Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/40/20241202

Comparative analysis of machine learning in diagnosing Parkinson’s: Utilizing vocal characteristics

Yu-Rou Yu

Hillsborough High School, Hillsborough, 08844, United States

yuy@htps.us

Abstract. Parkinson’s disease is a neurodegenerative disorder that affects movement. Diagnosing Parkinson’s disease has traditionally involved clinical assessments by neurologists, and this practice still persists today to a significant extent. However, clinical assessments can be prone to subjectivity. In this study, a comprehensive predictive modeling approach was undertaken, employing nine distinct machine learning algorithms and six different model evaluation metrics to identify the best performing algorithms. The findings reveal that, using only 12 vocal characteristics, KNeighborsClassfier (KNC), MLPClassifier (MLP), and XGBClassifier (XGBC) achieved the highest score of 0.87. This score is generally considered very good, indicating that the model is robust and possesses strong predictive power. This study marks a crucial initial step in leveraging machine learning techniques for more effective and potentially more accurate diagnosis of Parkinson’s disease based on patients’ vocal characteristics.

Keywords: Parkinson’s disease, machine learning, vocal characteristics.

1. Introduction

Parkinson’s disease is a neurodegenerative disorder that affects movement. It develops gradually, often starting with subtle tremors, stiffness, and difficulty with coordination. As it progresses, individuals may experience slowed movement and impaired balance. Diagnosing Parkinson’s disease has traditionally involved clinical assessments by neurologists from evaluating medical history, and observing symptoms to the more advanced diagnostic approaches including neuroimaging techniques like MRI or DaTscan, which can help visualize changes in the brain [1]. This practice still persists today to a significant extent. However, there isn’t a definitive test for Parkinson’s disease and these diagnostic approaches could be influenced by subjectivity.

Studies have indicated that vocal characteristics play a pivotal role in the diagnostic process for Parkinson’s disease. Alterations in speech patterns, including changes in pitch, frequency, amplitude variation, and articulation, often manifest as early symptoms. These changes, termed dysphonia, reflect the underlying neurodegenerative processes affecting the vocal cords, muscles, and control mechanisms in the brain [2].

In recent years, machine learning algorithms have emerged as powerful tools in various medical applications. These algorithms analyze vast amounts of patient data, allowing for predictive modeling, disease diagnosis, and outcome forecasting. Their ability to discern intricate patterns within complex datasets enables more accurate prognoses, aiding in making informed decisions. In this study, 9 distinct machine learning algorithms were employed to analyze data from 195 patients, considering their diverse vocal characteristics. Then, six different model evaluation metrics were utilized to assess the performance of the models in accurately identifying patients with Parkinson's disease. The aim is to potentially enable earlier detection and intervention for improved patient outcomes.

