Performance Evaluation of the Filter, Wrapper, Mutual Information Theory, and Machine Learning Feature Selection Methods for XGBoost-Based Classification Tasks

Jasim Mohammed Dahr[1] and Alaa Sahl Gaafar[2]
[1,2]Directorate of Education in Basrah, Basrah, Iraq
E-mail: Jmd20586@gmail.com, alaasy.2040@gmail.com

Keywords: statistical analysis, relationships, features selection, effectiveness, data mining, linear, nonlinear, machine learning, factors ranking

Received: February 9, 2025

Feature selection is a model for mining datasets to obtain and choose sensible and meaningful parameters and values required for building high-performance classification or regression tasks. Even more worthy of note is the fact that relevance, interactions of features, and reduction of noise and redundancy through the use of associations with ground truth values. The concept of feature selection is most appreciated for large size and complex datasets in which a set of attributes and matching values as contributing significantly to the determination of decisions made by machines or human agents. This paper compares the performances of machine learning algorithms, wrapper, filter and mutual information methods for features selection in data. The Diabetes dataset acquired from the Pima Indians Diabetes Database hosted by the National Institute of Diabetes and Digestive and Kidney Diseases was adopted for the validation. The outcomes revealed that, the XGBoost model classification's accuracy of 75.76%, precision of 64.63%, and F1-score of 65.43% were best due to others. Also, the mutual information theory or embedded technique offers the best recall score of 71.25% trailed by the filter technique. The mutual information provided the least false-positive of 23 followed by the filter technique at 27. The filter technique outcomes with two-tailed significance test score of p(0.059)<0.05), which are statistically significant at confidence value of 95%. Also, the filter feature selection technique further reduces the dimensionality, redundancy in the variables, and maintain the data variance. Moreso, the overfitting of model is minimized, but raising degree of freedom of the base model during classification tasks.

Povzetek: Narejena je primerjava zmogljivost filtrov, ovijalnih metod, teorije medsebojne informacije in strojnega učenja za izbiro značilk, ki pokaže, da filter in informacijska teorija izboljšujeta klasifikacijo XGBoost.

1 Introduction

The progress of the past 10 years enabled various digital devices to generate data that poses problems during analysis and representation. The high-dimensionality and volume of datasets further added to the complexities in data sciences and data mining. Even more problematic are the weaknesses of traditional approaches to address low accuracy, scarce memory, high costs of executions, and inability to select the significant features representative of entire data [1], [2].

Feature selection is an intelligent algorithm built on machine learning for the choosing data subset that produces optimal outcome [3], [4][5]. The purpose of feature selection is to minimize the noise and redundancy in data, which enhance the outcomes of classification and regression tasks [6]. Some common feature selection approaches are categorized under wrapper, embedded, filter and ensemble models for numerous features selection of mapping soil organic matter of complex geospatial-data [7]. According to [8], the process of optimizing decision-making during clinical diagnosis and prognosis are most desirable especially at time emergencies. The characteristics of patients and clinic presentations reinforce detection of critical and less-critical suffers like COVID-19. Many scholars turned to artificial intelligence and machine learning techniques to attain low-dimensionality and geometric features conservation in the raw datasets.

The concept of key performance indicators soft sensing was introduced as impacting significantly on the decision-making processes of the large industries. The soft sensors data-propelled with machine learning models are leading pact, though, many of underlying features of industrial data are less adaptive and noise-full, giving rise to the need for their selection strategies. Support vector regression, partial least-squares, optimized sparse partial least squares and multi-output least-squares support vector regression are commonly used to mine the features to determine quality of product [9]. The authors in [10] have applied autonomous feature selection technique (EAFS) with an ensemble model to reveal intrusion within network flows. The authors in [11] identified a solution to perennial problem of cancer detection and treatment in which selection of Gene from micro-array data due to its high-dimensionality. Equally, the majority of cancer data bound

for practice and research are laced with noise and low instances. Therefore, individual and mix methods for feature selections that accounted for highly relevant and informative genes are adopted for the classification. To this end, noisy data and irrelevant genes elimination could improve cancer detection. The fundamentals of feature selection categories and their matching techniques include:

Filter feature selection uses scores to minimize the highly dimensional datasets, less-computationally intensive, speedy, and achieves higher generality. Most commonly filter methods include:

Information gain select features in datasets through entropy and amount of information in random variable. Mutual information is tailored towards non-linear associations between two random variables. Conditional mutual information maximization selects features using an approximation measure that reduce the correlation between genes and features.

