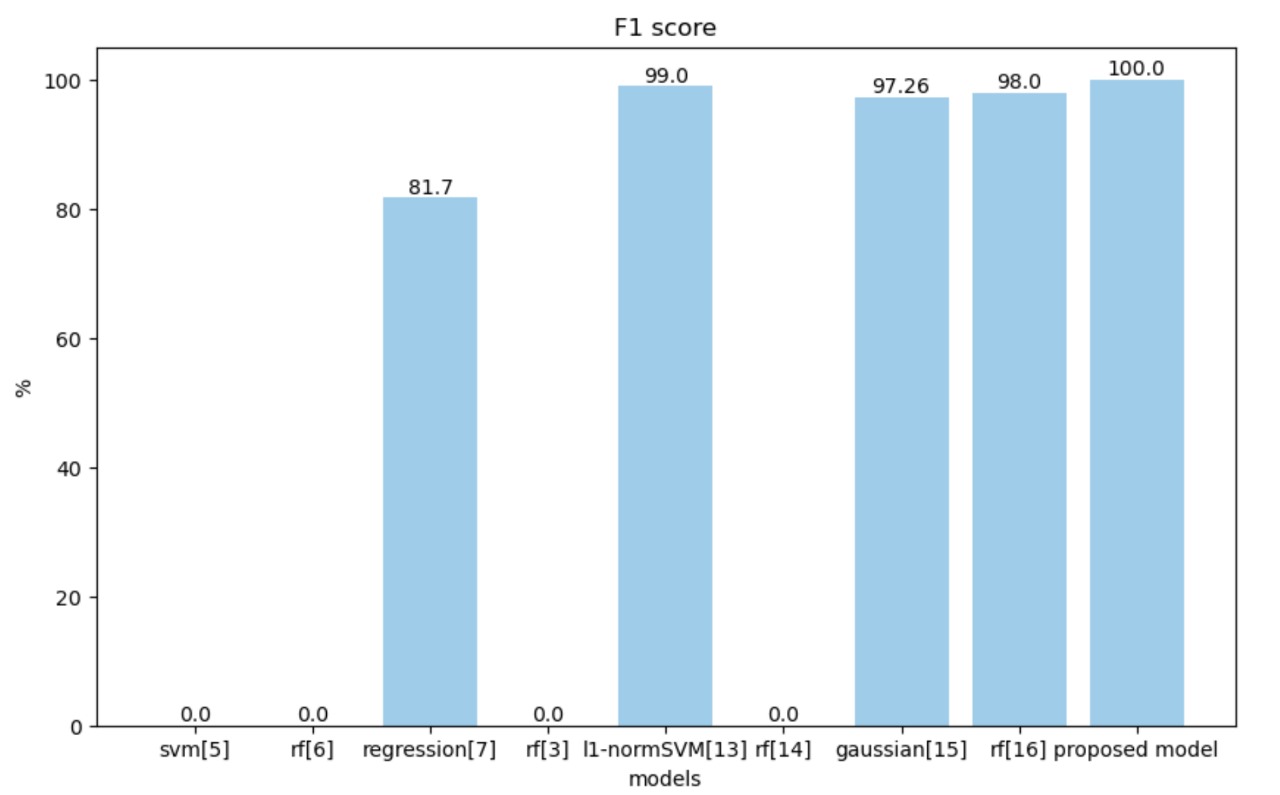
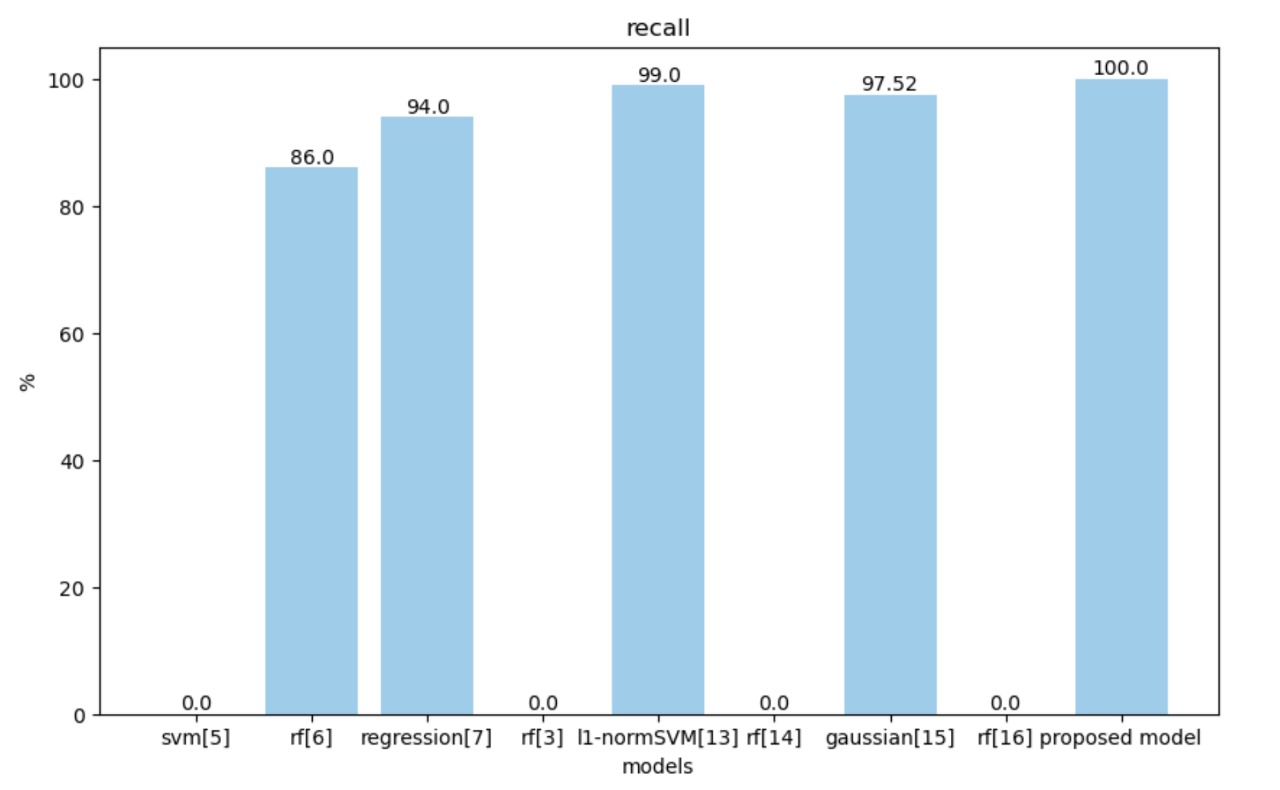