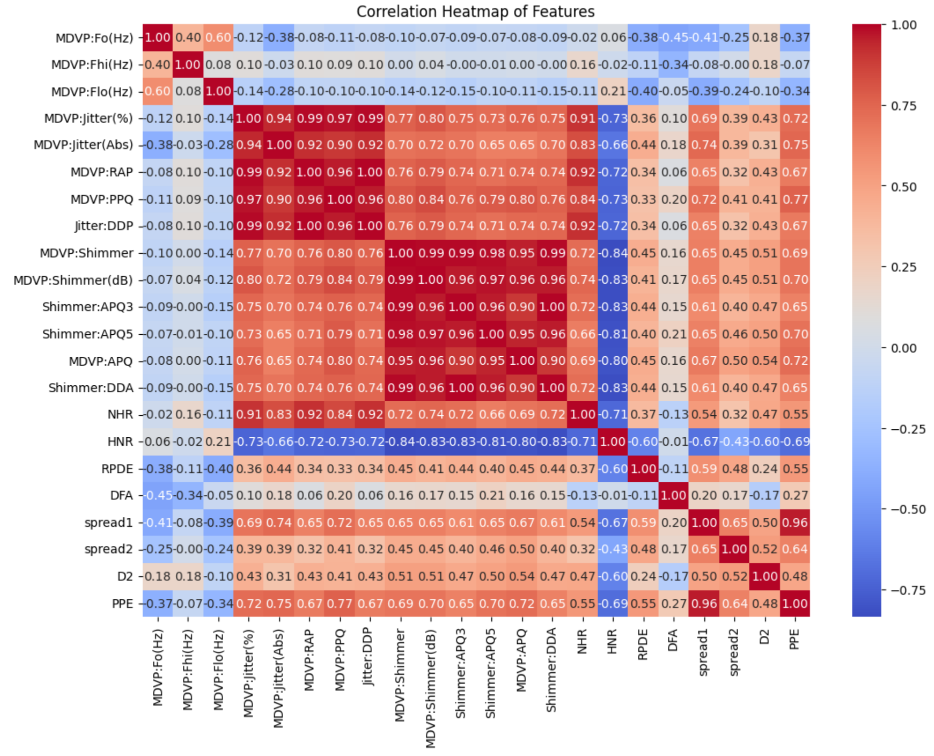
2. Methods

2.1. Data Acquisition

Provided by the University of California, Irvine and available through the university website and Kaggle, the dataset encompasses 22 vocal characteristic attributes for 195 patients, where 147 individuals have been diagnosed with Parkinson's disease [3]. These attributes collectively represent a diverse range of vocal traits, covering frequency, amplitude, noise components, jittering, shimmering, nonlinear dynamical complexity, and scaling properties, providing a holistic view of patients' vocal characteristics. The dataset also underwent scrutiny to identify potential missing values, extreme outliers, and any potential irregularities. Subsequently, adjustments were made, including the imputation of missing values and the application of aggregation methods.

2.2. Correlation Matrix

The correlation matrix serves as an initial exploration and analysis step in understanding the structure of the data. Figure 1 presents a correlation heatmap displaying the correlation strength and direction among the 22 vocal characteristic variables. When a correlation is highly positive, approaching 1, it indicates a strong positive linkage, signifying that an increase in one variable is accompanied by a concurrent increase in the other. In contrast, a substantial negative correlation, nearing -1, suggests a strong negative connection, indicating that an increase in one variable is linked to a decrease in the other. High correlations between features suggest redundancy or multicollinearity. Redundant features offer similar information, potentially affecting model performance and interpretation. Therefore, the aim is to identify highly correlated features and remove redundant features, reducing dimensionality without losing much information.