Minimum redundancy maximum relevance entails used of mutual information algorithm by computing the maximum relevance between features and target and minimum redundancy between two random variables. Random forest ranking takes advantage of decision trees to combine forecasts across a set of random trees and accuracy-based ranking. Fast correlation-based filter is most applicable to multivariate criteria for minimizing noise and retaining highly relevant data through symmetric uncertainty method.

Fisher's scoring algorithm or F-score is the selection technique based on F-distribution to choose each unique feature matching the target feature, that is, the smallest distance in terms of minimum interclass distance and maximum intraclass distance. Relief algorithm adopts correlation score between features, and their weights to choose the features to prediction. Outcomes are most impacted by noise and errors in the features subsets. Pareto optimization is a multi-objective optimization to create and reveal the collection of best acceptable solution from the solutions space by the decision-maker.

Wrapper feature selection uses classifier and machine learning algorithms to seek out optimal features subsets from the raw search space of features. It is most desirable for ranking of features for moderate-dimensional datasets, though, computing resource-intensive. The categories are discussed as follows:

Evolutionary-based technique is propelled by nature's evolutionary process: Flower pollination algorithm, genetic algorithm. Swarm-based technique realized from the animals' social behaviours: Artificial bee colony algorithm, cuckoo search, dragonfly algorithm, moth flame algorithm, particle swarm optimization, firefly algorithm, bat algorithm, and ant colony optimization, Human-based technique taken after activities and behaviour of humans: Teaching-learning-based algorithm, and learning automata. Physics-based technique following from physical processes of the nature: Black hole algorithm, and gravitational search algorithm. Music-based technique motivated from music instrument: Harmony search. Others: Crossover, and stacked autoencoder.

Also, the principal component analysis and model-based feature selections had recently been applied in hyperspectral image prediction tasks [12]. Information retrieval is a developing aspect of computer science in discovering important information form piles of the same. Natural language processing offers numerous relationships between textual data [13]. Machine learning models are holding sway in personalized disease risk forecasts such as early screening and treatment using single nucleotide polymorphisms associations [14]. The key contributions of this paper include:

i. To extract feature importance using the feature selection methods on the textual dataset.
ii. To determine the effectiveness of the feature selection methods using XGBoost model as base estimator.
iii. To compare the outcomes of XGBoost model based on the selected features using standard metrics.

The organization of the paper includes: section two is the related studies. Section 3 is the methodology. Section 4 is the results and discussion. The last section is the conclusion.

2 Related studies

In Arabic text classification problem, an improved Chi Square feature selection was applied on common classifiers for Arabic texts classification by [13]. Considering measuring metrics of precision, F-score, runtime, and recall, the performances of the classifiers were better after feature selection process. With pre-processing and improved CHI feature extraction, the RF model achieved precision of 0.958. NB obtained recall of 0.966. F1-score of 0.960 was attained by NBM model. Though, more optimization algorithms could augment feature selection methods performances. Other forms of dataset can be investigated.

The authors in [15] experimented the cause-effect of wrapper, filter, and embedded feature selection techniques on Random Forest, Artificial Neural Network and Gradient Boosting Machine. The outcomes revealed that, mutual information generated features, and the Gradient Boost Machine predictions for the wastewater components were the most superior using R square and mean square error. The GBM attained the biggest performance by R^2 of 0.58, RMSE of 0.092, and MAE of 0.017. However, the GBM is less sensitive to changes in features set when compared to RF and ANN models.

In [3], the authors highlighted in a survey about swarm intelligent feature selection technique in which Ant algorithm speedups execution, lower costs, and generates better results. But it lags behind in terms of accuracy against Cuckoo search, genetic algorithm, and Bees colony. The feature selection increased the detection rate of healthy and cancer patients through expression of gene profile by accuracy of 99.13% using enhanced Cuckoo search in 29 features against accuracy of 98.75% for original 34 features when applied on digital image of breast cancer dataset.

