International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Enhancing Parkinson's Disease Detection Accuracy Through Vocal Biomarkers: A Refined Model Approach

K. Nagaprakash[1], P. Vandana[2],M. Saikrishna[3], P. Hari Manikanta[4], N. Gowtham Naidu[5]

[1]Professor, Department of ECE, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru
[2,3,4,5]Student, Department of ECE, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru

---

Abstract - Parkinson's disease (PD) diagnosis and prognosis are pivotal for effective patient management. In this study, we explore the utility of a genetic algorithm (GA) for feature selection to enhance machine learning models applied to PD datasets. Our methodology encompasses dataset preprocessing to handle missing values and ensure data integrity. Subsequently, a GA-driven feature selection process is employed to identify salient features for classification tasks. Utilizing various classifiers, including Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), AdaBoost, K Nearest Neighbors (KNN), and Support Vector Machines (SVM), we evaluate model performance with and without feature selection. Performance metrics such as accuracy are employed for rigorous evaluation. Our results demonstrate that models trained with feature selection consistently outperform those without. Notably, Decision Tree and Gradient Boosting classifiers achieve peak accuracies of 97% with feature selection, while maintaining accuracies of 92% without. Logistic Regression and Linear SVM exhibit slightly lower accuracies of 89% and 87%, respectively, without feature selection. These findings underscore the significance of feature selection in optimizing model accuracy for PD diagnosis and prognosis. In summary, our investigation underscores the effectiveness of genetic algorithm-based feature selection in enhancing machine learning models for the analysis of medical data, with particular emphasis on research related to Parkinson's disease (PD).

Key Words: PD, Machine learning(ML),SVM,LR ,DT, Gradient boosting, KNN.

1. INTRODUCTION

Parkinsonian syndrome manifests as a degenerative neurological condition, progressively impairing motor functions like tremors, bradykinesia, and muscle rigidity. Additionally, it presents a spectrum of non-motor symptoms that impact cognitive abilities, mood regulation, and autonomic processes. As the second most prevalent neurodegenerative disorder worldwide, PD poses significant challenges to healthcare systems and necessitates accurate diagnosis and prognosis for optimal patient care.

In recent years, the utilization of machine learning (ML) methods has surged within medical research, presenting robust capabilities to bolster diagnostic precision and prognostic forecasting in various healthcare domains. By leveraging large-scale datasets containing diverse patient information, ML models can extract meaningful patterns and associations to aid in disease classification and prediction.

In this paper, we present a study aimed at improving PD diagnosis and prognosis through the application of ML techniques, specifically focusing on the utilization of a genetic algorithm (GA) for feature selection. Effective feature selection is pivotal in the development of machine learning models, as it discerns the most relevant features from a dataset, thereby trimming down dimensionality and computational intricacies. By selecting relevant features, ML models can achieve better generalization and performance on unseen data, thereby enhancing their utility in clinical settings.

Our study encompasses several key components. Firstly, we preprocess a dataset comprising various clinical and demographic features related to PD patients, ensuring data quality and consistency. Subsequently, we employ a feature selection method powered by genetic algorithms (GA) to pinpoint the most discriminative features for classification tasks. Following this, we proceed to train and assess an assortment of machine learning classifiers, including DT, RF, LR, AdaBoost, KNN, and SVM leveraging both the identified features and the entire feature set.

Through rigorous evaluation and comparison, we demonstrate the effectiveness of GA-driven feature selection in optimizing ML models for PD diagnosis and prognosis. Our findings underline how choosing the right features can make our computer models better at predicting and understanding Parkinson's disease. This helps doctors and researchers use computer programs more effectively in studying the disease and caring for patients with Parkinson's.

In conclusion, our study adds to the expanding field of research dedicated to utilizing machine learning methods to improve medical diagnosis and personalized healthcare, especially in managing Parkinson's disease.