	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)	Shimmer:APQ3	Shimmer:APQ5	MDVP:APQ	Shimmer:DDA	NHR	HNR	RPDE	DFA	spread1	spread2	D2	PPE
MDVP:Fo(Hz)	1.00	0.40	0.60	-0.12	-0.38	-0.08	-0.11	-0.08	-0.10	-0.07	-0.09	-0.07	-0.08	-0.09	-0.02	0.06	-0.38	-0.45	-0.41	-0.25	0.18	-0.37
MDVP:Fhi(Hz)	0.40	1.00	0.08	0.10	-0.03	0.10	0.09	0.10	0.00	0.04	-0.00	-0.01	0.00	-0.00	0.16	-0.02	-0.11	-0.34	-0.08	-0.00	0.18	-0.07
MDVP:Flo(Hz)	0.60	0.08	1.00	-0.14	-0.28	-0.10	-0.10	-0.10	-0.14	-0.12	-0.15	-0.10	-0.11	-0.15	-0.11	0.21	-0.40	-0.05	-0.39	-0.24	-0.10	-0.34
MDVP:Jitter(%)	-0.12	0.10	-0.14	1.00	0.94	0.99	0.97	0.99	0.77	0.80	0.75	0.73	0.76	0.75	0.91	-0.73	0.36	0.10	0.69	0.39	0.43	0.72
MDVP:Jitter(Abs)	-0.38	-0.03	-0.28	0.94	1.00	0.92	0.90	0.92	0.70	0.72	0.70	0.65	0.65	0.70	0.83	-0.66	0.44	0.18	0.74	0.39	0.31	0.75
MDVP:RAP	-0.08	0.10	-0.10	0.99	0.92	1.00	0.96	1.00	0.76	0.79	0.74	0.71	0.74	0.74	0.92	-0.72	0.34	0.06	0.65	0.32	0.43	0.67
MDVP:PPQ	-0.11	0.09	-0.10	0.97	0.90	0.96	1.00	0.96	0.80	0.84	0.76	0.79	0.80	0.76	0.84	-0.73	0.33	0.20	0.72	0.41	0.41	0.77
Jitter:DDP	-0.08	0.10	-0.10	0.99	0.92	1.00	0.96	1.00	0.76	0.79	0.74	0.71	0.74	0.74	0.92	-0.72	0.34	0.06	0.65	0.32	0.43	0.67
MDVP:Shimmer	-0.10	0.00	-0.14	0.77	0.70	0.76	0.80	0.76	1.00	0.99	0.99	0.98	0.95	0.99	0.72	-0.84	0.45	0.16	0.65	0.45	0.51	0.69
MDVP:Shimmer(dB)	-0.07	0.04	-0.12	0.80	0.72	0.79	0.84	0.79	0.99	1.00	0.96	0.97	0.96	0.96	0.74	-0.83	0.41	0.17	0.65	0.45	0.51	0.70
Shimmer:APQ3	-0.09	-0.00	-0.15	0.75	0.70	0.74	0.76	0.74	0.99	0.96	1.00	0.96	0.90	1.00	0.72	-0.83	0.44	0.15	0.61	0.40	0.47	0.65
Shimmer:APQ5	-0.07	-0.01	-0.10	0.73	0.65	0.71	0.79	0.71	0.98	0.97	0.96	1.00	0.95	0.96	0.66	-0.81	0.40	0.21	0.65	0.46	0.50	0.70
MDVP:APQ	-0.08	0.00	-0.11	0.76	0.65	0.74	0.80	0.74	0.95	0.96	0.90	0.95	1.00	0.90	0.69	-0.80	0.45	0.16	0.67	0.50	0.54	0.72
Shimmer:DDA	-0.09	-0.00	-0.15	0.75	0.70	0.74	0.76	0.74	0.99	0.96	1.00	0.96	0.90	1.00	0.72	-0.83	0.44	0.15	0.61	0.40	0.47	0.65
NHR	-0.02	0.16	-0.11	0.91	0.83	0.92	0.84	0.92	0.72	0.74	0.72	0.66	0.69	0.72	1.00	-0.71	0.37	-0.13	0.54	0.32	0.47	0.55
HNR	0.06	-0.02	0.21	-0.73	-0.66	-0.72	-0.73	-0.72	-0.84	-0.83	-0.83	-0.81	-0.80	-0.83	-0.71	1.00	-0.60	-0.01	-0.67	-0.43	-0.60	-0.69
RPDE	-0.38	-0.11	-0.40	0.36	0.44	0.34	0.33	0.34	0.45	0.41	0.44	0.40	0.45	0.44	0.37	-0.60	1.00	-0.11	0.59	0.48	0.24	0.55
DFA	-0.45	-0.34	-0.05	0.10	0.18	0.06	0.20	0.06	0.16	0.17	0.15	0.21	0.16	0.15	-0.13	-0.01	-0.11	1.00	0.20	0.17	-0.17	0.27
spread1	-0.41	-0.08	-0.39	0.69	0.74	0.65	0.72	0.65	0.65	0.65	0.61	0.65	0.67	0.61	0.54	-0.67	0.59	0.20	1.00	0.65	0.50	0.96
spread2	-0.25	-0.00	-0.24	0.39	0.39	0.32	0.41	0.32	0.45	0.45	0.40	0.46	0.50	0.40	0.32	-0.43	0.48	0.17	0.65	1.00	0.52	0.64
D2	0.18	0.18	-0.10	0.43	0.31	0.43	0.41	0.43	0.51	0.51	0.47	0.50	0.54	0.47	0.47	-0.60	0.24	-0.17	0.50	0.52	1.00	0.48
PPE	-0.37	-0.07	-0.34	0.72	0.75	0.67	0.77	0.67	0.69	0.70	0.65	0.70	0.72	0.65	0.55	-0.69	0.55	0.27	0.96	0.64	0.48	1.00