The author in [12] adopted an enhanced binary equilibrium optimizer algorithm to raise the feature

selection tasks and minimize the dimension of the datasets. The KNN and SVM classifiers, wrapper methods, were used to counter the over-fitting challenge. The outcomes after statistical analysis were better with the introduced method. Using the vegetation indices, and principal component analysis feature selection approaches to improve the prediction of plant water content using partial least square regression, random forest, and backpropagation neural network. Consequently, the hybrid feature selection of model-based features, and principal component analysis raised the performance of the enhanced backpropagation neural network model (BPNN-PCA-MF) by R2 of 0.998, and RMSE of 25.20% respectively. Future works could extend the model to others areas of applications.

In [16], they experimented mutual information, univariate ROC-AUC, and fisher score filter feature selection methods with thirty-two machine learning approaches. Authors leveraged on multiple correlation approach to disband analogous features, reduce dimensionality, and increase the performance significantly in terms of accuracy of 26.50%, F1-score of 70.90%, and ROC of 26.74%. The mutual information technique with random forest correlation having the strongest influence on the outcomes.

In [17], they developed a general learning equilibrium optimizer (GLEO) founded on the wrapper feature selection strategy for diseases diagnosis. Authors utilised the new strategy to uncover a collection of 9 important biological features hidden in the volume of 16 attributes. The outcomes were outstanding based on the new feature selection procedure with regards to accuracy, fitness value, and size of feature. The informative features selection increases accuracy of biological classification jobs in which the GLEO applied on Leukemia dataset achieved accuracy of 0.9986. Future studies could investigate other metaheuristic algorithms and their hybridizations schemes.

The authors in [4] conducted feature selection on large-size DDoS attacks on 5G core network infrastructure using machine learning method to decimate time, noise and latency. Authors leveraged on the feature selection to ride-off noisy features from the data which increase the accuracy of the detection. The Stacking based feature selection produced accuracy of 97.183%, precision of 97.442%, F1-score of 97.065%, recall of 96.724%, and AUC of 98.075% for reduced features of 10 of DDoS intrusion dataset. Whereas, only the filter, wrapper, and ML model were investigated.

The authors in [18] noted that, radiomics technique is a common quantitative features extraction for medical imagery. These distribution of features raises the accuracy of the prognostics and classification of diseases at the early stages like lung cancer. Though, further refinement of radiomics including consistency, cohort power, harmonization of feature value, etc., to increase stratification of risk, and management of patients. But the study did not provide empirical explanations about the optimal technique of the multiple features selection approaches.

In [10], the authors designed an ensemble autonomous feature selection known as EAFS for intrusion detection tasks. The EAFS computed the feature importance, which were used up by normalized score of mixed (NSOM) to create the subset from the volume on basis of high score. Upon validation, the classification accuracy of the ensemble model at 99.34% was better with the UNSW-NV15 datasets. Similarly, false alarm rate of 1.69% was observed.

The authors in [14] discussed the advantage of undertaking feature selection that extract most informative feature within the noisy non-informative, redundant, and irrelevant features. The concept targets genetic sequencing for improving precision medicine and patients' healthcare through distinct genetic makeups. The multivariate filter methods (chi square test) were used to explain the relationship between single nucleotide polymorphisms known as epistasis effects. In this way, complex diseases were detected by eliminating genetic features associated with insignificant heritability of individuals. However, there is no empirical results from study.

In [7], the authors utilised four feature selection techniques for curating optimum parameter subsets within initial data for mapping soil organic matter in multifaced forest landscape of Southern China. Authors leverage on wrapper, embedded, filter and ensemble models to create optimum variable of the datasets. The outcomes indicated that, XGBoost was superlative ensemble performer than wrapper and filter models using root mean square error (RMSE) and correlation coefficient (R2). The RRA-based ensemble model performance improved by R2 of 5.77% and RMSE of 9.16%. The reason for this trend can be explain by the use of multiple selectors and robust rank aggregation procedure.