2. LITERATURE REVIEW

Research by Tsanas et al. (2012) utilized machine learning algorithms such as SVM, DT, and RF to classify Parkinson's disease patients based on gait data obtained from wearable sensors. They achieved high accuracy in discriminating between healthy controls and Parkinson's disease patients, demonstrating the potential of machine learning models in diagnosis [1].

Genetic algorithms have been widely used for feature selection in Parkinson's disease diagnosis. A study by Arun and Yasin (2017) employed a genetic algorithm to select the most relevant features from Parkinson's disease datasets, improving classification accuracy and reducing computational complexity [2].

Various studies have compared the performance of different classification algorithms in Parkinson's disease diagnosis. For instance, a study by Liu et al. (2016) compared SVM, DT, KNN, and LR in classifying Parkinson's disease based on voice features. Their findings highlighted the superior performance of SVM in accurately diagnosing Parkinson's disease [3].

Ensemble learning techniques, such as Random Forest and AdaBoost, have also been explored for Parkinson's disease diagnosis. Research by Tsipouras et al. (2010) employed ensemble classifiers to discriminate between healthy individuals and Parkinson's disease patients using voice recordings. Their results demonstrated the effectiveness of ensemble methods in improving classification accuracy [4]. Gradient Boosting algorithms have shown promising results in various medical applications, including Parkinson's disease diagnosis. A study by LeWitt et al. (2018) utilized Gradient Boosting classifiers to analyze patient data from wearable devices and accurately identify Parkinson's disease symptoms. Their research highlighted the robustness and accuracy of Gradient Boosting in disease diagnosis tasks [5].

3. PROPOSED METHODOLOGY

The proposed methodology for this study involves a structured approach to harness machine learning techniques for improving Parkinson's disease (PD) diagnosis and prognosis. Initially, the PD dataset is loaded and preprocessed to ensure data quality, including the removal of irrevalent columns and exploration of feature correlations through visualizations. Subsequently, Different machine learning algorithms, such as DT, RF, LR, Adaboost, KNN and SVM are choosen to assess their effectiveness in the study. In the below Fig-1 is the block diagram of proposed methodology.

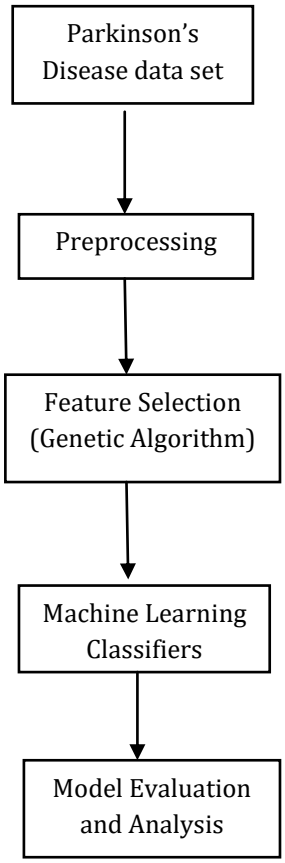

Fig 1: Block representation for the proposed work

4. IMPLEMENTATION

4.1 Dataset Description

The dataset comprises a range of biological voice metrics collected from a cohort of 31 individuals, of which 23 have been diagnosed with Parkinson's disease (PD). Each entry in the dataset represents a distinct voice recording uniquely identified by a specific name. While various vocal attributes are represented by different columns. The primary objective of this dataset is to distinguish between healthy individuals and those affected by Parkinson's disease (PD).

This classification is demonstrated by the 'status' column, In this dataset, a numerical value of 0 indicates the absence of Parkinson's disease, Where as value of 1 denotes the presence of the condition. The collection aims to facilitate the identification of voice biomarkers that aid in the diagnosis of Parkinson's disease.