Figure 1. Correlation Heatmap with 22 Features.

In the heatmap, MDVP: Jitter(%), MDVP: Jitter(Abs), MDVP: RAP, MDVP: PPQ, and Jitter: DDP exhibit very strong correlations with each other (correlation coefficients close to 1), indicating redundant information. In addition, these variables represent fundamentally similar vocal characteristics, prompting the decision to retain one and drop the others. The utilization of feature importance scores from Decision Trees and Random Forests aids in this selection process. Among these correlated features, Jitter: DDP has the highest scores, leading to its retention in the model while discarding the other four. Similarly, MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, and Shimmer: DDA are also very highly correlated, with Shimmer: APQ5 having the highest feature importance scores. Consequently, Shimmer: APQ5 is retained, and the remaining five are eliminated. Finally, NHR and HNR exhibit a strong negative correlation, and NHR has higher feature importance scores. Thus, NHR is retained, and HNR is removed from the model. Figure 2 displays the updated correlation heatmap after dropping these 10 features, resulting in a model with only 12 features.

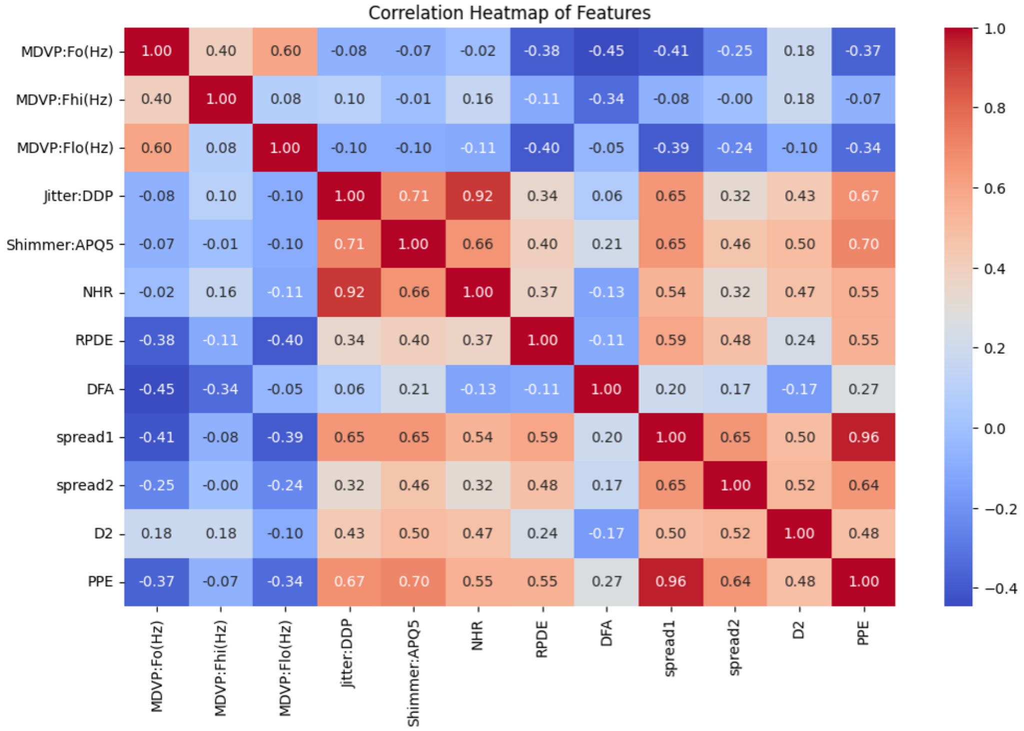


Figure 2. Correlation Heatmap with 12 Features

2.3. *Dataset Partitioning: Training and Testing Datasets*

Splitting the dataset into training and testing sets is essential to evaluate the model's performance and its ability to generalize to unseen data. The training set is used to teach the model the patterns and relationships within the data, allowing it to learn to make predictions. Meanwhile, the testing set acts as an unseen dataset used to assess how well the model generalizes to new, previously unseen data. This study employs a typical division, allocating 70% of the data for training and reserving the remaining 30% for testing. This split aims to ensure that the model has enough data to learn from while also having sufficient unseen data to evaluate its performance. Additionally, the validation dataset is also utilized during the model's training phase to fine-tune hyperparameters and validate different models. It helps prevent overfitting by providing an additional checkpoint to evaluate the model's performance during training.

2.4. *Data Standardization*

Standardization is another crucial preprocessing step in machine learning. It transforms the numerical features to have a mean of zero and a standard deviation of one. This normalization ensures that all features contribute equally to the model training process, preventing any particular feature from dominating simply due to its larger scale or magnitude. Without standardization, features with larger scales might disproportionately influence the model, overshadowing the impact of smaller-scaled features. In addition, by standardizing features, the model becomes more robust and less sensitive to the scale of input variables. This step helps machine learning algorithms perform optimally, especially those that rely on optimization techniques like gradient descent. For instance, K-Nearest Neighbors benefits significantly from standardized features, as it speeds up convergence and ensures fair consideration of all features during the training process.