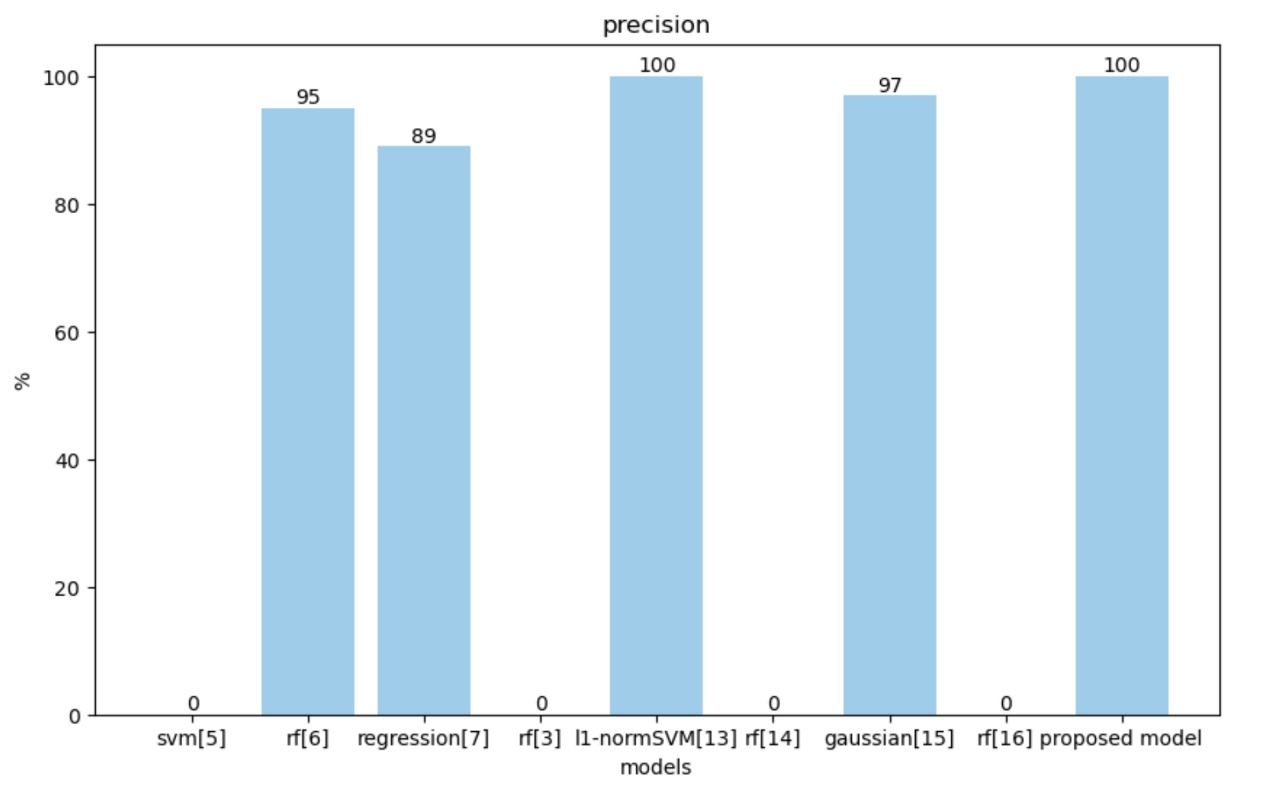
*F1 Score* gives the harmonic mean of precision and recall.