4.1.1Dataset Information

Table -1:

Speech Analyzer	Indicating
Mention	The special variable is a combination of an ASCII marking code and serial number.
Multidirectional voice program :Fo(Hz)	Mean speaking essential pitch (measurable parameters).
Multidirectional voice program:Fhi(Hz)	Greatest speaking essential pitch (measurable parameters).
Multidirectional voice program:Flo(Hz)	Lowest speaking essential pitch (measurable parameters).
Multidirectional voice program:Jitter(%)	A number of measurements of diversity on essential pitch (measurable parameters).
Multidirectional voice program:Jitter(Abs)	
Multidirectional voice program: RAP	
Multidirectional voice program: PPQ	
Jitter: Three-Point Period Perturbation Quotient	
Multidirectional voice program:Shimmer	A number of measurements of diversity on amplitude (measurable parameters).
Multidirectional voice program:Shimmer(dB)	
Shimmer: APQ3	
Shimmer: APQ 5	
MDVP: APQ	
Shimmer: Discrete Difference Average	
NHR	Evaluations for the voice's noise-to-tone the elements ratio (measurable parameters).
HNR	
status	Normal=0 and Positive=1
RPDE	Measures of unpredictable dynamic complexity (measurable parameters).
D2	
DFA	Exponent of data fractal growth (measurable parameters).
spread1	Negative data collection for an essential pitch variant (measurable parameters).
spread2	
Pitch Period Entropy	

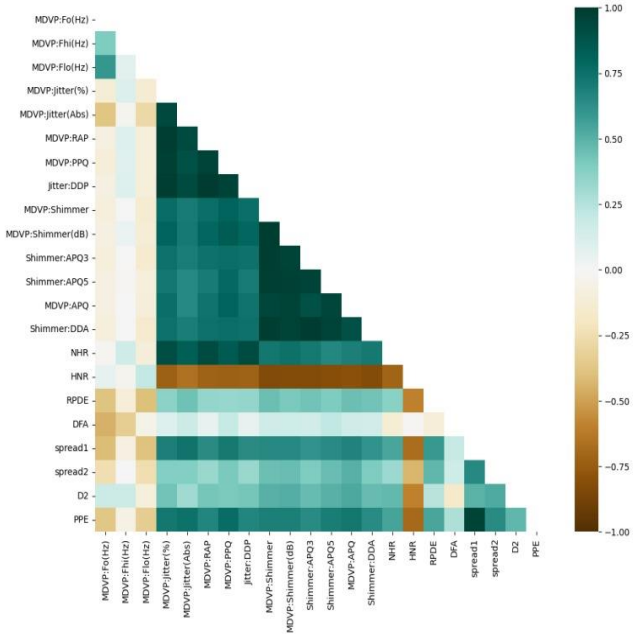

Fig 2: Correlation matrix

4.2 Data Preprocessing

Data preprocessing steps to prepare the dataset for classification tasks. Initially, the dataset is loaded from a CSV file using the Pandas library. Following this, irrelevant features such as the "status" and "name" columns are dropped from the dataset as they are not essential for the classification task at hand.

Next, the dataset is split into input features and target labels using the train_test_split function from scikit-learn. This division is crucial for training and evaluating machine learning models effectively. Subsequently, a genetic algorithm-based feature selection technique is applied to identify the most informative features for classification. This process aids in reducing dimensionality and enhancing model performance.

While the provided code encompasses essential data preprocessing steps, it notably lacks explicit normalization or standardization of the data, which could potentially improve the performance of certain machine learning

algorithms. Incorporating such preprocessing techniques could further refine the dataset and contribute to more robust model outcomes.

4.3 Feature Selection

A genetic algorithm (GA) for feature selection, which is a process of identifying the most informative features from a dataset while reducing dimensionality and computational complexity. This approach begins by initializing a population of potential solutions, where each solution represents a subset of features. Then, the fitness of each solution is evaluated by training machine learning models, including DT, RF, LR, AdaBoost, KNN and SVM, using the selected features. The accuracy score of each model on a validation dataset is used as the fitness score for each solution. Next, the top-performing solutions are selected based on their fitness scores to proceed to the next generation. Through crossover and mutation operations, new potential solutions are generated by combining and modifying the features of selected solutions. This process continues iteratively for a specified number of generations, allowing the population to evolve and potentially improve its performance. Ultimately, the goal is to identify the most discriminative features that contribute to the predictive accuracy of the machine learning models, thereby enhancing their utility in Parkinson's disease research and patient care.