2.5. Evaluating Machine Learning Models

Multiple evaluation metrics are used in this study to assess and compare the performance of different models. Table 1 lists those evaluation metrics alongside their definitions and suitability for different scenarios [4]. False positives (FP) in Parkinson's prediction could lead to unnecessary treatments or stress, while false negatives (FN) might result in delayed or missed diagnoses. Each of these metrics plays a role in different aspects of evaluating a predictive model for Parkinson's disease, addressing specific concerns such as correctly identifying positive cases (Parkinson's) or negative cases (non-Parkinson's), balancing errors, and considering the overall predictive quality. Therefore, the decision is to assign equal weight to each metric, but slightly downgrading the importance of Accuracy due to the dataset's class imbalance, where Parkinson's disease cases significantly outnumber non-Parkinson's cases.

Table 1. Evaluation Metrics

Metric	Definition/Formula	Suitability
Precision	Precision quantifies the ratio of correctly predicted positive observations to the total predicted positive observations. Precision = TP / (TP + FP)	Useful when the focus is on minimizing false positives (FP), common in scenarios like disease diagnosis. Helps in avoiding unnecessary treatments by minimizing false positives
Specificity	Specificity measures the ratio of correctly predicted negatives to all actual negatives. Specificity = TN / (TN + FP)	Crucial when avoiding false positives (FP) is a priority, used in scenarios like medical screenings.
Recall (Sensitivity)	Recall represents the ratio of correctly predicted positive observations to all actual positives. Recall = TP / (TP + FN)	Crucial when missing actual positive instances (FN) is critical, like medical diagnoses. Helps in avoiding delayed or missed diagnoses.
F1 Score	F1 score is the harmonic mean of precision and recall, offering a balance between the two metrics. F1 Score = 2 * (Precision * Recall) / (Precision + Recall)	Effective when the focus is on balancing precision and recall, suitable in various scenarios.
Matthews Correlation Coefficient (MCC)	MCC measures the correlation between predicted and actual classifications, considering all four confusion matrix elements. MCC = (TP * TN - FP * FN) / sqrt((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))	Suitable for assessing overall classification performance in both balanced and imbalanced datasets.