The authors in [8] attempted to apply machine learning in finding evidence of COVID-19 from sets of indicators within blood tests of suffers. The idea was to reduce poor morbidity and prognosis of COVID-19, which achieved through the use of matrix factorization and Random Forest algorithm. The former reduced the high-dimensionality in blood biomarker space of COVID-19 patients in which Arterial Blood Gas O2 saturation and C-reactive Protein are held high as important features to prognosis. There are prospects of using ML for clinical and pathogenesis of COVID-19 but, there could be loss of features.

In [9], they developed a causal based on autonomous selection of features of the soft sensing in key performance indicators. This technique is motivated by non-linear causal model and information theory to address the interleave between original industrial dataset's key performance indicators and every feature importance. The AdaBoost ensemble method constructed the soft sensors' key performance indicators after a non-zero causal-and-effect autonomous features selection, which realized RMSE of 2.215, and R2 of 61.1. Though, it was discovered that the ML models are weak for dataset containing causal features.

The authors in [19] utilised ranker-based feature selection method with 9 machine learning to correctly classify eye diseases. The outcomes revealed that, the SVM outpaced with accuracy of 99.11%, followed by LR at 98.58%. A

nested ensemble selection method comprising wrapper and filter feature selection techniques for precision, efficiency and simplicity by [20]. This approach was capable of extracting relevant variables from irrelevant, and correlated from redundant features. Using the real-life and synthetic datasets, the nested ensemble selection attained relatively higher accuracy even for multiclass cases.

In [21], they developed an ensemble feature selection method for combining the effectiveness of single feature selection algorithms such as correlations between forest aboveground biomass and features, and the multicollinearity of selected features. The stability heterogeneity correlation-based ensemble method ranked and generated subsets of important features for high accuracy of prediction. The XGBoost model generated better R2 of 0.61 and RMSE of 15.32 against RF model at R2 of 0.57 and RMSE of 16.16 for 26 of 46 features extracted.

In [22], the authors leveraged on correlated and frequent items, a text feature selection for subsequent classification tasks. The proposed strategy account for importance and feature interactions with association to determine the interrelationship between features and target class. The validation process adopted SMS spam dataset that produced 95.12% accuracy of text classification after optimal features of 6%.

Summary of the key related studies including author(s), objective(s), techniques, results, and weaknesses are presented in Table 1.

Table 1: Summary of the key related studies

Author(s)	Objective(s)	Techniques	Results	Weaknesses
Alshaer et al. (2021)	Arabic text classification.	Feature extractor: Chi-Square. Estimator: NBM.	Accuracy: 0.966. F1-score: 0.960.	More datasets and optimizers required.
Bagherzadeh et al. (2021)	Cause-effect wastewater.	Wrapper, filter and ML.	GBM best features selector by R^2: 0.58. RMSE: 0.0092, MAE: 0.017.	GBM is less adaptive to changes in data features.
Jameel & Abdullah (2021)	Cancer gene profiling and diagnosis.	ML: Cuckoo Search, Genetic algorithm and Bees Colony.	Accuracy: 99.13% for 29 features against 98.75% for original 34 features.	Digital dataset only.
Too & Mirjalili (2021)	Diseases diagnosis.	Wrapper: GLEO applied on Leukemia dataset.	Accuracy: 99.86%.	More metaheuristic algorithms can be investigated.
Jain & Saha (2022)	Data attributes extraction.	Mutual information, ML, and filter.	Accuracy: 26.50%. F1-score: 70.90%. ROC: 26.74% for mutual information technique.	Low performance of mutual information technique.
Zhang et al. (2022)	Intrusion detection tasks.	ML: Ensemble autonomous features selection.	Accuracy: 99.34%, false positive: 1.69% with UNSW-NV15 dataset.	Only ML feature selected was investigated.
Saberi-Movahed et al. (2022)	COVID-19 evidence in blood sample.	ML: matrix factorization. and random forest.	Dimensionality reduction. Importance feature scores	No-textual dataset.
Sun et al. (2023)	Causal based autonomous features selection in soft sensing.	Information theory model.	No empirical results.	More data complexity.
Kamalov et al. (2023)	Relevant features selection.	Nested ensemble approach with filter and wrapper.	No empirical results.	More complexity in dataset.
Marouf et al. (2023)	Eye disease features classification.	ML approaches such as SVM and LR.	SVM: Accuracy of 99.11%. LR: 98.58%.	Technique limited to ML.
Zhang et al. (2023)	Text feature selection.	Forest aboveground biomass and multicollinearity of selected features.	No empirical outcomes.	Domain non-specific.
Mamdouh & El-Hafeez (2023)	Text features selection in SMS spam dataset.	Importance and interaction between features and target class.	Accuracy: 95.12%.	Unknown technique.