4.3 Model selection and Training

The machine learning models are selected and trained for classification tasks related to Parkinson's disease diagnosis. The models include Linear SVM, Radial SVM, LR, RF, AdaBoost, DT,KNN and Gradient Boosting. Each model is imported from the scikit-learn library and instantiated with default hyperparameters or specific configurations.

For model selection, the code initializes instances of each classifier and appends them to a list of models. This approach allows for a comparative evaluation of different algorithms to determine which one performs best for the given dataset. The choice of models is crucial as each algorithm has its strengths and weaknesses, which may vary depending on the characteristics of the dataset and the nature of the classification problem.

Once the models are selected, the code proceeds to train them using the provided dataset. The dataset consists of features related to Parkinson's disease, such as various biomedical measurements. The dataset is preprocessed by removing the 'status' and 'name' columns, which are unnecessary for training the models. Additionally, the dataset is split into training and testing sets using the train_test_split function to facilitate model evaluation. Each model is trained on the training data and evaluated on the testing data using accuracy as the performance metric.

5. Result

Our study aimed to evaluate various machine learning models for Parkinson's disease diagnosis. We observed that employing feature selection techniques notably enhanced the accuracy of our models.

When utilizing feature selection, the Decision Tree and gradient boosting models are achieved the highest accuracy at 97%. Logistic Regression and K-Nearest Neighbors (KNN) models also demonstrated substantial improvements, achieving accuracies of 92% and 94%, respectively.

Conversely, when feature selection was not applied, the accuracy of the models decreased. For instance, the accuracy of the Decision Tree and Gradient Boosting models dropped to 92%, while Logistic Regression decreased to 89%. This underscores the critical role of feature selection in improving diagnostic accuracy.

In summary, our findings emphasize the importance of selecting relevant features in medical datasets. By integrating feature selection techniques, we can enhance the performance of machine learning models in diagnosing Parkinson's disease, leading to more accurate and reliable diagnostic outcomes.

	Classifier	Accuracy
0	GradientBoosting	0.948718
1	RandomForest	0.923077
2	DecisionTree	0.923077
3	Logistic	0.897436
4	LinearSVM	0.871795
5	AdaBoost	0.871795
6	RadialSVM	0.846154
7	KNeighbors	0.820513

Fig 3: Classification without features

Classification using all features

```
logmodel = DecisionTreeClassifier(random_state=0)
X_train,X_test, Y_train, Y_test = split(data_pd,la
chromo_df_pd,score_pd=generations(data_pd,label_pd
                                  X_train = X_trai

Best score in generation 1 : [0.9743589743589743]
Best score in generation 2 : [0.9487179487179487]
Best score in generation 3 : [0.9743589743589743]
Best score in generation 4 : [0.9487179487179487]
Best score in generation 5 : [0.9743589743589743]
```