Table 1. (continued)

Accuracy	Accuracy represents the proportion of correctly predicted instances among the total instances. Accuracy = (TP + TN) / (TP + TN + FP + FN)	Suitable when class distribution is relatively balanced across classes.

3. Results

3.1. Best Performing Algorithms

In this study, nine machine learning algorithms were chosen. First, each algorithm was applied on the training set to familiarize the model with the data's patterns and relationships, facilitating its ability to make predictions. Subsequently, the model's performance was evaluated on the testing set using the six distinct evaluation metrics as mentioned previously. These nine machine learning algorithms include: DecisionTreeClassifier (DTC), a method that builds a tree-like structure to make decisions by partitioning the dataset based on features; GaussianNB (GNB), a classifier based on Bayes' theorem, assuming features are independent and follow a Gaussian distribution; KNeighborsClassifier (KNC), a classifier based on the majority class among its k nearest neighbors in the feature space; LGBMClassifier (LGBMC) Light Gradient Boosting Machine, a boosting algorithm that uses gradient boosting framework and is optimized for efficiency; LogisticRegression (LR), modeling the probability of a binary outcome using a logistic function and linear regression; MLPClassifier (MLP) Multi-Layer Perceptron, a type of neural network with multiple layers of nodes, used for classification tasks; RandomForestClassifier (RFC), constructing multiple decision trees and merging their outputs to improve accuracy and control overfitting; SVC (SVC) Support Vector Classifier, a classifier based on finding the hyperplane that best separates classes in a high-dimensional space; XGBClassifier (XGBC) Extreme Gradient Boosting, a boosting algorithm known for its speed and performance in tabular datasets [5-9].

Figure 3 presents a summary of the model evaluation outcomes. Both the DTC and LR algorithms consistently display comparatively lower scores across all metrics when compared to the average. GNB also exhibits relatively lower performance across most metrics. Conversely, the KNC, MLP, and XGBC algorithms consistently demonstrate comparatively higher scores than the average across all metrics. These three algorithms showcase robustness and reliability across different evaluation criteria, standing out as top performers for predicting Parkinson's disease.

Figure 4 highlights the weighted average All Metrics score for each of the nine algorithms, where KNC, MLP, and XGBC achieve the highest score of 0.87, while GNC, LR, and DTC attain the lowest scores of 0.79, 0.8, and 0.81, respectively. Although the ideal score or range of scores can vary depending on the specific use case, a general guideline is that a score of 0.8 or higher is often considered a good score for most model evaluation metrics [10].