***Before Feature Selection:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifier | accuracy | precision | recall | F1 Score | Specificity |
| LightGBM | 84.74 | 84.44 | 95.00 | 96.55 | 81.81 |
| Voting ensemble | 84.74 | 84.44 | 95.00 | 89.41 | 63.15 |
| Stacking ensemble | 83.05 | 84.09 | 92.5 | 88.09 | 63.15 |

***After Feature Selection:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| classifier | accuracy | precision | recall | F1 Score | Specificity |
| LightGBM  (K-means) | 98.86 | 100 | 97.77 | 98.87 | 100 |
| LightGBM  (SOM) | 97.70 | 100 | 95.55 | 97.7 | 97.6 |
| Voting ensemble  (apriori) | 100 | 100 | 100 | 100 | 100 |
| Stacking ensemble  (apriori) | 100 | 100 | 100 | 100 | 100 |



1. **Comparison with the state-of-the-art results**

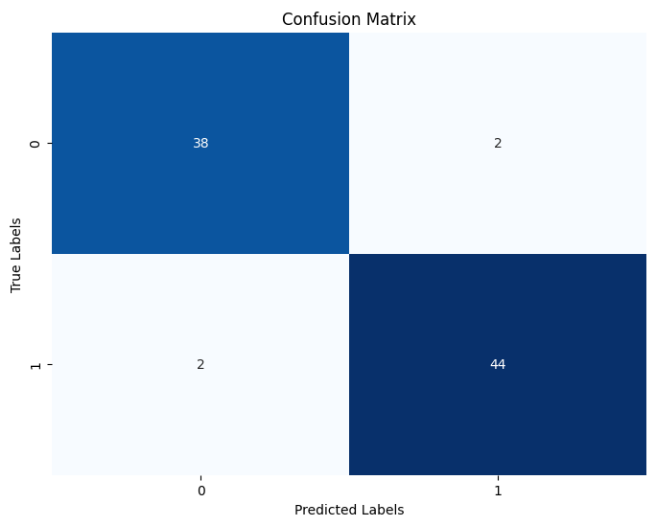
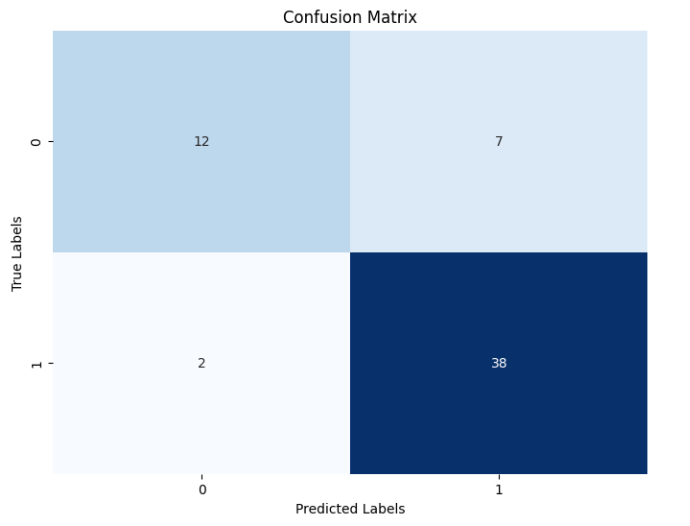
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **s.no.** | **Name** | **year** | **method** | **Proposed algorithm** | **accuracy** | **precision** | **recall** | **F1-score** |
| [5] | Early diagnosis of Parkinson’s disease using machine learning algorithms | 2020 | SVM, ANN, CART | SVM | 93.84% | NA | NA | NA |
| [6] | Early detection of Parkinson’s disease using machine learning | 2023 | Support Vector Machine, (SVM), Random Forest, K-Nearest Neighbors (KNN) and Logistic Regression models | Random Forest | 91.83% | 95% | 86% | NA |
| [7] | EARLY DETECTION OF PARKINSONS USING MACHINE LEARNING | 2024 | Logistic regression, SVM, Random Forest Regressor and K nearest neighbors | Logistic Regression | 85.70% | 89% | 94% | 81.7% |
| [3] | A Comparison of Machine Learning Algorithms for Parkinson’s Disease Detection | 2023 | Decision Tres, Random Forest, Logistic Regression, SVM, Naïve BAyes | Random Forest | 95.42% | NA | NA | NA |
| [2] | Parkinson disease prediction using machine learning-based features from speech signal | 2024 | Random Forest,Logistic regression, KNN, DNN, Vote | Vote | 95.00% | NA | NA | 97% |
| [10] | Diagnostic classification of Parkinson’s disease based on non-motor manifestations and machine learning strategies | 2024 | AdaBoost, Bagging, DT, HNN, MLP, NB, RF, RIPPER, SVM | SVM | 86.30% | NA | NA | 90% |

