A Comparative Analysis of Feature Selection Algorithms for Parkinson's Disease Classification

T. N. Chitradevi[1], Dr. P. Swathypriyadharsini[2], S. Jansi Rani[3], S. Gokila[4], S. Senthilnathan[5]

[1] *Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamil Nadu, India. Email: chitradevi.04@gmail.com*
[2] *Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamil Nadu, India. Email: swa.pspd@gmail.com*
[3] *Assistant Professor (Sl. G), Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India. Email: jansi.sankar@srec.ac.in*
[4] *Assistant Professor, Department of Information Technology, Karpagam Institute of Technology, Coimbatore, India. Email: gokilasadhasivan@gmail.com*
[5] *Assistant Professor, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamil Nadu, India. Email: senthil.sn1985@gmail.com*

KEYWORDS
Parkinson's disease, Feature Selection, Classification, SVM, AdaBoost, Decision Tree, Random Forest, Logistic Regression, XGBoost

ABSTRACT
Parkinson's disease is a progressive neurological disorder that adversely affects the quality of life of an individual and is hard to diagnose and treat at its initial stages. It helps in proper diagnosis and targeted treatment by classifying the acoustic features extracted from people suffering Parkinson's disease and healthy subjects responsible for this disease. Conventional classification techniques have considerable challenges with the low sample size and high dimensionality of acoustic features data of Parkinson's disease. Techniques for feature selection have been developed as efficient tools that may eliminate redundancy and discover the most relevant features. The purpose of this paper is to analyze three types of feature selection techniques, namely, filters, wrappers, and embedded methods, in order to identify the important acoustic features. These features are utilized to train and test the models for classification, SVM, AdaBoost Classifier, Decision Tree Classifier, Random Forest Classifier, Logistic Regression and XGBoost. For all these classification models, by this feature selection process, their precision improved by quite a margin even when features have been minimized. Of all these models, Random Forest and XGBoost Classifier has high precision and is very reliable. AdaBoost Classifiers also demonstrate good performance, showing that they may be very promising for dependable classification in such datasets. The present study determines key acoustic features that are crucial in the course of Parkinson's disease and brings forth new information on the molecular mechanisms of the disease. Integration of feature selection methods with machine learning algorithms helps to enhance diagnostic accuracy and optimizes computational efficiency.

1. Introduction:

Parkinson's disease (PD) is a progressive neurodegenerative disorder primarily affecting motor function and gradually decreasing the quality of life for millions of people around the world. Early diagnosis is critical for symptom management and slowing down disease progression, but this is difficult due to the overlap of symptoms with other diseases and the absence of any definitive biomarkers [1]. The very high dimensionality of the data combined with a rather small sample size poses enormous challenges for the traditional classification algorithms, calling for robust feature selection techniques that can efficiently address these challenges.

Feature selection is one of the important steps in the analysis of high-dimensional data, as it identifies the most relevant and non-redundant features that are most relevant to the classification model. This improves the accuracy of the classification model and reduces the computational complexity as well as training time. Statistical techniques that are commonly used for feature relevance determination include ANOVA, Chi-Square, and Information Gain. Two more wrapper methods are Forward Selection and Backward Elimination, which evaluate the subsets of features iteratively to make sure the dataset is refined to give the best performance of the classifier. These are two important methods in avoiding overfitting and in providing better interpretability of complex datasets, which is worthwhile in Parkinson's disease research [2].

Feature selection is an important step in the high-dimensional data analysis because it determines the most relevant and discriminating features classification models require. This makes the accuracy of the classification model much better while reducing the time it takes for training and its computational complexity. Some of the commonly used statistical methods to check the relevance of features include ANOVA, chi-squared, and information gain. Two more wrapper approaches that iteratively refine the dataset with promising subsets for highest classifier performance are Backward Elimination and Forward Selection. These methods are highly important for research in Parkinson's disease because they enhance the explainability of complex datasets and contribute to reducing the risk of overfitting.

These were used to classify Parkinson's disease effectively: Support Vector Machines (SVM), Logistic Regression, Random Forest, Decision Trees, AdaBoost Classifier, and XGBoost. These algorithms handle high-dimensional data complexity pretty well and have shown robust performances in similar domains. Random Forest and XGBoost stand out for their ability to handle large feature sets with high accuracy and efficiency. This addition of feature selection techniques greatly enhances the performance of such classifiers by noise reduction and focusing on the more informative features [3].

Reduction in difficulties in diagnosing patients with Parkinson's disease, combining feature selection methods, is achieved with sophisticated machine classifiers. The combination of different techniques, such as statistics, wrappers, and embedding, ensures a deep level of feature selection approach which will make the discovery of important biomarkers possible. This research focuses attention on the importance of feature selection and machine learning application to boost diagnostic precision and provide more profound insights into the molecular pathways responsible for Parkinson's disease. Focused therapy and early diagnostics tools of this disabling disease could be formulated by putting emphasis on features which are relevant for the diagnosis.

2. Related Work

Parkinson's disease is one of the most challenging neurological disorders affecting millions worldwide, mostly the aged above 60. During its progression, motor disability and speech impairment are manifested to worsen. Early detection and diagnosis are important, and the advancement in machine learning (ML) and data mining techniques has shown great potential results in enhancing the accuracy of classification of PD [4]. Several studies have investigated the performance of different classifiers, including SVM, RF, and decision trees, for the diagnosis of PD from acoustic features extracted from voice recordings [6].

Feature selection is one of the key factors in the PD classification process by reducing the dimensionality of the dataset and improving the performance of classifiers. The most common methods of feature selection used to find out the most relevant features for the classification of PD are ANOVA, Chi-square, and Information Gain [4, 7]. Furthermore, wrapper-based approaches, including forward selection and backward elimination, have been found to improve model performance through selecting the best possible features for the classifier [11]. The latest advanced approaches of LGBM and XGBoost are now also increasingly being used for feature selection since they handle large data volumes and also increase model performance [10].

In recent studies, the integration of feature selection methods and ensemble learning techniques has also been explored. Saeed et al. (2023) [9] proved that using ensemble classifiers with optimal feature subsets, such as stacking and bagging, improved the accuracy of PD detection. The application of evolutionary algorithms in feature selection to enhance classifier performance was further explored in other works [12]. These methods help reduce unwanted or unnecessary features, which improve generalization and then lead to more accurate predictions [13].

Some other studies on the classifier's performance were evaluated on different algorithms that detect PD from voice data. Among common classifiers, SVM and random forests were identified. The reason why SVM works so well with high-dimensional datasets is because of its ability to deal with complicated feature spaces [6]. Recursively, recent research has shown that the developed gradient boosting algorithm, XGBoost for PD detection gives very high accuracy and robust results [15, 9]. Another ensemble classifier is AdaBoost and GBM. It has also proved to be performing well in the classification of PD because they handle problems of the relationship between linear and nonlinear data [14, 23].

The integration of machine learning algorithms with feature selection techniques is the key to improving the diagnosis and prediction of Parkinson's disease. An increase in the use of advanced methods such as XGBoost, LGBM, and ensemble learning will increase the accuracy of detecting PD, thereby improving patient outcomes by providing early diagnosis and intervention. Future research should focus on perfecting these techniques to be applicable in real-world settings.

3. Methodology:

The methodology of this project in figure 1 is structured into two phases, namely: (i) feature selection from acoustic features data for training and testing and (ii) evaluation of the effectiveness of the selected features using a variety of classifiers. In capturing different dimensions of the significance of features, this research combines filter, wrapper, and embedded methods.

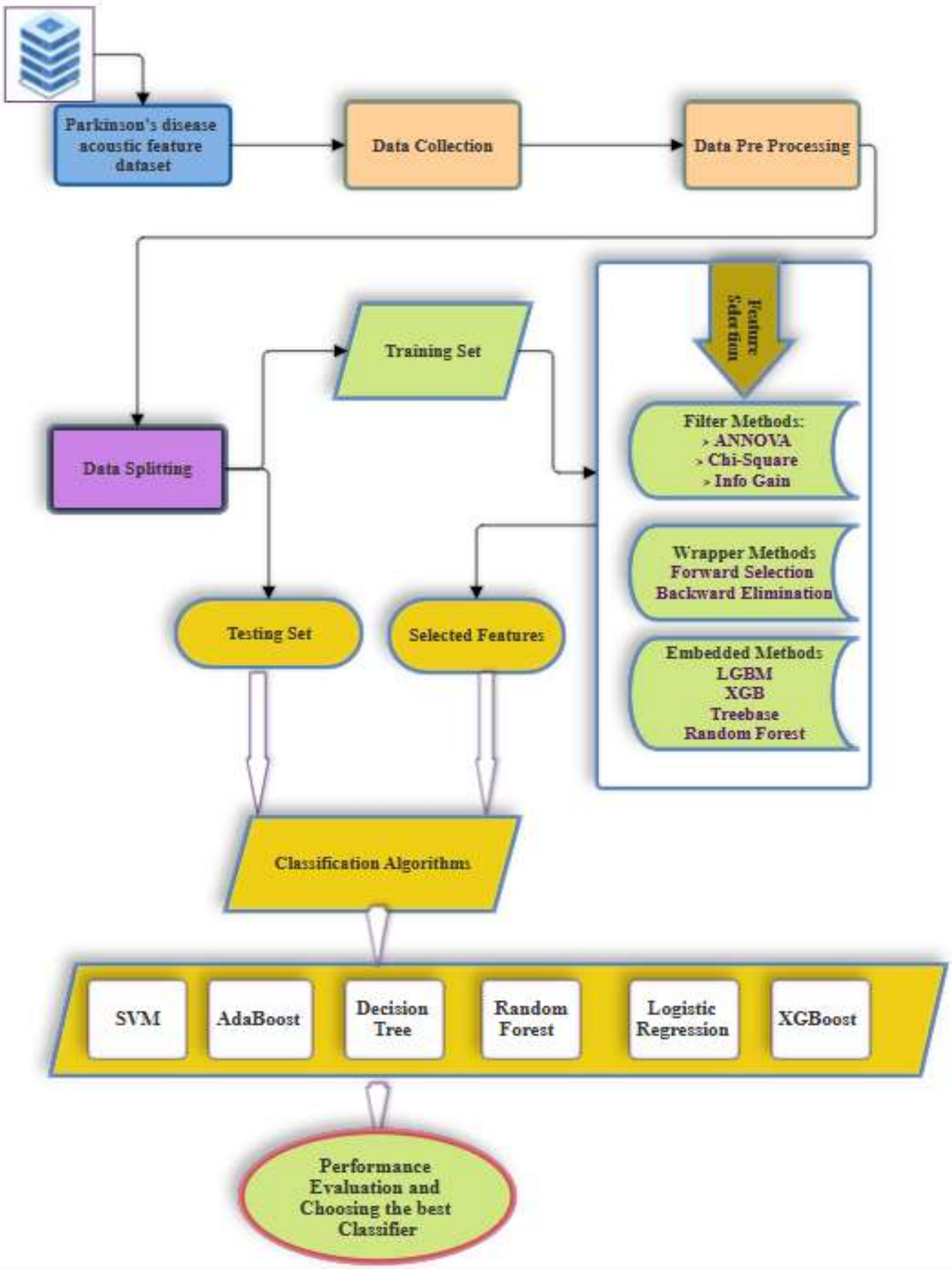

Figure 1: Methodology of this project

3.1 Feature Selection

3.1.1 Filter Methods

ANOVA, Chi-Square, and Information Gain evaluate the feature's importance according to statistical metrics without classifier dependence. Such methods are more appropriate for high dimensional datasets like acoustic features datasets as they have the effect of reducing the computation complexity and preserving the intrinsic properties of the data. The use of these methods makes data analysis faster and efficient, enhancing the model efficiency while leading to a better classification result. Their straightforward implementation makes them a valuable choice in biomedical applications, including Parkinson's disease diagnosis.