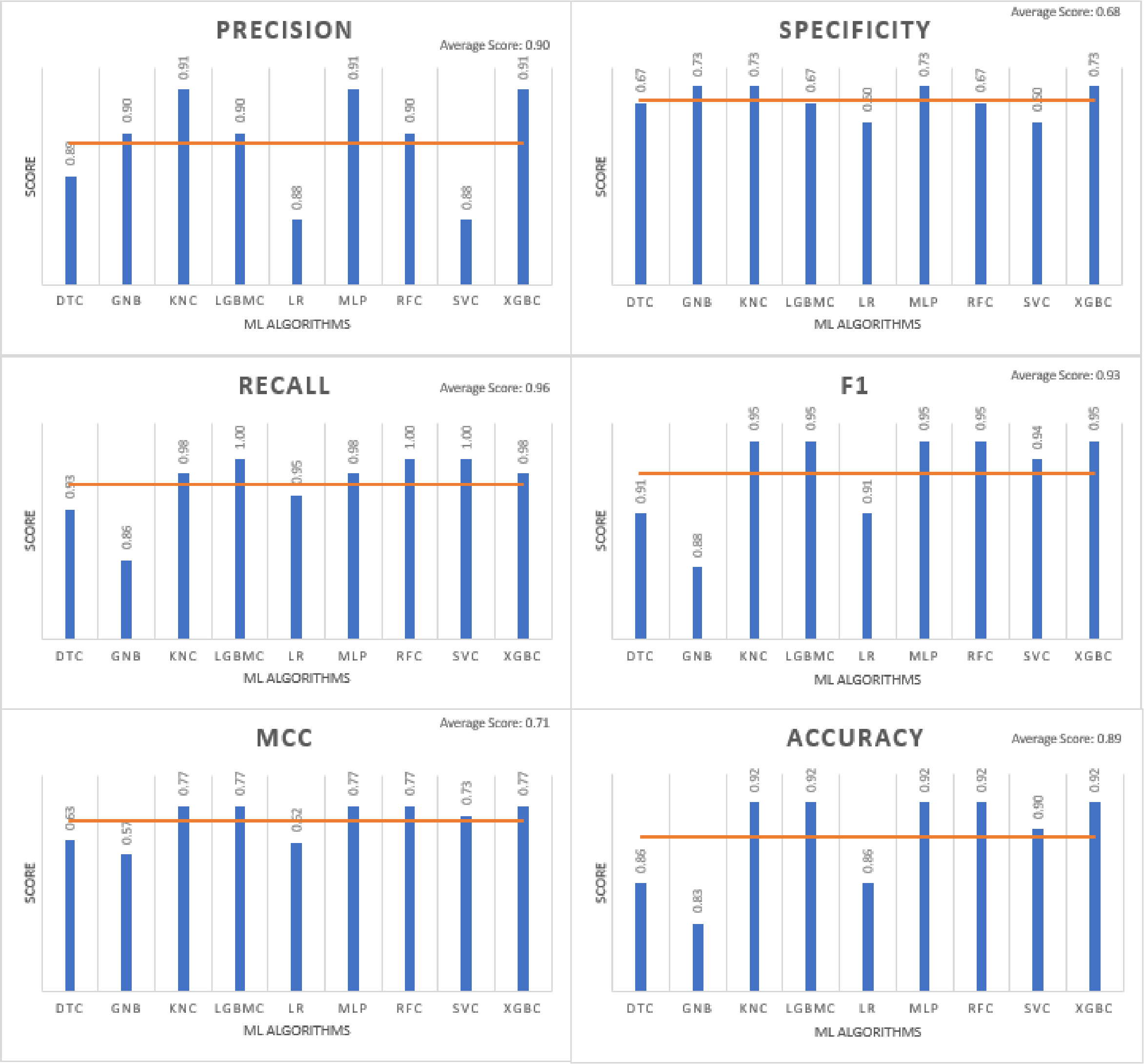


Figure 3. Summary of Model Evaluation Outcomes

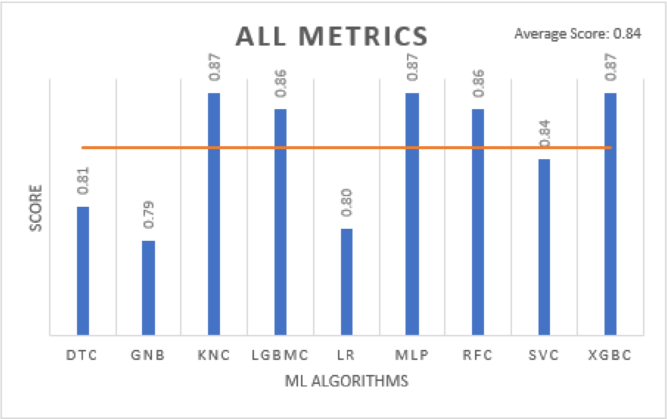


Figure 4. Weighted Average of All Metrics for All Nine Algorithms

3.2. Impact of Feature Reduction on Model Performance

Furthermore, the entire process of model fitting and model evaluating was replicated using the original dataset containing all 22 features. Displayed in Figure 5 are the weighted average scores of all metrics using both 22 and 12 features across the nine algorithms. Dropping the previously identified 10 redundant features resulted in either identical or improved model performance in 6 out of the 9 algorithms. For the remaining 3 algorithms, there was only a slight decrease in model performance, which was more than compensated by the advantage of a more streamlined model, reducing redundancy and multicollinearity. This observation validates the earlier decision to eliminate those 10 features.