1. **Observation:**

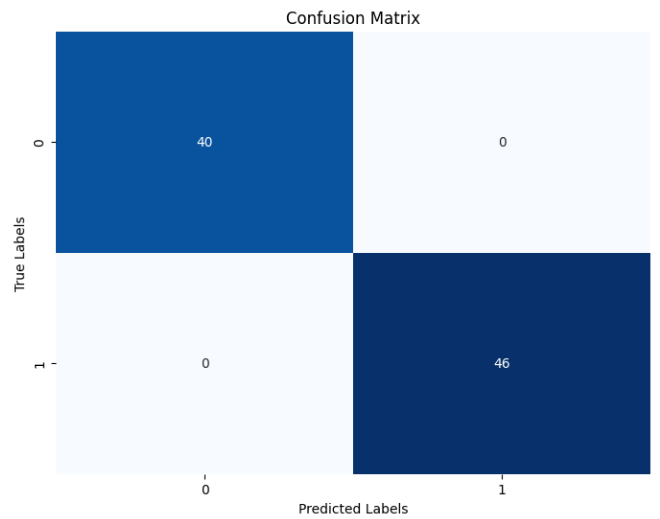
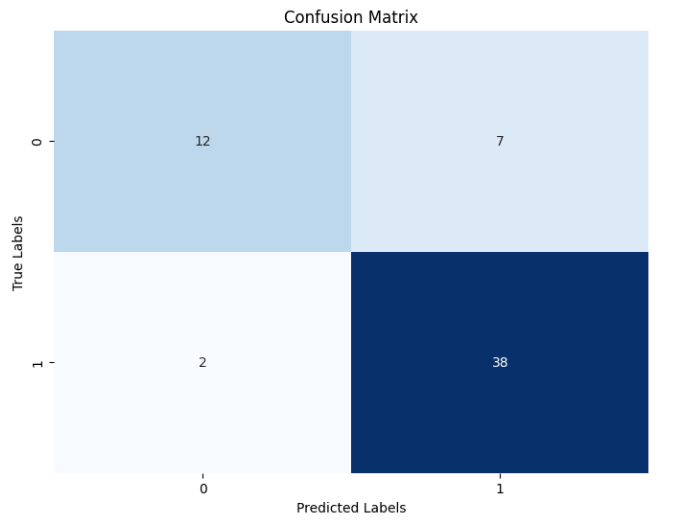
Feature selection was performed for the diagnosis of PD via the phonetic features. There are 22 phonetic features extracted from the speech signals of PD patients and the healthy people. The dataset was rearranged to include less number of columns (features) for an efficient classification. 15 features were selected from FS using correlation based KNN clustering and SOM(self organizing maps) and used in LGBMClassifier and LIME. 10 features were selected from FS using apriori algorithm and used in voting and stacking ensemble. FS provided about 15 % improvement in accuracy and for KMeans and about 12% improvement in accuracy for SOM based on LGBMClassifier . It produced about 16% for both voting ensemble and stacking ensemble.