3.1.2 Wrapper Methods

Methods, like Forward Selection and Backward Elimination, improve feature subsets based on the collective contribution they make to the classifier. They are particularly good at discovering feature interdependencies, so that a more global selection is achieved. Forward Selection adds features one at a time, whereas Backward Elimination removes less important features one at a time, trying to optimize the feature set for maximum model accuracy. Although computationally intensive, it is superior because the selection made for feature selection will focus on the needs of that classification model, hence efficient for complex datasets.

3.1.3 Embedded Methods

There are two algorithms which actually perform feature selection within the training process of a Model-Lasso Regression, and Random Forest. Lasso Regression applies L1 regularization: this causes smaller coefficients of features to get reduced to zero; therefore, it is one of the efficient techniques used in the case of high-dimensional sparse data. The Random Forest algorithm can rank the importance of features by modeling nonlinear interactions and managing complex relationships. These methods improve the interpretability of the model and also result in efficient feature selection as their training process optimizes both for predictive accuracy and computational efficiency.

After feature selection, subsets of varying sizes, for example, 50, 100, 200 are evaluated following the ranking of the best-performing features. With classifiers such as Support Vector Machines (SVM), Logistic Regression, Decision Trees, XGBoost, AdaBoost, and Random Forest, the 20 most informative features are selected to further investigate. The strength of each feature selection method in maximizing the accuracy of prediction while decreasing computational cost is determined by benchmarking the performances of different classifiers.

3.2 Classifier

3.2.1 SVM

Support Vector Machine (SVM) is a very powerful supervised learning algorithm primarily used for classification tasks. SVM works by mapping data points into some high-dimensional feature space and then finding the optimal hyperplane that maximizes the margin between classes. The main goal of SVM is to maximize the gap between classes, while the boundary is determined by the support vectors, which are the data points closest to the hyperplane. This attribute allows SVM to perform generalization well, even in a high-dimensional space.

SVM applies kernel functions to change non-linearly separable data into a higher-dimensional space in which it becomes linearly separable. These include the use of linear, polynomial, and Radial Basis Function (RBF) types of kernels. The model depends on the kernel chosen, so the margin created in producing an accurate result. SVM has shown great predictive performance in high-dimensional data, so it is a good option to use in areas such as acoustic features classification, where accuracy and robustness are a must.

3.2.2 AdaBoost Classifier:

AdaBoost is an ensemble learning method which is adaptive in nature, combining a series of weak classifiers to come up with a strong classifier. This core idea of adaptive boosting is to train a sequence of weak learners, normally decision stumps, iteratively. Each subsequent model concentrates on the errors made by the previous ones. The process is achieved by increasing the weights of misclassified instances so that they are important in the next iteration. The model becomes stronger and more accurate with time, through correcting the previous misclassifications.

Although AdaBoost is adaptive in nature, it is great at improving predictive accuracy, especially with weak classifiers. It unfortunately is sensitive to noisy data and outliers due to the distortion of model performance with incorrect weight adjustments. However, AdaBoost often results in better classification accuracy than other algorithms for most binary tasks in classification. Its efficiency, ability to handle different types of data, and excellent ability in classifying predictions based on data features make it popular for a vast number of classification problems.

3.2.3 Decision Tree:

The decision tree represents an approach for supervised learning that can also be used for regression and classification problems. It recursively breaks down the data to determine the best way to split and optimize information gain or decrease in impurity at each node. With internal nodes representing decisions based upon features, branches representing the decision rules, and result or class labels by representing leaves, the model remains incredibly easy to understand as well as use due to its tree-like structure.

While decision trees are very effective at recognizing complex patterns in data, they tend to overfit the data, especially when it comes to high-dimensional data. Such pruning techniques that have been incorporated into Decision Trees and modified for generalization include restrictions on tree depth and removal of branches due to low predictive value. Using methods such as Random Forest, which averages the output of several trees and lowers the impact of overfitting, decision trees can truly come to excellent accuracy in two-class and multiclass classifications.

3.2.4 Random Forest:

In the training phase, Random Forest ensemble learning technique builds multiple decision trees. It tries to reduce association between trees by choosing random subsets of features at each node and constructs each tree using bootstrapped samples from the dataset. This will give a varied collection of decision trees, thus improving the model's overall resilience. Once the trees are built, the algorithm combines their predictions by averaging them for regression tasks or by majority voting for classification tasks.

The ability of Random Forest to combine multiple models, namely decision trees, boosts the performance and reduces overfitting, which accounts for its accuracy. Ensuring diversity in the trees, every tree is trained on a random subset of data points throughout the training phase. For predictions, the average value (for regression) or the majority class prediction (for classification) acts as the basis for the final output. A random forest is a reliable alternative to complex datasets with high dimensionality because it evaluates feature relevance, handles high-dimensional data well, and produces great accuracy.

3.2.5 Logistic regression

Logistic regression is a statistical model applied to classification problems where the problem is one of two types. Logistic regression is quite different from linear regression, because it produces a continuous output, not a continuous value. Instead, logistic regression uses a logistic function (sigmoid) to map input features to a probability between 0 and 1. That probability corresponds to the likelihood that a given data point falls into one class or the other. The model applies a log-odds transformation in order to estimate the relationship of the dependent variable (binary outcome) with independent variables (features).

Logistic regression uses maximum likelihood estimation for fitting the best line or hyperplane between two classes. This gives a chance of the presence of a data point belonging to the positive class through applying the logistic function and calculating the coefficients for each attribute within the dataset. In terms of prediction, a sample is classified as positive when the computed probability is larger than 0.5. Otherwise, it is assigned as negative. Logistic regression is very interpretable with high computational efficiency. A clear picture of how each component can affect the anticipated result arises from it. It is the most reliable model for the purpose of binary classification problems. There are techniques for further increases in accuracy such as feature scaling and regularization.

3.2.6 XGBoost:

XGBoost is a strong machine learning technique specifically developed for supervised learning scenarios like classification and regression. Its architecture is based on gradient-boosting concepts. That being said, multiple weak learners namely decision trees are used through an iterative process, by which the mistake that prior trees have committed to minimize the loss. Therefore, XGBoost will utilize the learning process of gradient descent by which it enhances the accuracy of the model. One of its prominent characteristics is the use of regularization terms that help in generalization and avoid overfitting. Therefore, it works very well with a wide range of datasets, especially complicated, high-dimensional data.

XGBoost creates trees one after the other to fix the mistakes caused by the previous tree. In the weighted sum of trees it employs, every tree tries to reduce the prediction errors. XGBoost uses column subsampling, shrinkage, and proper handling of missing data. The final model uses the entire output from all trees to create predictions that are incredibly accurate. XGBoost is extensively applied both in competitions and practice due to its excellent computational efficiency, scalability, and handling capacity of large-scale multi variation data.

4. Experimental Results

4.1 Dataset

Publicly available acoustic feature data drawn from voice recordings of people with Parkinson's disease and healthy controls to explore early detection in vocal patterns [24]. It uses the features that include jitter, shimmer, harmonics-to-noise ratio, and fundamental frequency, important for the detection of very minor vocal changes, which are symptoms of this disease given in Table 1. The split for consistency was 70-30 train-test. This gives enough data for training the models while keeping the test set unbiased. The 70-30 split is helpful for the comparison of feature selection techniques and classifiers under controlled conditions, simplifying the experimental process with reliable and consistent results for all experiments. This methodology enables the effective comparison of classifier performance, reflecting real-world applications in the detection of Parkinson's disease.

Table 1: Feature Description

Names	Role	Variable type	Description
Subject	ID	Categorical	Subject identification
Status	Target	Categorical	PD = 1; Healthy subject = 0
Age	Feature	Numerical - Discrete	Patients’ age in years
Sex	Feature	Feature Categorical	Men = 0; Woman = 1
Jitter	Feature	Feature Numerical - Continuous	Jitter
Shimmer	Feature	Feature Numerical - Continuous	Shimmer

LZ-2	Feature	Feature Numerical - Discrete	Lempel ziv complexity (order 2)
CPP	Feature	Feature Numerical - Continuous	Cepstral peak prominence
Hurst	Feature	Feature Numerical - Continuous	Hurst's exponent
MFS	Feature	W Feature Numerical - Continuous	Multifractal spectrum width
Shannon	Feature	Feature Numerical - Continuous	Shannon entropy
Permutation Permutation entropy	Feature	Feature Numerical - Continuous	
PPE	Feature	Feature Numerical - Continuous	Pitch period entropy
FMMI	Feature	Feature Numerical - Discrete	First minimum in mutual information
FZCF	Feature	Feature Numerical - Discrete	First zero in correlation function
GNE	Feature	Feature Numerical - Continuous	Glottal to noise excitation
ZCR	Feature	Feature Numerical - Continuous	Zero crossing rate
D2	Feature	Feature Numerical - Continuous	Correlation dimension
HNR	Feature	Feature Numerical - Continuous	Harmonic to noise ratio
RPDE	Feature	Feature Numerical - Continuous	Recurrence period density entropy
GQ_prc5_95	Feature	Feature Numerical - Continuous	Glottal quotient 5 to 95 percentile
GQ std cycle open	Feature	Feature Numerical - Continuous	Glottal quotient stand. deviation open
GQ std cycle closed	Feature	Feature Numerical - Continuous	Glottal quotient stand. deviation closed
MFCC0	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 0
MFCC1	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 1
MFCC2	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 2
MFCC3	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 3
MFCC4	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 4
MFCC5	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 5
MFCC6	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 6
MFCC7	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 7
MFCC8	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 8
MFCC9	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 9
MFCC10	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 10
MFCC11	Feature	Feature Numerical - Continuous	Me frequency cepstral coefficient 11
MFCC12	Feature	Feature Numerical - Continuous	Mel frequency cepstral coefficient 12

4.2 Result and Discussion

The baseline performance was evaluated directly on the dataset using classifiers including Support Vector Machine (SVM), AdaBoost Classifier, Decision Tree, Logistic Regression, Random Forest, and XGBoost. Afterward, feature selection techniques were included to improve the classification process. Filter methods were used in assessing and ranking the importance of the features through Information Gain, ANOVA, and ChiSquare. For wrapper methods, Forward Selection and Backward

Elimination have been used by iteratively refining the feature subset. Additionally, the use of embedded methods explores the inherent feature selection capabilities that models such as LGBM and Tree-based algorithms inherently possess. Combining all these techniques means having to assure that feature selection strategies have had an all-round study for improved classification accuracy concerning Parkinson's disease prediction.