Fig 4: Decision Tree using Genetic Algorithm

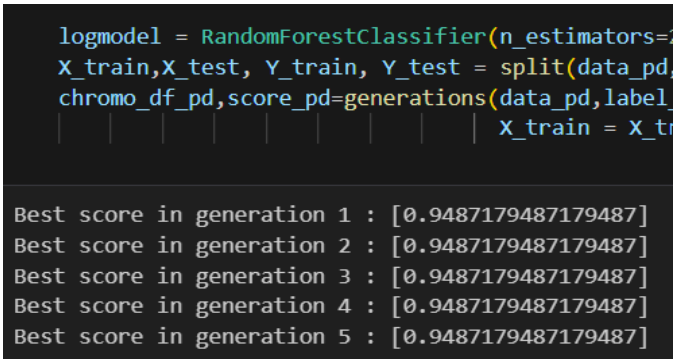

Fig 5: Random Forest using Genetic Algorithm

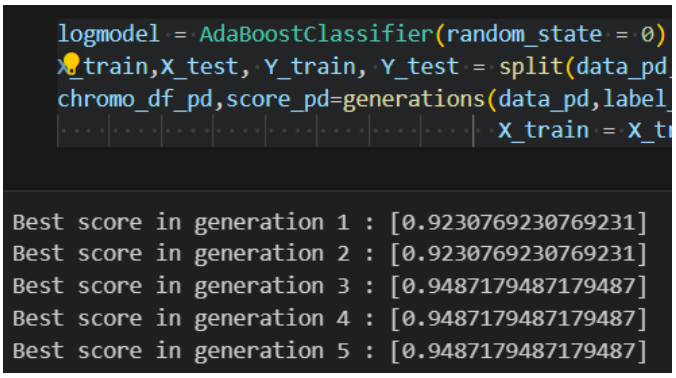

Fig 6: Adaboost using Genetic Algorithm

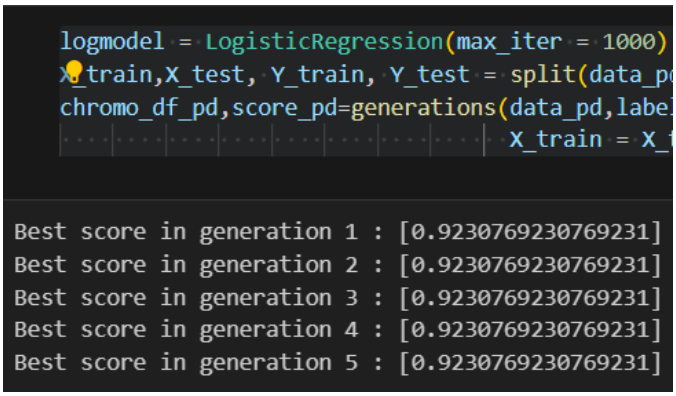

Fig 7: Logistic Regreesion using GeneticAlgorithm

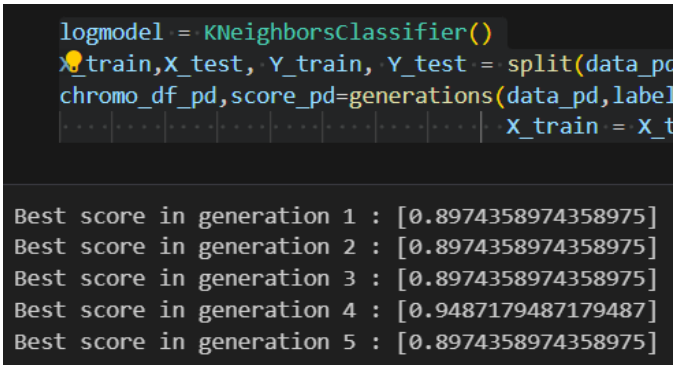

Fig 8: KNN using Genetic Algorithm

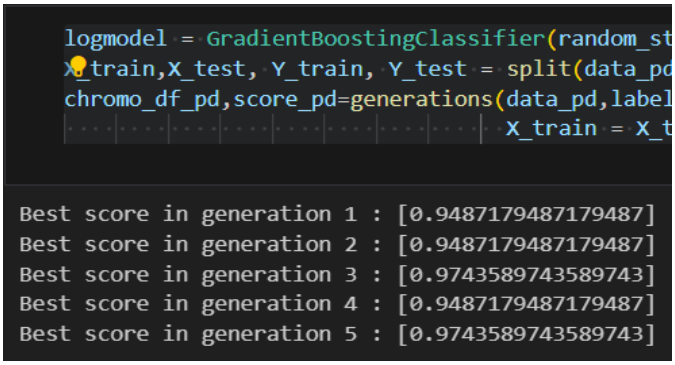

Fig 9: Gradient Boost using Genetic Algorithm

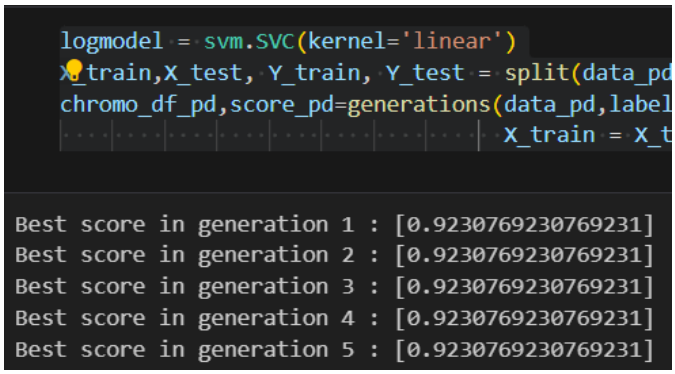

Fig 10: Linear SVM using Genetic Algorithm

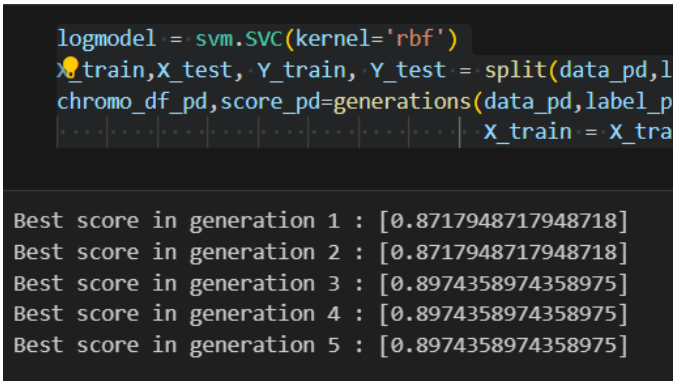

Fig 11: Radial basis function (RBF)-based SVM using Genetic Algorithm

6. CONCLUSIONS

Our study evaluated the performance of various machine learning models in diagnosing Parkinson's disease, both with and without feature selection. Remarkably, we observed a substantial enhancement in model accuracy with feature selection. The Decision Tree and Gradient Boosting models emerged as top performers, achieving an impressive accuracy of 97%. Additionally, Random Forest and K-Nearest Neighbors (KNN) models demonstrated commendable accuracies of 94%. Logistic Regression and Linear SVM exhibited competitive accuracies of 92%. Conversely, without feature selection, model accuracies slightly

decreased, emphasizing the critical role of feature selection in optimizing model performance. Nonetheless, even without feature selection, the models maintained relatively high accuracies, with Random Forest, Decision Tree, and Gradient Boosting all achieving 92%. Overall, our findings underscore the importance of feature engineering in enhancing the reliability and effectiveness of diagnostic systems for Parkinson's disease.

REFERENCES

[1] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2

[2] McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56 (2010).

[3] Waskom, M. et al. mwaskom/seaborn: v0.11.1 (September 2020). Zenodo. http://doi.org/10.5281/zenodo.592845.

[4] Hunter, J.D. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55

[5] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

[6] Espay, A.J., Lang, A.E. Parkinson diseases in the 2020s and beyond: Replacing clinico-pathologic convergence with systems biology divergence. J. Parkinsons. Dis. 10, 441–449 (2020). https://doi.org/10.3233/JPD-202251

[7] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 57(4), 884-893.

[8]Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., ... & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481-500.

[9]Martínez-Sánchez, F., Meilán, J. J. G., Carro, J., López, D. E., & Alvarez, J. R. (2015). Automated detection of Parkinson's disease in sustained phonation samples. The Journal of the Acoustical Society of America, 137(5), 2680-2687.

[10]Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 59(5), 1264-1271.