**Confusion Matrices***:*

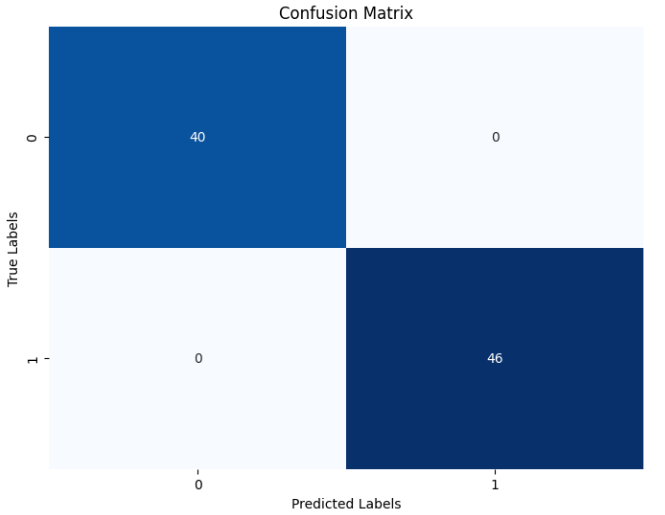
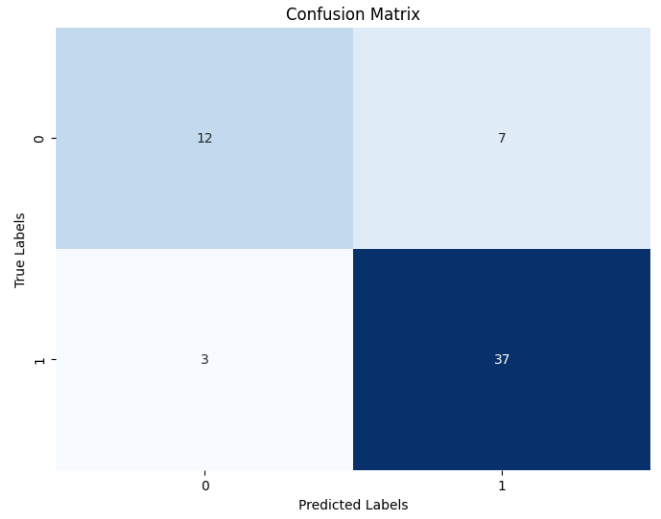
**Before Feature Selection: After Feature Selection:**



LightGBM Classifier Using KMeans



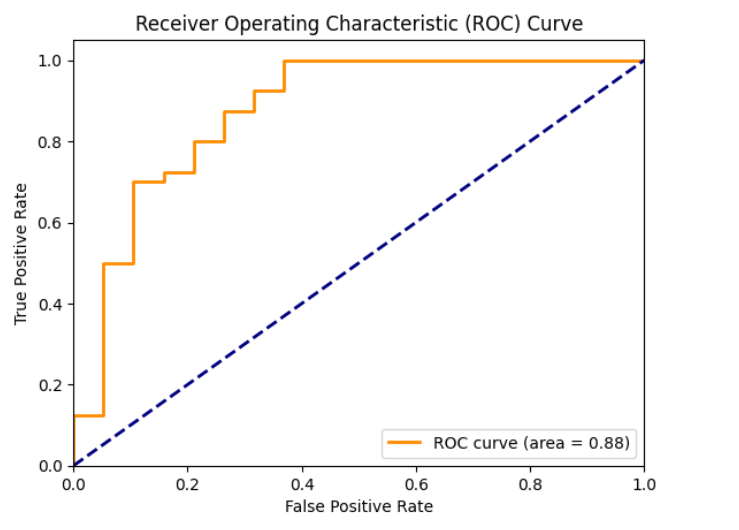
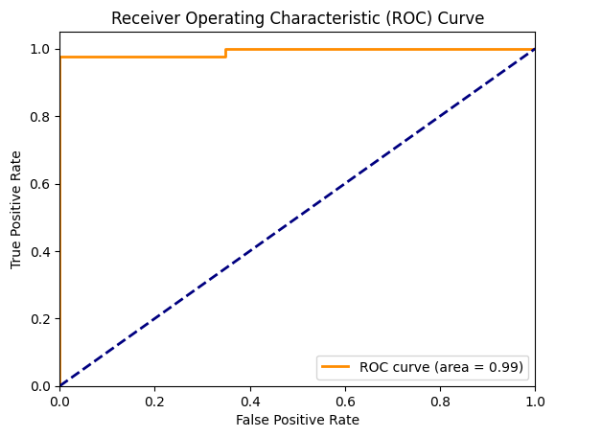
Voting Ensemble Using Apriori