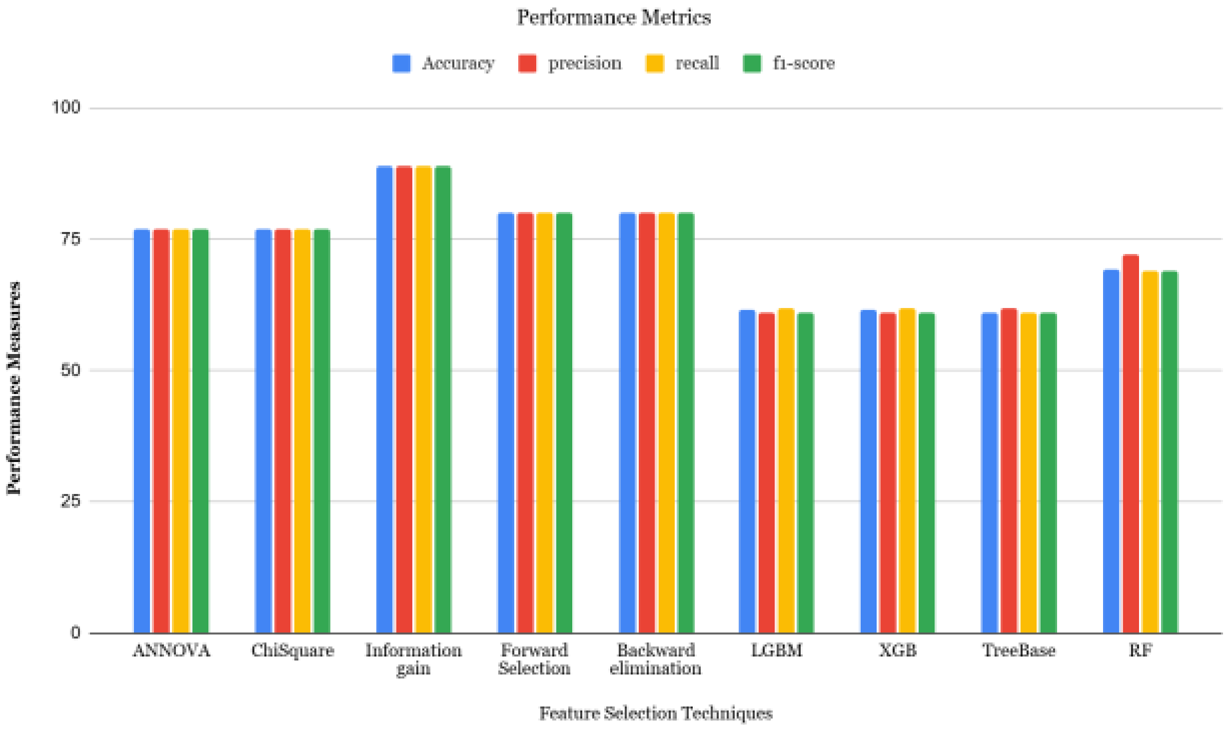

Figure 2: SVM Performance

Figure 2 illustrates the performance evaluation of the SVM model with various feature selection methods, namely filter methods: Information Gain, ANOVA, and ChiSquare; wrapper methods: Forward Selection and Backward Elimination; and model-based approaches: LGBM, XGB, TreeBase, and Random Forest. Information Gain achieved the highest performance with an accuracy of 88.88% and consistent precision, recall, and F1-score values of 89.00%. Both ANOVA and ChiSquare had moderate results with accuracy rates of 76.92% and respective metrics of 77.00% for precision, recall, and F1-score. The wrapper methods, Forward Selection and Backward Elimination, were also equal in terms of performance with the accuracy of 80.00% and identical metrics. However, the model-based approaches showed poor performance since LGBM and XGB achieved an accuracy of 61.53%, precision, recall, and F1-scores of around 61.00%, while TreeBase was scored equally at 61.00%. Random Forest performed the best with accuracy at 69.23%, precision at 72.00%, recall at 69.00%, and F1-score at 69.00%.

The following ROC curves Figure 3 illustrates the trade-off between True Positive Rate (Sensitivity) and False Positive Rate across thresholds, giving insights into the performance of SVM using different feature selection techniques. The maximum area under the ROC curve was achieved by Information Gain, showing its strength in classifying the classes well. ANOVA and ChiSquare also obtained average AUC values as their performance is consistent but not outstanding in terms of classification. Wrapper methods such as Forward Selection and Backward Elimination had slightly improved AUC, indicating the iterative refinement of relevant features. Model-based approaches were LGBM and TreeBase, whose AUC values were low and, therefore, reflected its suboptimal performance. Whereas AUC values of Random Forest and XGB were different, Random Forest was much better for several reasons: higher precision and recall values. This shows that feature selection is very important in raising the sensitivity and specificity of the classifier, and Information Gain worked best in this experiment for this task.

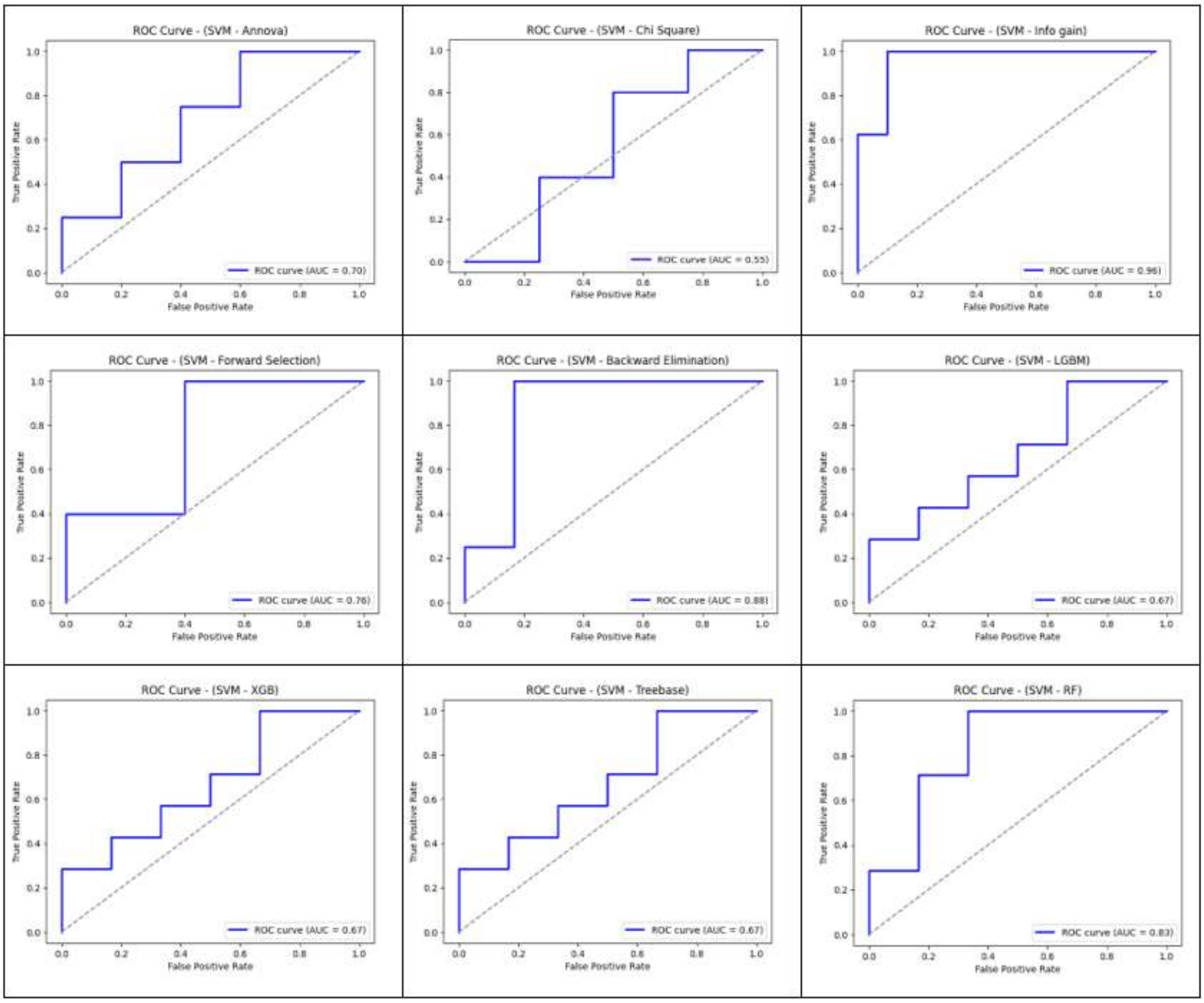

Figure 3: ROC Curves of SVM with Feature Selection Techniques

Figure 4 illustrates the performance of the AdaBoost Classifier with respect to different feature selection methods such as filter methods- ANOVA, ChiSquare, Information Gain, wrapper methods- Forward Selection, Backward Elimination and model-based approaches-LGBM, XGB, TreeBase, Random Forest. Among the techniques, the feature selection method based on ANOVA was seen as the best that produced 88.89% accuracy. Precisely, the values for precision, recall, and F1 were all 91.00%, 89.00%, and 89.00%, respectively. ChiSquare had a moderate value with an accuracy of 76.92%, while precision, recall, and F1-scores were at 78.00%, 77.00%, and 77.00%. Information Gain had an accuracy of 83.33%, with the precision, recall, and F1-scores measured at 84.00%, 83.00%, and 83.00%. Wrapper methods, which are Forward Selection and Backward Elimination, performed less well with an accuracy score of 80.00% with slightly varied metrics, where the F1-score decreased by less for Backward Elimination at 79.00%. Among the model-based methods, the XGB and TreeBase resulted in good performance with an accuracy of 84.62% and all other metrics at 85.00%. LGBM resulted in a satisfactory accuracy of 77.78%, but with overall performance inferior to that obtained by XGB and TreeBase. Random Forest resulted in 77.78% accuracy; precision at 84.00%, recall at 78.00%, and F1-score at 76.00%.