From Table 1, the majority of the studies had identified the propensity of feature selection based on various methods classification tasks with the sole goal of denoising and redundancy removal using information theory, and machine learning approaches. Most prominent among the techniques are information theory, ML, wrapper, filters, and ensemble methods.

3 Research methodology

This paper investigates the performances of features selection techniques based on machine learning and information theory. On the part of the machine learning, the XGBoost was selected due to its decision-making capability in which weak classifiers are augmented. While, Maximal Information Coefficient was elected from information theory for choosing features within datasets based on non-functional and functional associations among variables. The proposed features selection approach is shown in Figure 1.

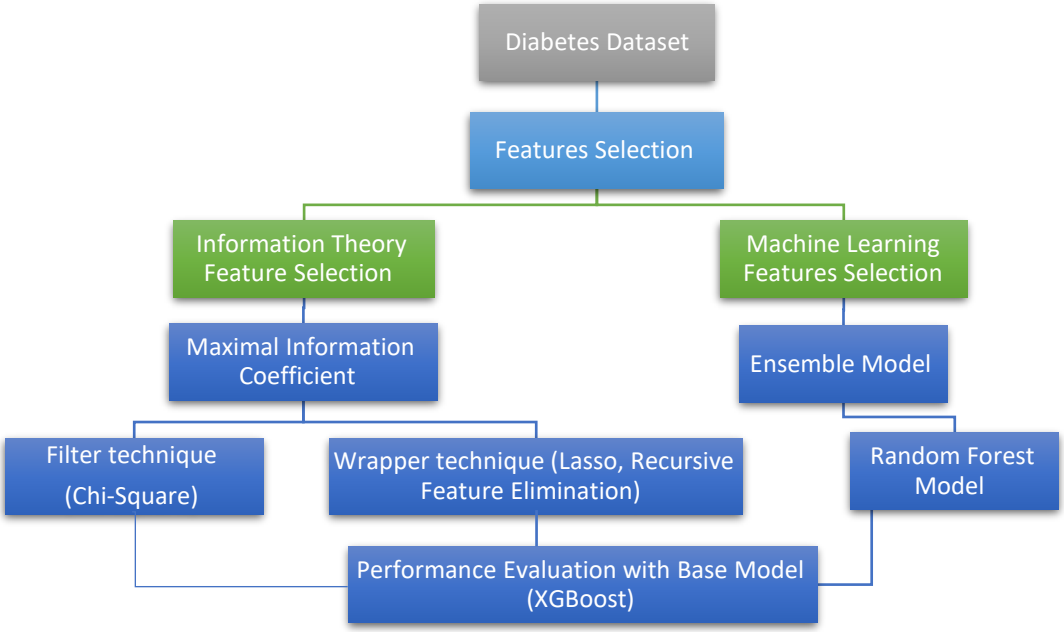

Figure 1: The layout of the features selection strategies performance comparisons study

From Figure 1, the main approaches considered for the Diabetes features selection include: wrapper (Lasso, Recursive feature elimination), filter (Chi-Square), and machine learning algorithm (XGBoost). The performance evaluation parameters related to the features selection technique are utilised for study. The key features selection approaches are discussed as follows:

Maximal information coefficient is an information theory-founded measure of association by accounting several non-functional and functional relationships existing within variables. Generally, it is equal to the coefficient of determination (R^2) as given by Equation 1.

$$R^2 = \frac{\sum(a-\bar{a})*(b-\bar{b})}{\sqrt{\sum(a-\bar{a})^2 * \sum(b-\bar{b})^2}} \quad (1)$$

Again, it uses values between 0 and 1, which explains 0 as statistical independence, and 1 as an entirely noiseless relationship. It can be represented by Equation 2.