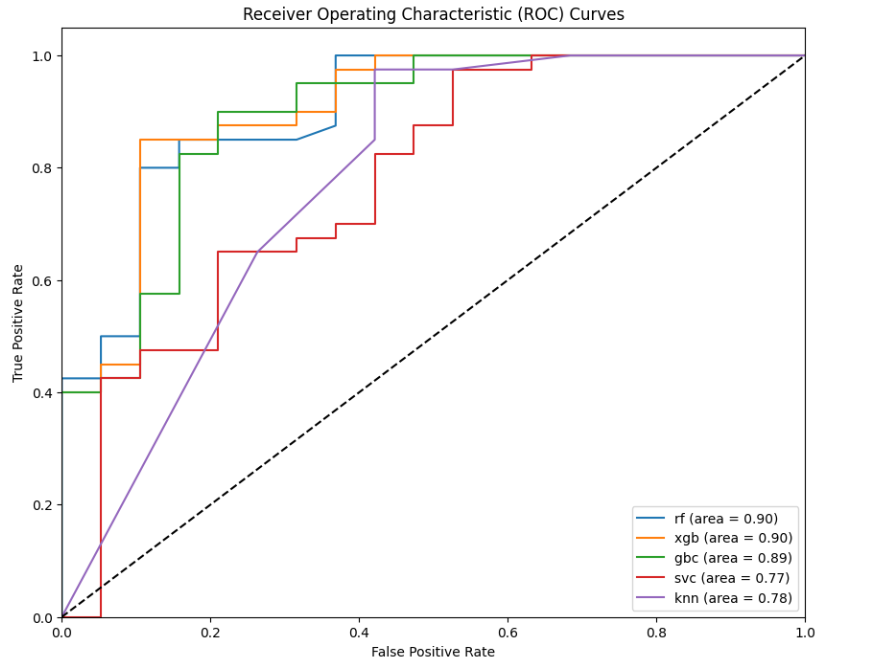
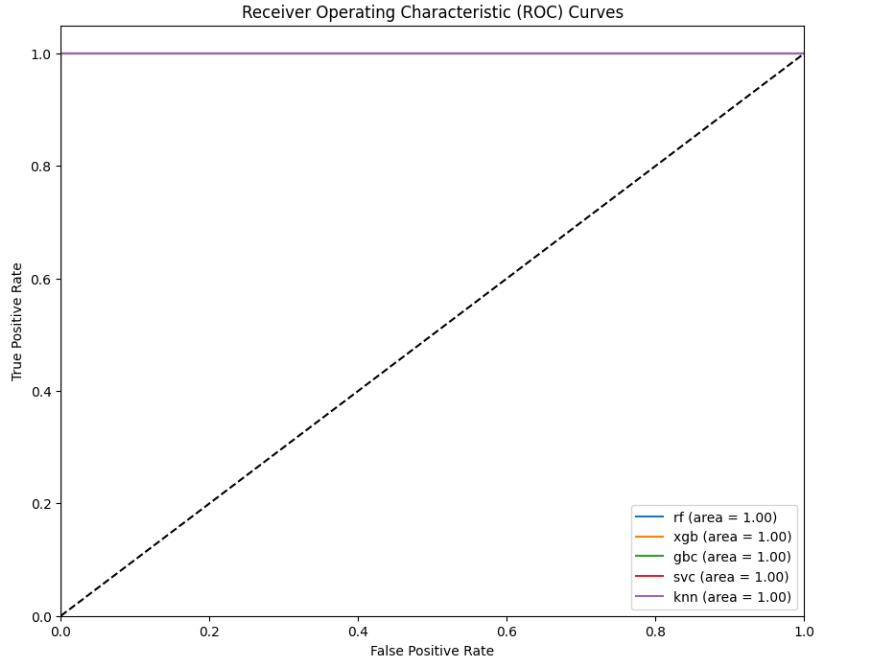
Stacking Ensemble Using Apriori