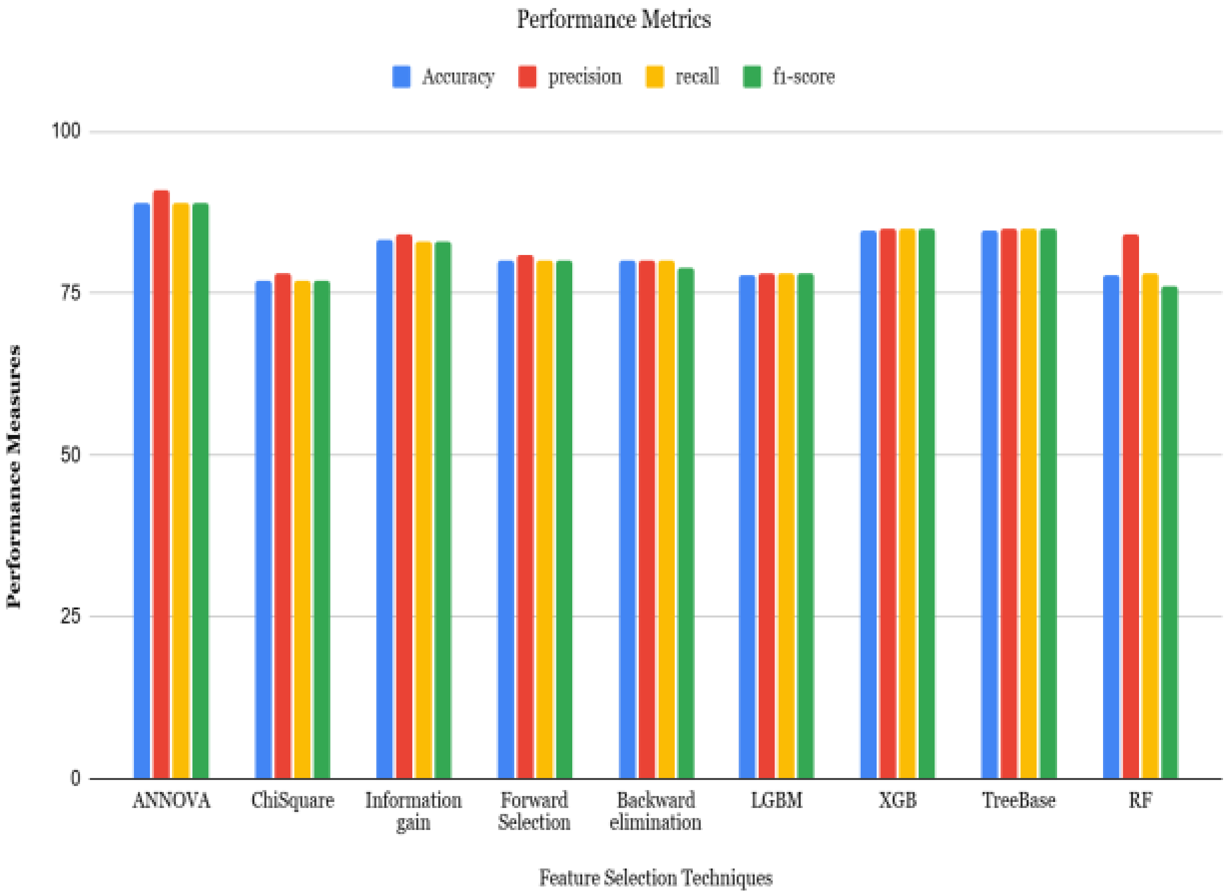

Figure 4: AdaBoost Classifier Performance

The following ROC curves Figure 5 illustrates the trade-off between True Positive Rate (Sensitivity) and False Positive Rate across thresholds, providing insights into the performance of the AdaBoost Classifier with different feature selection techniques. The maximum area under the ROC curve (AUC) was achieved by Information Gain, highlighting its effectiveness in distinguishing between classes. ANOVA and ChiSquare showed moderate AUC values, indicating consistent but average classification performance. Wrapper methods, such as Forward Selection and Backward Elimination, demonstrated slight improvements in AUC due to the refinement of selected features. Among the model-based methods, XGB stood out with a significantly high AUC, reflecting its superior classification capabilities. In contrast, TreeBase exhibited a lower AUC, signifying weaker performance, while LGBM had satisfactory results but fell short of XGB. Random Forest displayed balanced AUC values, supporting its reliable precision and recall scores. These findings emphasize the critical role of feature selection in enhancing the sensitivity and specificity of classifiers, with Information Gain and XGB being the most effective methods in this experiment for the AdaBoost Classifier.

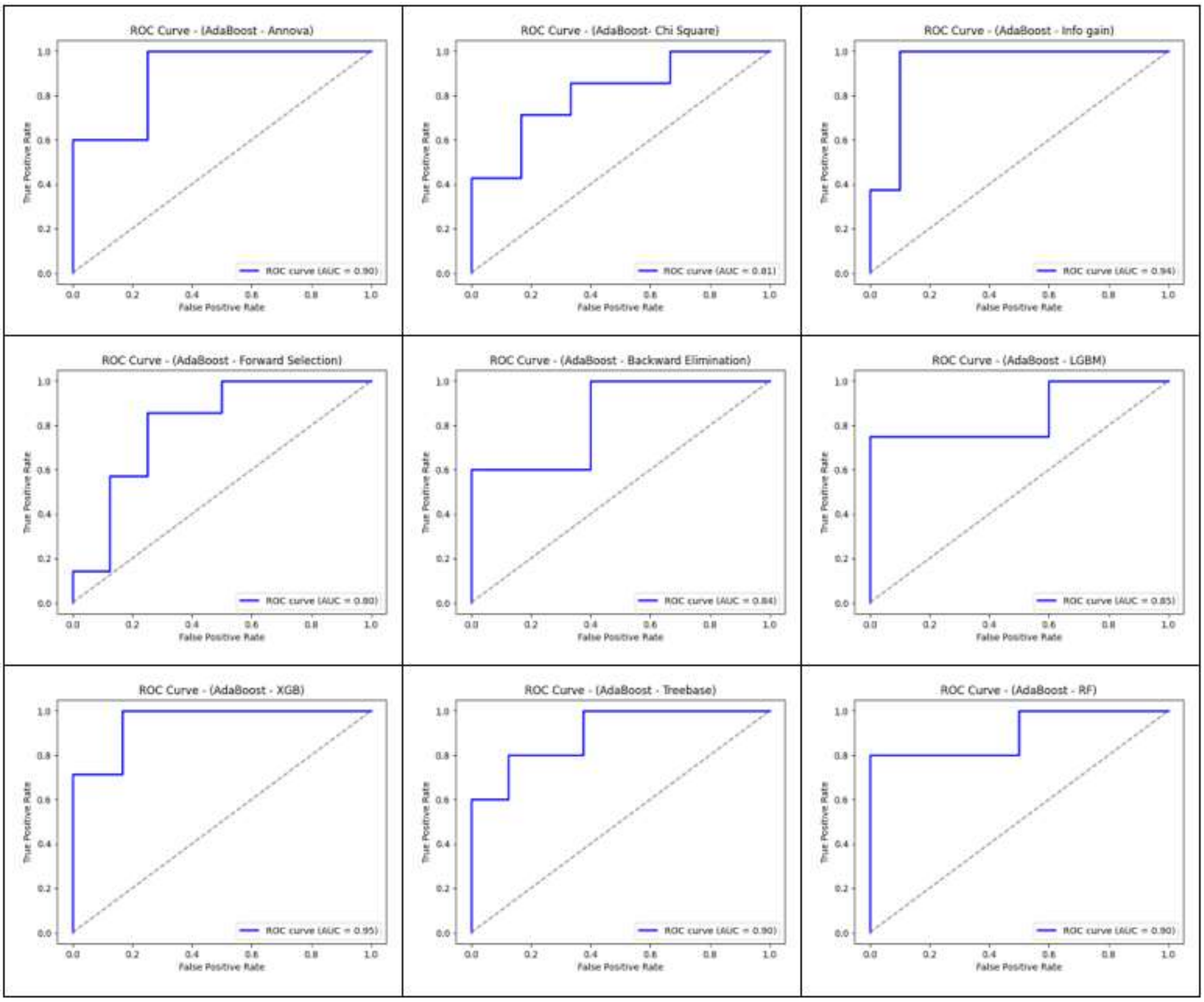

Figure 5: ROC Curves of AdaBoost Classifier with Feature Selection Techniques

Figure 6 shows the performance of the Decision Tree Classifier using various feature selection methods, such as filter methods (ANOVA, ChiSquare, Information Gain), wrapper methods (Forward Selection, Backward Elimination), and model-based methods (LGBM, XGB, TreeBase, Random Forest). Among the filter methods, ANOVA turned out to be the most successful with an accuracy of 84.62%, precision, recall, and F1-scores all of 85.00%. ChiSquare gave a reasonable result with accuracy of 76.92% and precision, recall, and F1-scores of 78.00%, 77.00%, and 77.00%, respectively. Information Gain gave good results with an accuracy of 83.33%, and precision, recall, and F1-scores of 89.00%, 83.00%, and 84.00%, respectively. Among the wrapper methods, Forward Selection performed well with an accuracy of 80.00% and precision, recall, and F1-scores around 80.00%. However, Backward Elimination resulted in lower performance compared to the earlier results with an accuracy of 86.67%, and a precision of 90.00%. Model-based techniques such as XGB, TreeBase, and Random Forests all obtained accuracy of 88.89% with a precision of 92.00%, and recall of 89.00%, and an F1-score of 89.00%. LGBM, though providing good results, also achieved an accuracy of 88.89%, but on the whole, model-based methods and filter methods like ANOVA and ChiSquare proved to be the best for optimizing the performance of the Decision Tree Classifier.