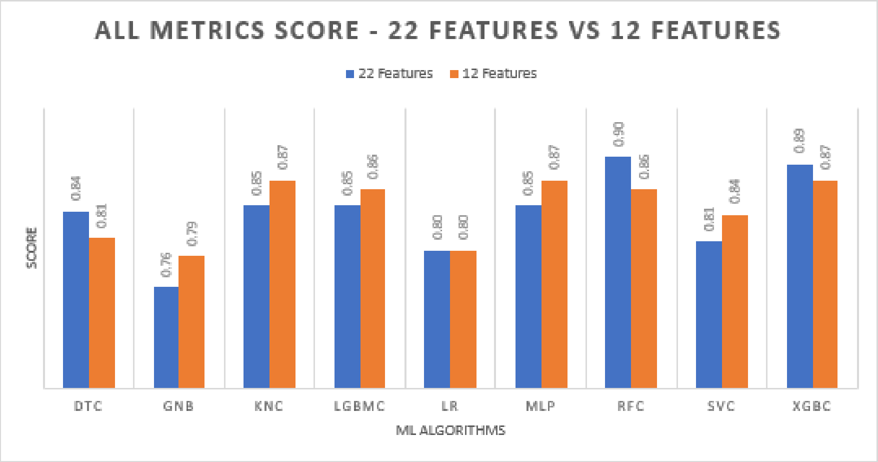


Figure 5. All Metrics Score of Models with 22 Features and 12 Features

4. Conclusion

Diagnosing Parkinson’s disease has traditionally involved clinical assessments by neurologists, and this practice still persists today to a significant extent. However, clinical assessments can be prone to subjectivity. In this study, a comprehensive predictive modeling approach was undertaken, employing nine distinct machine learning algorithms and six different model evaluation metrics to identify the best-performing algorithms. The findings reveal that, using only 12 vocal characteristics, KNeighborsClassifier (KNC), MLPClassifier (MLP), and XGBClassifier (XGBC) achieved the highest score of 0.87. This score is generally considered very good, indicating that the model is robust and possesses strong predictive power. This study marks a crucial initial step in leveraging machine learning techniques for more effective and potentially more accurate diagnosis of Parkinson’s disease based on patients’ vocal characteristics.

References

[1] Pagano G, Niccolini F, Politis M. Imaging in Parkinson’s disease. Clin Med (Lond), 2016,16(4): 371-5.

[2] Rusz J, Tykalová T, Novotný M, et al. Distinct patterns of speech disorder in early-onset and late-onset de novo Parkinson’s disease. NPJ, 2021.

[3] Kaggle, Parkinson’s Disease Dataset. (2021), Available online at: https: //www.kaggle.com/datasets/gargmanas/parkinsonsdataset

[4] Naidu G, Zuva T, Sibanda, E. M. A Review of Evaluation Metrics in Machine Learning Algorithms In Artificial Intelligence Application in Networks and Systems. Springer, 2023.

[5] Iqbal H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI, 2021, 2 (160).

[6] Bonaccorso G. Machine Learning Algorithms. Packt Publishing, 2017

[7] Song YY, Lu Y. Decision tree methods: Applications for classification and prediction. Shanghai Arch Psychiatry, 2015, 27(2): 130-5.

[8] McCarty DA, Kim HW, Lee HK. Evaluation of Light Gradient Boosted Machine Learning Technique in Large Scale Land Use and Land Cover Classification. Environments, 2020, 7(10): 84.

[9] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics, 201, 15(1): 41-51.

[10] Vidhya, 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023). (2023), Available online at: https: //www.analyticsvidhya.com/ blog/2019/08/11-important-model-evaluation-error-metrics/