**ROC Curves:**

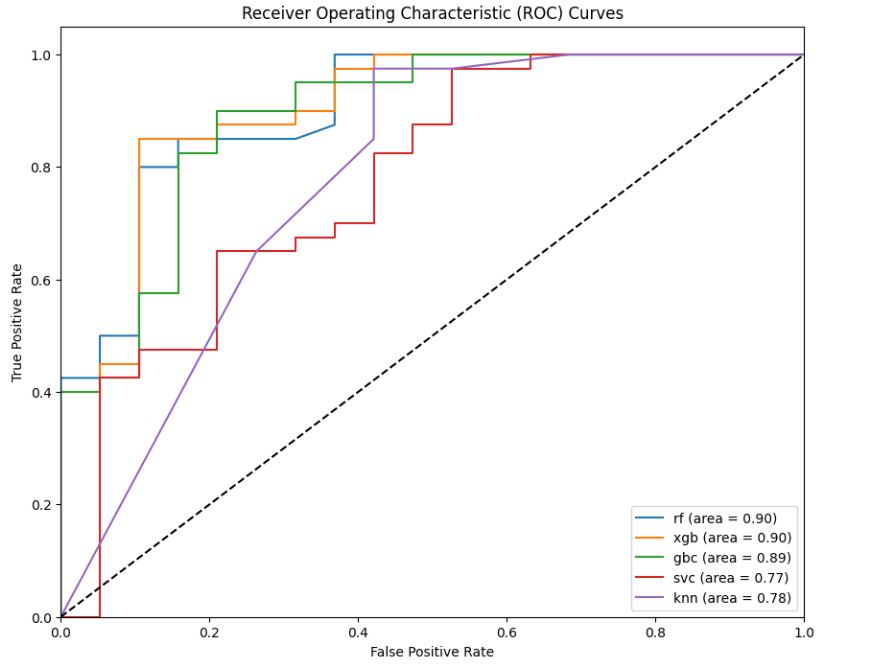
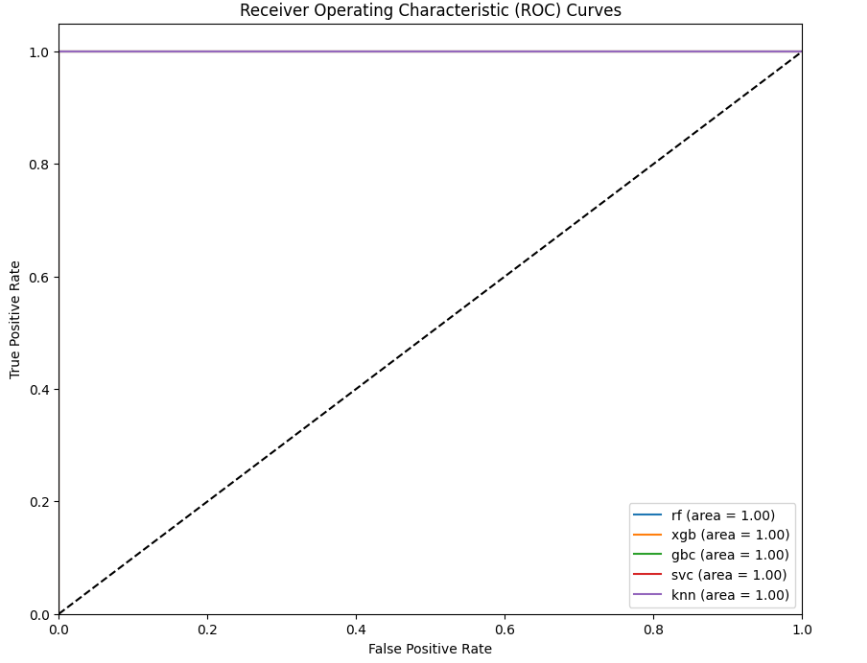
**Before Feature Selection: After Feature Selection:**

LightGBM Classifier Using KMeans

Voting Ensemble Using Apriori

Stacking Ensemble Using Apriori

1. **Conclusion:**

By utilizing various Feature Selection (FS) methods, we identified the most significant features from our dataset. This process led to a substantial reduction in computational cost, processing time, and dimensionality, while also enhancing the accuracy of our models. This refined methodology allows for a more precise and reliable analysis of speech patterns, thereby improving our ability to diagnose Parkinson’s disease with high accuracy at its early stages.

Early diagnosis is crucial as it enables timely intervention and treatment, potentially preventing the progression of the disease and improving patient outcomes and quality of life. Through this enhanced approach, healthcare professionals can leverage advanced machine learning techniques to make more informed decisions and provide better care for individuals at risk of developing Parkinson’s disease.

1. **References:**

[1] Islam, N., Turza, M.S.A., Fahim, S.I. *.* Single and Multi-modal Analysis for Parkinson’s Disease to Detect Its Underlying Factors. *Hum-Cent Intell Syst* **4**, 316–334 (2024). <https://doi.org/10.1007/s44230-024-00069-z>

[2] Yuan, L., Liu, Y. & Feng, HM. Parkinson disease prediction using machine learning-based features from speech signal. *SOCA* **18**, 101–107 (2024). <https://doi.org/10.1007/s11761-023-00372-w>

[3] Vu, T. A A Comparison of Machine Learning Algorithms for Parkinson’s Disease Detection. 2023 (ICCAIS) 122023, pp. 316-321. https://doi.org/10.1109/ICCAIS59597.2023.10382406

[4] Raval, S., Balar, R., & Patel, V. A Comparative Study of Early Detection of Parkinson's Disease using Machine Learning Techniques. (ICOEI) 4, 2020, pp. 509-516. https://doi.org/10.1109/ICOEI48184.2020.9142956

[5] Zehra Karapinar Senturk, Early diagnosis of Parkinson’s disease using machine learning algorithms, Medical Hypotheses 138, 109-603,(2020). <https://doi.org/10.1016/j.mehy.2020.109603>.