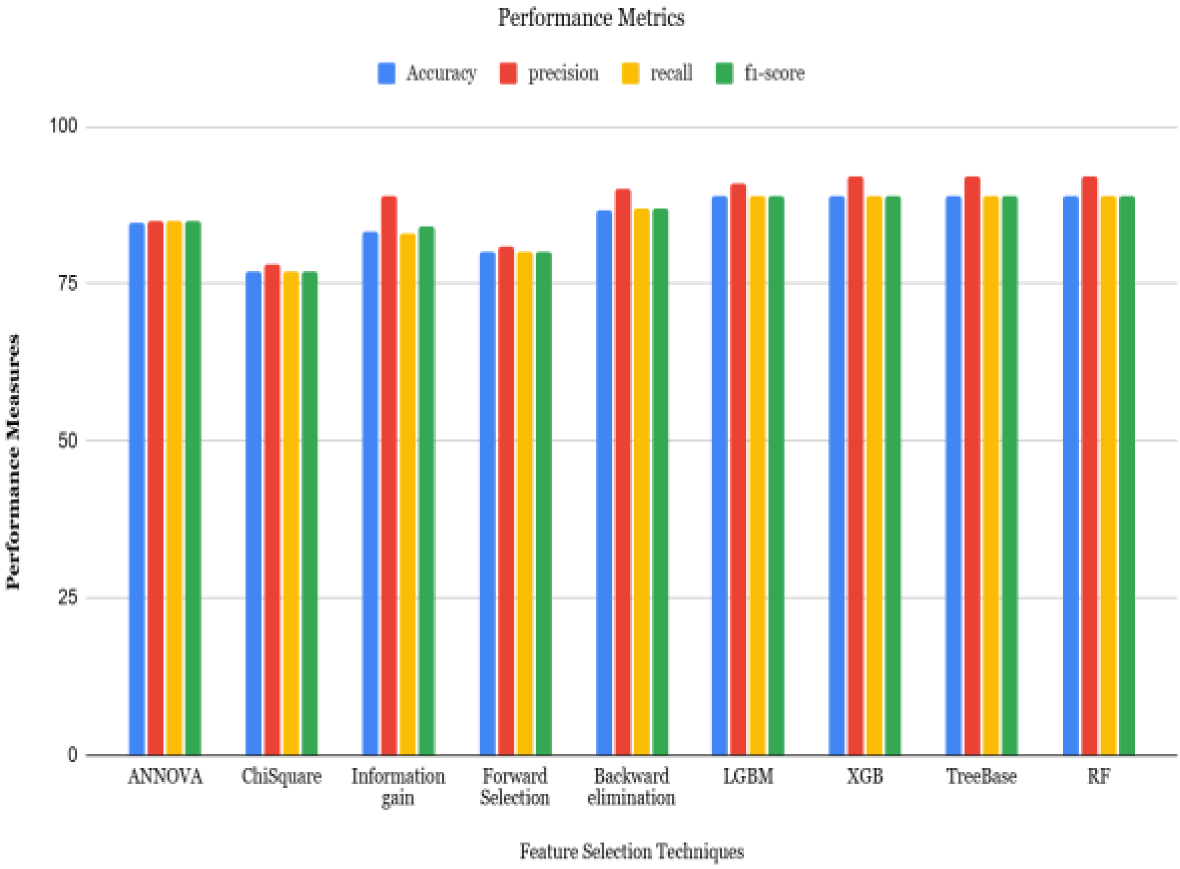

Figure 6: Decision Tree Classifier Performance

The following ROC curves Figure 7 dipicts the trade-off between True Positive Rate and False Positive Rate across thresholds, providing insight into how the Decision Tree Classifier performs with regard to different feature selection techniques. ANOVA and ChiSquare showed the highest AUC for filter methods as they have consistent and reliable classification performances. Information Gain, although less pronounced, demonstrates middle values of AUC; hence, it works properly in particular examples. Wrapper techniques, such as Forward Selection and Backward Elimination, significantly increase the value of AUC, where a better result can be achieved through Forward Selection as it includes feature selection with much detail. Model-based approaches clearly present superior AUC values such as XGB, TreeBase, and Random Forest that could classify perfectly well but LGBM was moderate. The importance of appropriate feature selection, the results stress the sensitivity and specificity of the Decision Tree Classifier; ANOVA, ChiSquare, and model-based methods like XGB and Random Forest, in that order, appear to be the most effective.

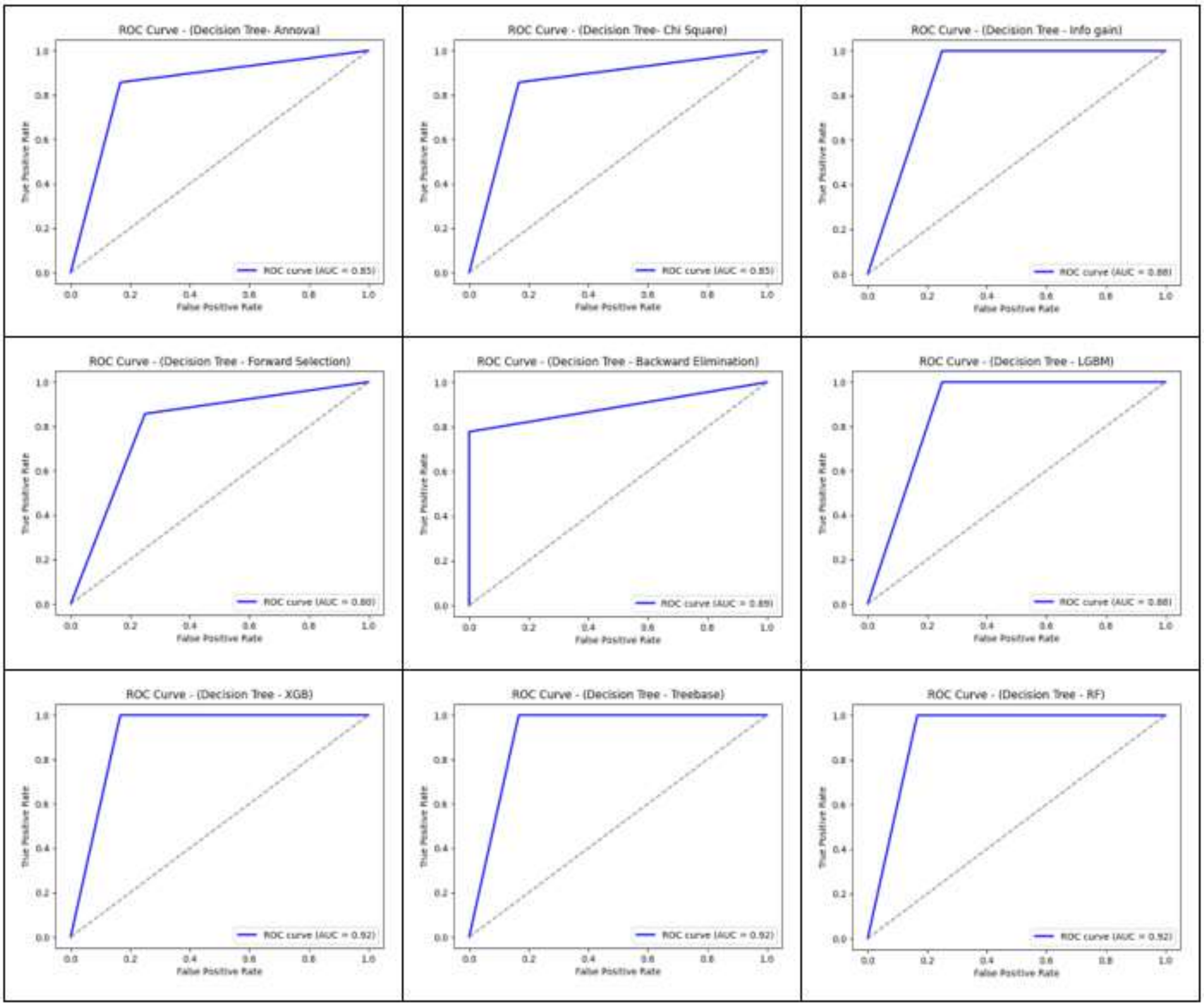

Figure 7: ROC Curves of Decision Tree Classifier with with Feature Selection Techniques

Figure 8 shows the performance of the Random Forest Classifier when evaluated with various feature selection techniques: filter methods, including ANOVA, ChiSquare, and Information Gain; wrapper methods, including Forward Selection and Backward Elimination; and model-based approaches, such as LGBM, XGB, TreeBase, and Random Forest. Among the filter methods, the best result was obtained by using Information Gain, which produced an accuracy of 94.44%, and also had good precision, recall, and F1-scores at 95.00%, 94.00%, and 94.00%, respectively. ANOVA also performed very well with an accuracy of 88.89% and balanced metrics, while ChiSquare showed moderate performance with an accuracy of 77.78%, and precision, recall, and F1-scores of 78.00%. Wrapper methods, Forward Selection and Backward Elimination, also gave the same results; both had an accuracy of 80.00% and their respective precision, recall, and F1-scores were about 80.00%. The model-based methods include LGBM, XGB, TreeBase, and Random Forest, which have a perfect accuracy score of 100.00%, precision of 1.00%, recall of 1.00%, and an F1-score of 1.00%. This outcome reveals that model-based methods, such as Random Forest itself, and filter methods like Information Gain and ANOVA are the best options for boosting the performance of the Random Forest Classifier.

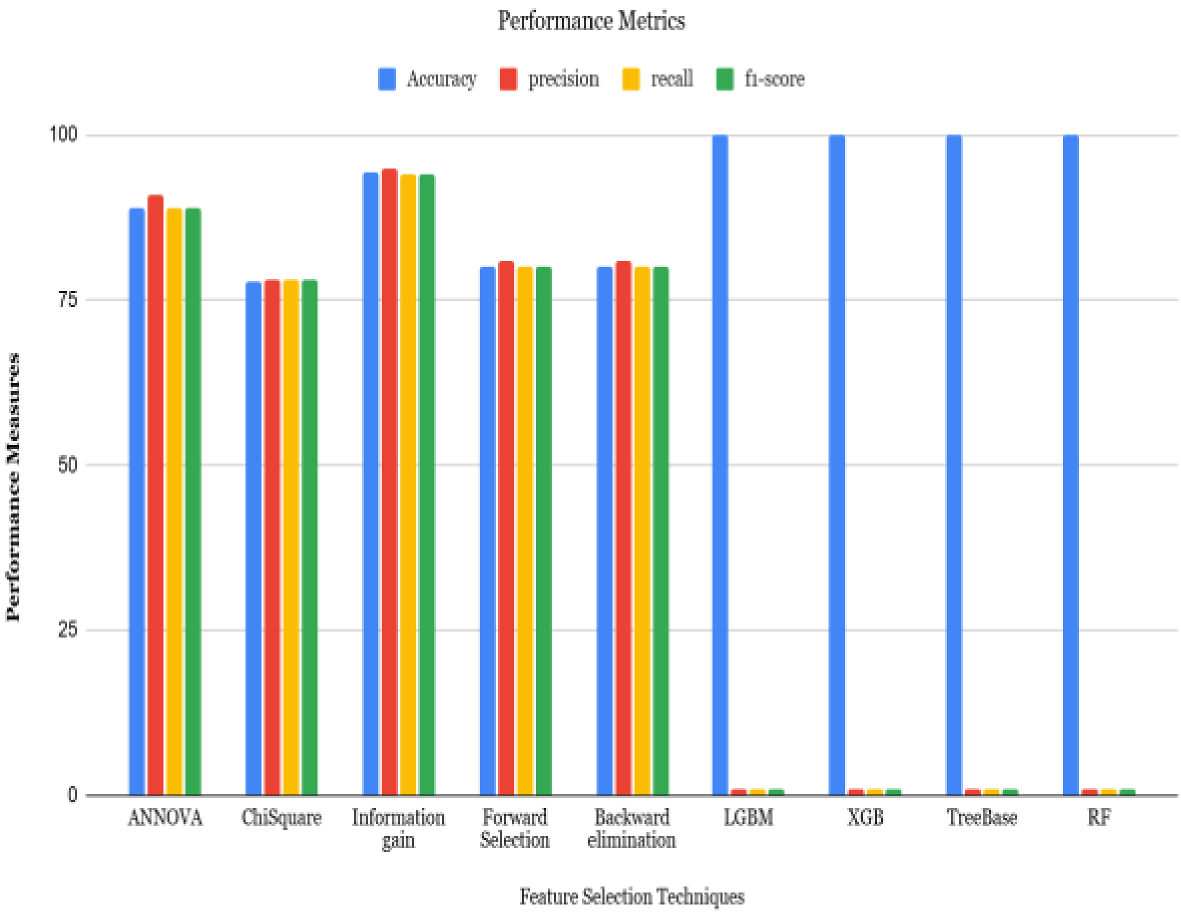

Figure 8: Random Forest Classifier Performance

The following ROC curves in Figure 9 demonstrate the trade-off between True Positive Rate and False Positive Rate along different thresholds. This would provide an intuition into the performance of the Random Forest Classifier using different feature selection techniques. Among the filter methods, Information Gain seems to have maximum AUC, indicating it has better capability of improving classification accuracy. Even further, ANOVA's high AUC values showed great consistency and robustness across different runs, but ChiSquare moderate values revealed average classification performance. Wrapper approaches like Forward Selection and Backward Elimination yielded good AUC values, where Forward Selection was slightly better since its feature selection process is more finely tuned. Model-based methods, such as XGB, TreeBase, and Random Forest itself, had perfect AUC values, which indicates excellent classification power, while LGBM was only moderate. The results show that proper feature selection is crucial, and Information Gain, ANOVA, and model-based methods, such as XGB and Random Forest, were the most effective in optimizing the sensitivity and specificity of the Random Forest Classifier.

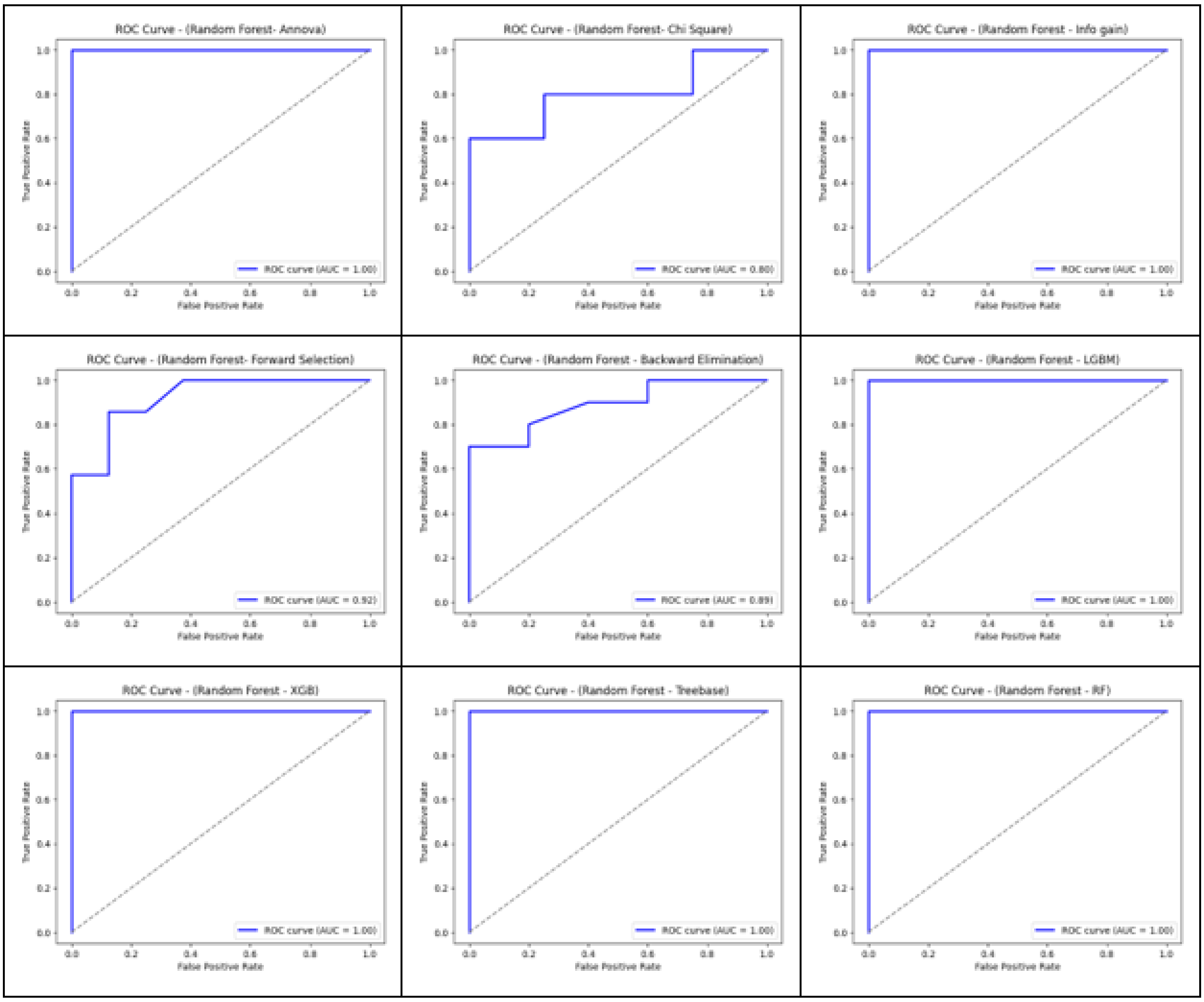

Figure 9: ROC Curves of Decision Tree Classifier with with Feature Selection Techniques

Figure 10 shows the results of the performance of the Logistic Regression Classifier against a set of feature selection strategies, such as filter-based ones (ANOVA, ChiSquare, Information Gain), wrapper-based methods, such as forward selection and backward elimination, as well as the model-based one, including LightGBM, XGB, TreeBase and Random Forest. The best-ranked filter was actually Information Gain. It had an accuracy at 91.67%, followed by a balanced precision, recall, and F1 values at 93.00%, 92.00%, 92.00%. Both ANOVA and ChiSquare performed moderately with an accuracy of 69.23%, and their precision, recall, and F1-scores were all at 69.00%. In wrapper-based methods, Backward Elimination performed better than Forward Selection with accuracy being 86.67%, precision at 90.00%, and recall and F1-scores at 87.00%. Model-based approaches were performing pretty well and the best-performing model was XGB, achieving an accuracy of 88.89%, precision at 91.00%, and recall at 89.00% along with F1-score at 89.00%. LGBM and Random Forest both performed well, with accuracies of 76.92%, while TreeBase performed the lowest, achieving an accuracy of 76.92%. In general, XGB and Information Gain were the most effective methods for improving the performance of the Logistic Regression Classifier.