[6] Aditi Govindu, Sushila Palwe, Early detection of Parkinson's disease using machine learning, Procedia Computer Science, 2023,Pages 249-261, <https://doi.org/10.1016/j.procs.2023.01.007>.

[7] D. M. Kumar, R. Arthi, A. Rajeev, A. Ranjith, A. Murali and A. K, "Early Detection of Parkinsons Using Machine Learning," 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), 2024, pp. 562-565, <https://doi.org/10.1109/ESIC60604.2024.10481533>.

[8] Saleh, S., Cherradi, B., El Gannour, O. *et al.* Predicting patients with Parkinson's disease using Machine Learning and ensemble voting technique. *Multimed Tools Appl* **83**, 33207–33234 (2024). <https://doi.org/10.1007/s11042-023-16881-x>

[9] Barukab, O.; Ahmad, A.; Khan, T.; Thayyil Kunhumuhammed, M.R. Analysis of Parkinson’s Disease Using an Imbalanced-Speech Dataset by Employing Decision Tree Ensemble Methods. Diagnostics **2022**, 12, 3000. <https://doi.org/10.3390/diagnostics12123000>

[10] Martinez-Eguiluz, M., Arbelaitz, O., Gurrutxaga, I. *et al.* Diagnostic classification of Parkinson’s disease based on non-motor manifestations and machine learning strategies. *Neural Comput & Applic* **35**, 5603–5617 (2023). <https://doi.org/10.1007/s00521-022-07256-8>

[11] Little,Max. (2008). Parkinsons. UCI Machine Learning Repository. <https://doi.org/10.24432/C59C74>.

[12] Lamba, R., Gulati, T., Alharbi, H.F. *et al.* A hybrid system for Parkinson’s disease diagnosis using machine learning techniques. *Int J Speech Technol* **25**, 583–593 (2022). https://doi.org/10.1007/s10772-021-09837-9

[13] [ICIMMI '23: Proceedings of the 5th International Conference on Information Management & Machine Intelligence](https://dl.acm.org/doi/proceedings/10.1145/3647444)November 2023Article No.: 99Pages 1–8<https://doi.org/10.1145/3647444.3647926>

[14] S. Raval, R. Balar and V. Patel, "A Comparative Study of Early Detection of Parkinson's Disease using Machine Learning Techniques," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 509-516, doi: 10.1109/ICOEI48184.2020.9142956.

[15] Bhattacharjee, A. K., & Mukherjee, S. (2017). Parkinson's Disease Clustering and Classification with Multi Layer Perceptron using Best Ranked Voice Feature and Associative Rule Generation. \*International Journal of Computer Science Issues (IJCSI)\*, 14(5), 31-36. <https://doi.org/10.20943/01201705.3136>

[16] Gürüler, H. A novel diagnosis system for Parkinson’s disease using complex-valued artificial neural network with k-means clustering feature weighting method. Neural Comput & Applic 28, 1657–1666 (2017). https://doi.org/10.1007/s00521-015-2142-2

[17] Berus, L.; Klancnik, S.; Brezocnik, M.; Ficko, M. Classifying Parkinson’s Disease Based on Acoustic Measures Using Artificial Neural Networks. *Sensors* 2019, *19*, 16. https://doi.org/10.3390/s19010016

[18] Saleh, S., Cherradi, B., El Gannour, O. et al. Predicting patients with Parkinson's disease using Machine Learning and ensemble voting technique. Multimed Tools Appl 83, 33207–33234 (2024). <https://doi.org/10.1007/s11042-023-16881-x>

[19] Yifeng Yang, Long Wei, Ying Hu, Yan Wu, Liangyun Hu, Shengdong Nie, Classification of Parkinson's disease based on multi-modal features and stacking ensemble learning, Journal of Neuroscience Methods, 2021,109-019, <https://doi.org/10.1016/j.jneumeth.2020.109019>.

[20] Nguyen, H.V.; Byeon, H. Prediction of Parkinson’s Disease Depression Using LIME-Based Stacking Ensemble Model. *Mathematics* 2023, *11*, 708. <https://doi.org/10.3390/math11030708>

[21] Dhar, J. An adaptive intelligent diagnostic system to predict early stage of parkinson's disease using two-stage dimension reduction with genetically optimized lightgbm algorithm. Neural Comput & Applic 34, 4567–4593 (2022). https://doi.org/10.1007/s00521-021-06612-4