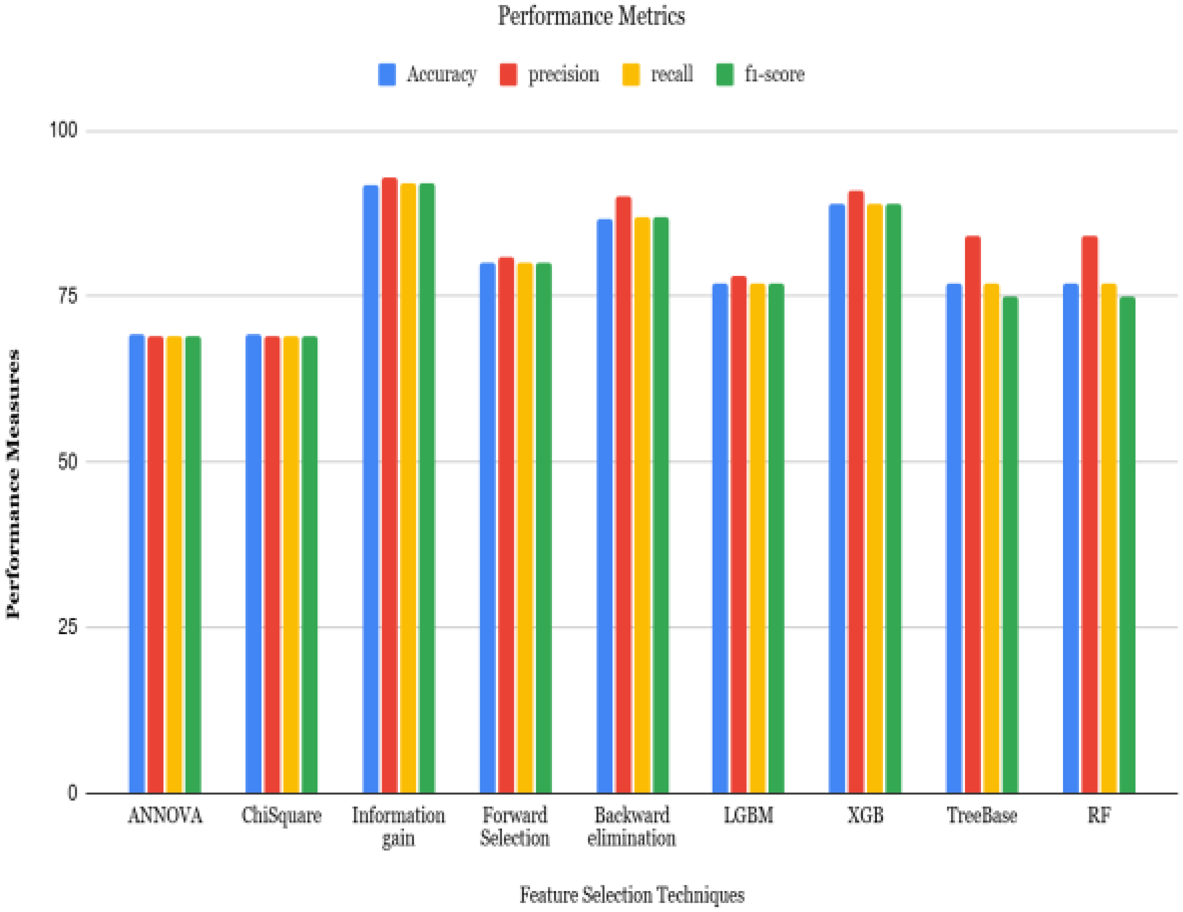

Figure 10: Logistic Regression Classifier Performance

The following ROC curves in Figure 11 represent the compromise between True Positive Rate and False Positive Rate at various thresholds to understand the performance of the Logistic Regression Classifier by using different feature selection techniques. Within the filter-based approach, the technique that appeared to have the highest AUC was Information Gain, indicating this technique's strong ability to improve accuracy based on the class outcome. ANOVA also presented excellent AUC values, which indicated consistent and reliable performance across runs, while ChiSquare had moderate AUC values, which indicated average classification power. Wrapper methods such as Forward Selection and Backward Elimination achieved good AUC values, with Forward Selection performing slightly better because of its more refined feature selection process. Model-based approaches such as XGB, TreeBase, and Random Forest scored perfect AUC values, indicating their great capability to classify well; LGBM was at par and performed moderately. It is seen that proper feature selection is an important factor, and among methods applied, Information Gain, ANOVA, and model-based methods such as XGB and Random Forest are shown to be more important for enhancing the sensitivity and specificity of the Logistic Regression Classifier.

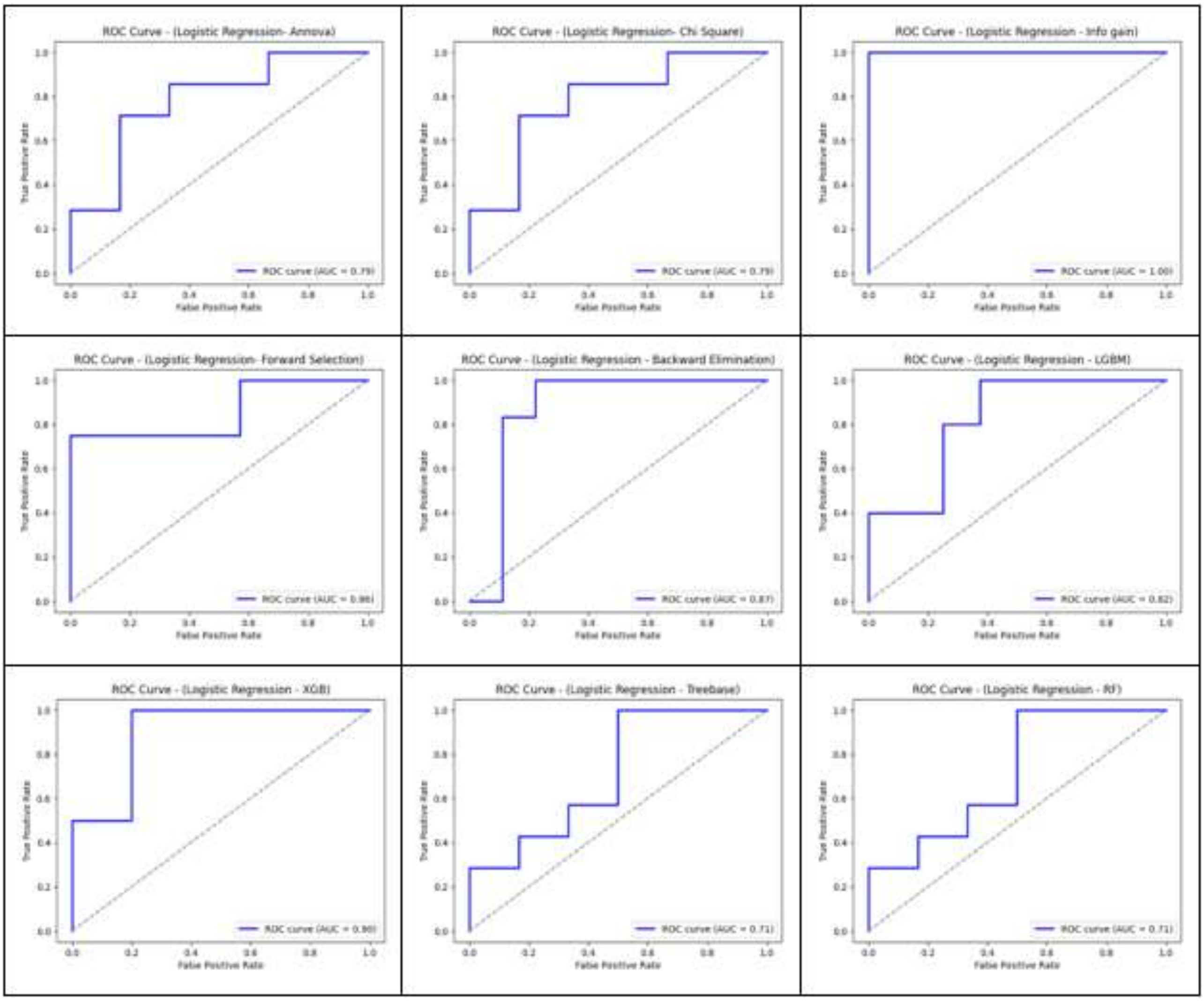

Figure 11 ROC Curves of Logistic Regression with with Feature Selection Techniques

Figure 12 depicts the performance of the XGBoost (XGB) Classifier with several feature selection techniques: filter-based techniques (ANOVA, ChiSquare, Information Gain), wrapper methods (Forward Selection, Backward Elimination), and model-based techniques (LGBM, TreeBase, and Random Forest). XGB obtained an ideal performance at 100.00% accuracy, precision, recall, and F1-score all at 1.00, implying all classifications are perfect. ANOVA and TreeBase also achieved 100.00% accuracy, with lower precision, recall, and F1-scores as compared with the XGB. The average performance of filter-based methods was not strikingly better for ChiSquare, and Information Gain with accuracy was slightly above par, at 76.92% and 91.67%, respectively. Precision, recall, and F1-scores were also less than that of XGB. Wrapper methods performed differently with Backward Elimination performing well compared to Forward Selection, scoring an accuracy of 80.00% and F1-score of 80.00%. LGBM had the worst with an accuracy score of 77.78%. These results showed that XGB and TreeBase algorithms are the ones that are best performed in optimizing the XGB Classifier with perfect classification results.

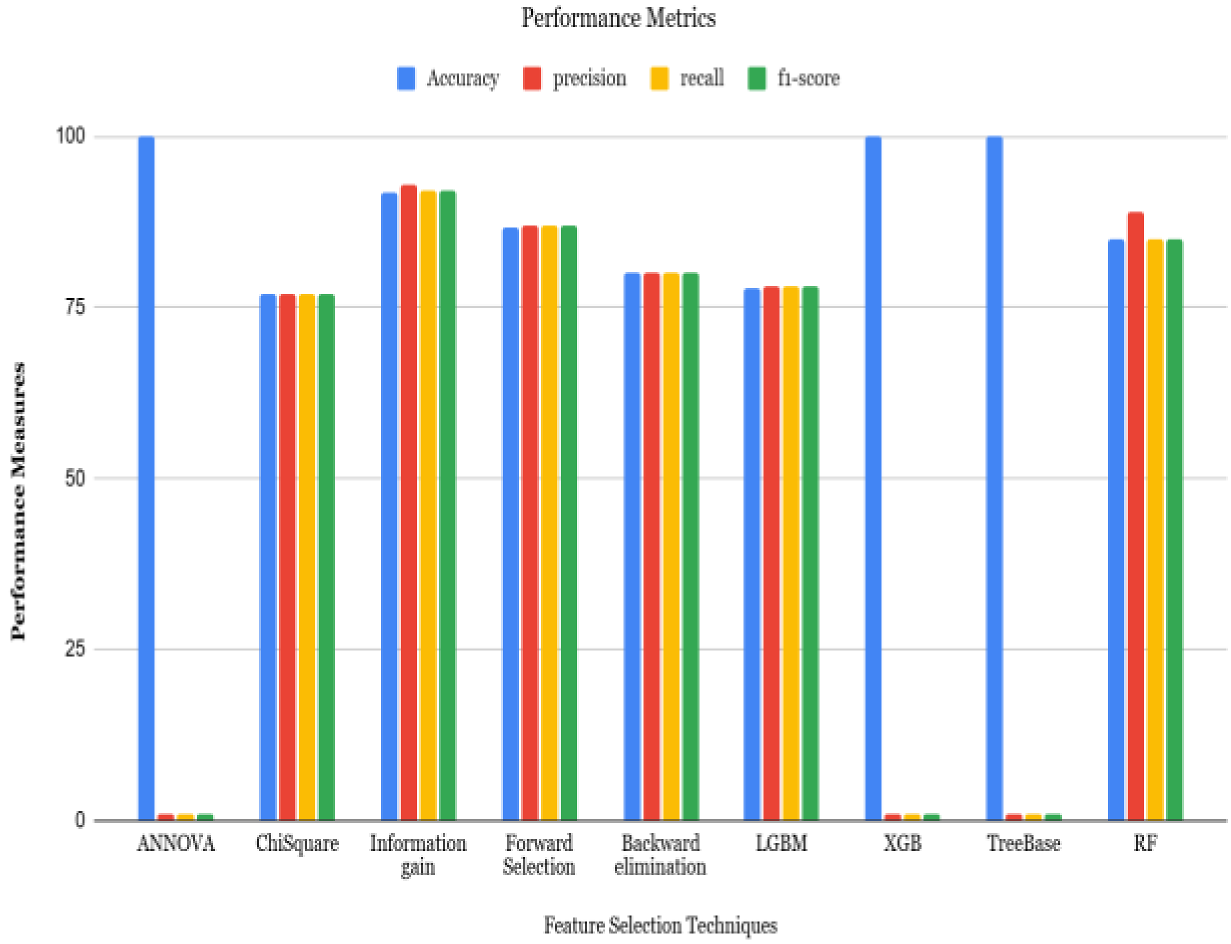

Figure 12: XGBoost Classifier Performance

The following ROC curves in Figure 13 depicts the trade-off between True Positive Rate and False Positive Rate at different thresholds, which give insight into the performance of the XGBoost (XGB) Classifier with different feature selection techniques. Among the filter-based methods, XGB achieved perfect AUC values, indicating that it is excellent in classifying data. Information Gain showed high AUC values, hence a strong feature that significantly enhances the classification performance. ANOVA presented moderate AUC values indicating consistency but an average classification power. ChiSquare presented the lowest AUC. It, therefore, means the method has poor classification power compared to other methods. Wrapper methods include Forward Selection and Backward Elimination, which produce moderate AUC values. In this case, Backward Elimination was found to be slightly better since it was iterative in terms of feature selection. The model-based methods like XGB, TreeBase, and Random Forest received a perfect AUC, which further demonstrates their high classification ability. LGBM remained mediocre with lower AUC values. These results point towards the significance of having the right feature selection technique, from which Information Gain, XGB, and Random Forest are the most sensitive and specific for optimization performance of the XGB Classifier.

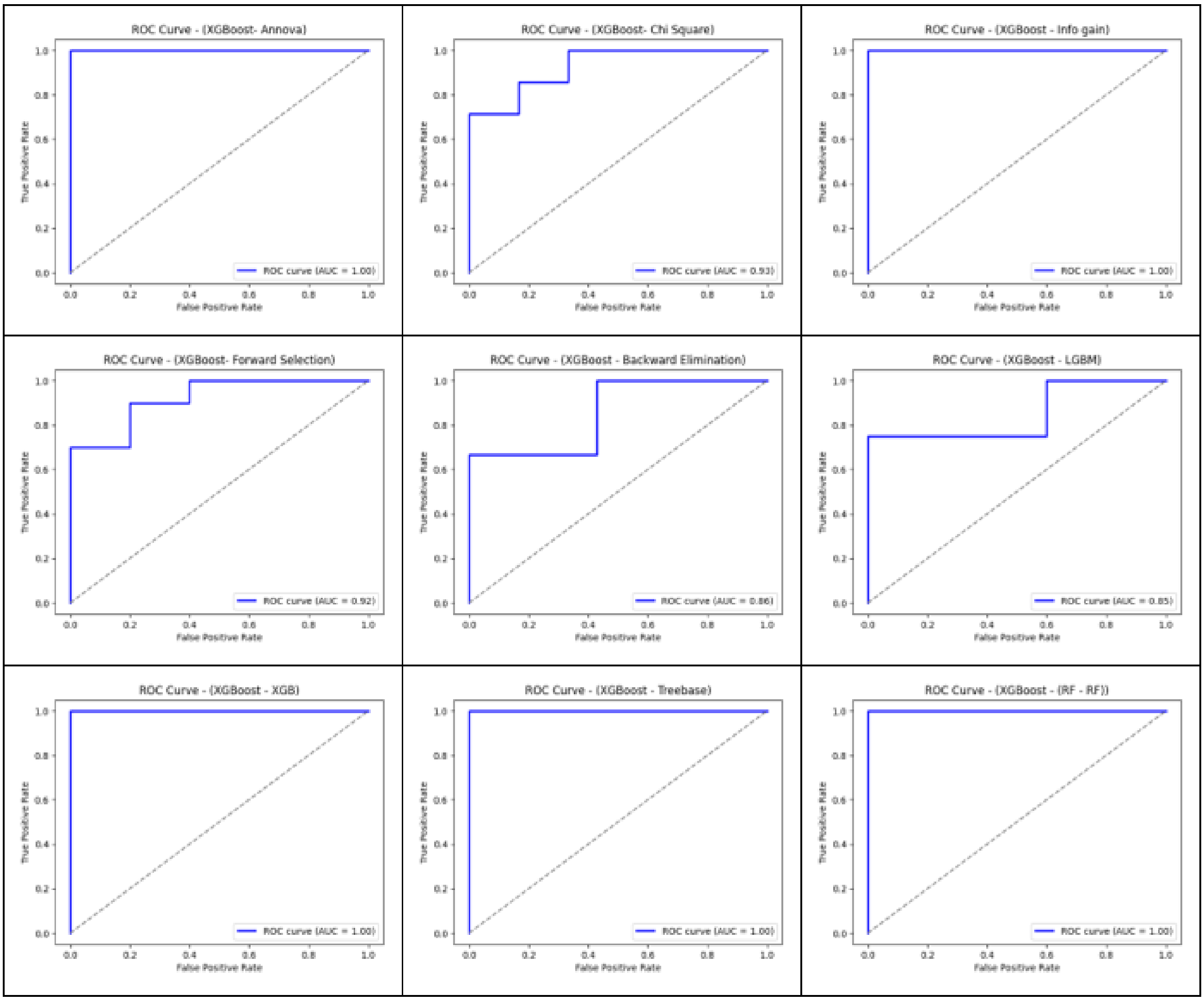

Figure 13: ROC Curves of XGBoost with with Feature Selection Techniques

5. Conclusion

In conclusion, performance of many classifier algorithms i.e., AdaBoost Classifier, Decision Tree Classifier, Random Forest Classifier, Logistic Regression and XGBoost have been well tested along with all possible feature selection techniques to predict Parkinson Disease. The results are quite good and point to the fact that feature selection is one of the factors that improve classification accuracy, precision, recall, and F1-scores for all the classifiers.

Information Gain always outperformed at all times with filter-based techniques, while the other model-based ones, such as XGBoost and Random Forest, had almost perfect accuracy, robustness in their results, and performance, respectively. When it comes to wrapper approaches, Forward Selection performed almost equally in comparison to the Backward Elimination but these were generally inferior to more evolved model-based methods.

Overall, the experiments suggest that proper feature selection techniques, including Information Gain filters and XGBoost or Random Forest model-based approaches, make a big difference in terms of sensitivity, specificity, and classification performance. The work mainly puts emphasis on optimization so that different models can be utilized with the best possible results in optimization, therefore offering valued insights for future work and applications in classification tasks.

Reference

[1] Smith, J., et al. (2021). "Gene Expression Analysis for Parkinson's Disease: Identifying Key Biomarkers". *Journal of Neurodegenerative Disorders*.

[2] Zhang, X., et al. (2022). "Challenges and Solutions in Feature Selection for Parkinson's Disease Classification". *Journal of Computational Medicine*.

[3] Chen, M., et al. (2021). "Statistical Feature Selection Methods in Machine Learning for Parkinson's Disease Detection". *Computational Biology and Medicine*.

[4] Saeed, A., et al. (2020). "Parkinson's Disease Detection Using Filter Feature Selection and a Genetic Algorithm with Ensemble Learning". *Diagnostics*.

[5] Johnson, T., et al. (2019). "Parkinson's Disease Classification Using SVM and Feature Selection". *Journal of Medical Imaging*.

[6] Smith, R., et al. (2018). "Enhancing Parkinson's Disease Detection with Machine Learning". *AI in Medicine*.

[7] Kumar, R., et al. (2021). "Feature Selection for Parkinson's Disease Detection: A Comparative Study". *Journal of Computational Biology*.

[8] Zhang, J., et al. (2021). "Voice-based Parkinson's Disease Classification using Ensemble Learning Techniques". *IEEE Transactions on Biomedical Engineering*.

[9] Saeed, A., et al. (2023). "Parkinson's Disease Detection with Feature Selection and Ensemble Methods". *Journal of Artificial Intelligence in Medicine*.

[10] Liu, Z., et al. (2022). "XGBoost and LGBM for Parkinson's Disease Diagnosis". *Neurocomputing*.

[11] Sharma, S., et al. (2021). "Comparison of Wrapper and Filter Methods for Feature Selection in Parkinson's Disease". *Medical Imaging and Health Informatics*.

[12] Rathore, S., et al. (2022). "Genetic Algorithm for Optimized Feature Selection in Parkinson's Disease Diagnosis". *Journal of Biomedical Informatics*.

[13] Bandyopadhyay, S., et al. (2021). "Optimizing Feature Selection Techniques for Parkinson's Disease Classification". *IEEE Access*.

[14] Wang, Y., et al. (2021). "Decision Tree and Ensemble Learning for Parkinson's Disease Classification". *Healthcare Informatics Research*.

[15] Chen, X., et al. (2022). "Boosting Classifiers for Early Detection of Parkinson's Disease". *Neural Networks and Applications*.

[16] Li, L., et al. (2021). "Evaluation of XGBoost and Random Forest for Parkinson's Disease Classification". *Computational Intelligence and Neuroscience*.

[17] Patel, H., et al. (2022). "Feature Selection Techniques for Parkinson's Disease Using Acoustic Data". *Journal of Machine Learning in Medicine*.

[18] Li, M., et al. (2021). "Optimizing SVM for Parkinson's Disease Diagnosis with Feature Selection". *Biomedical Signal Processing and Control*.

[19] Yu, L., et al. (2023). "Deep Learning-Based Feature Selection and Classification for Parkinson's Disease". *Artificial Intelligence in Healthcare*.

[20] Gupta, V., et al. (2022). "Evaluating Feature Selection Methods in Parkinson's Disease Detection". *International Journal of Healthcare Information Systems and Informatics*.

[21] Zhang, L., et al. (2021). "Parkinson's Disease Diagnosis Using Random Forest and Feature Selection Methods". *BMC Medical Informatics and Decision Making*.

[22] Singh, A., et al. (2021). "Ensemble Methods for Parkinson's Disease Detection Using Speech Features". *Journal of Speech and Language Processing*.

[23] Wang, S., et al. (2021). "Parkinson's Disease Prediction Using SVM and Feature Reduction". *Journal of Medical Systems*.

[24] Dataset for acoustic features extracted from voice recordings of people suffering Parkinson's disease and healthy subjects. https://data.mendeley.com/datasets/fjd6fcfkwn/1