$$m(a,b) = \frac{\max\{I(a,b)\}}{\log_2 \min\{n_x, n_y\}} \quad (2)$$

Where:
m = Maximal Information Coefficient,
a,b = random factors with minimum joint entropy normalization,
$\bar{a}$, $\bar{b}$ = mean score of the random factors
n = mutual information.
m explains the mutual information between variables a and b including generalization over a range and the equitability property.
m allocates similar score to evenly noisy associations without consideration of the form of association.
The Maximal Information Coefficient is capable of capturing different types of relationships, linear and non-linear (such as cubic, exponential, sinusoidal, superposition of functions). It leverages on mutual information, that is, symmetric nature. It removes any

assumptions concerning the variables distribution. The mutual information nature offers robustness to outliers. The coefficient computation is between 0 and 1 for better comparisons and understanding of outcomes. Though, it is incapable of depicting the direction or nature of the relationships. It is extensive in terms of computation. It is statistical less-effective against the Pearson's and distance correlation for independent data.

XGBoost is a renown gradient boosting framework which combines several decision trees. It is in the class of the ensemble models used for feature selection, compute the function with most influencing factors from the datasets, generate regression tree for understanding the importance of factors, higher Z-score values within the random probes are considered as importance factors, and produce the weights of regression trees for prediction accuracy. The ensemble model is a form regression model with larger learning abilities, and relatively smaller sensitivity for noisy datasets.

The process of ranking of factor selection by importance is conducted by mixing the Robust Aggregate rule. Given that the normalized rank factors of $y_{(1)} \leq \cdots \leq y_{(z)}, z = (1 \ldots i, \ldots z)$ signifies the ranking techniques. The rank vector $y_{(\grave{\iota})\leq} y_{(z)}$ as given by Equation 3.

$$R_{i,z}(Xt) = \sum_{f=j}^{z} \binom{z}{f} . Xt^{f} (-(Xt+1))^{z-f} \qquad (3)$$

$y_{(\grave{\iota})}$ is the defined as the ordering statistic for z unique factors shared over evenly between 0 and 1 for sorted sets, $R_{i,z}(Xt)$. The eventual score computed for the factors' ranks, y, whose p-value is given by Equation 4.

$$\partial(y) = \min_{i=1,\ldots,z} R_{i,z}(Xt) \qquad (4)$$

where, ∂(y) is p-value of computed factors' ranks. The value of $\partial \rightarrow 0$ explains a relative higher importance of the factors ranked distributed on the scale of 0 – 1 for the data distribution ranking importance for low variance of interclass and highest variance of the intraclass.

Data collection: The Pima Indians Diabetes Database is a collection originally maintained by the National Institute of Diabetes and Digestive and Kidney Diseases. It is composed of data of 768 women in Phoenix, Arizona, USA. The test's results showed that 258 tested positive and 500 tested negative. Again, there is one target (dependent) variable and the eight other factors including: Pregnancies, Oral Glucose Tolerance Test (OGTT), BloodPressure, SkinThickness, Insulin, Body Mass Index (BMI), Age, and PedigreeDiabetesFunction. The Pima population has been under study by the National Institute of Diabetes and Digestive and Kidney Diseases at intervals of 2 years since 1965. The Dataframe Class:0 = 6373, Class:1 = 1627.

Experimental parameters: The optimal requirements for the personal computer used for experimentation include:

Hardware: AMD E1-1200 APU Processor with RadeomTM Graphics 1.40 GHz, 64-bit Operating System, 4.00 GB RAM, x64-based processor.

Software: 3.5 Windows Experience Index, Windows 10 Single Language 2012.

Hyperparameters settings: The base model, XGBoost, hyperparameters values include: learning_rate (0.5), max_depth (3, 12), min_child_weight (1, 10), subsample (0.5, 1.0), colsample_bytree (0.5, 1.0), gamma (0, 5), and n_estimators (50, 500). Classifier = XGBoost, n_estimators = 100, max_depth = 3, n_jobs =- 1, random_state = 42, Metric: accuracy_score, f1_score, recall_score, precision_score, confusion_matrix, Classification function = Fitting, Data preprocessing function = Standard Scalar. The default values of the hyperparameters of Random Forest, Lasso model, and RFE were utilised.

4 Results and discussion

Lasso Model is a Mutual Information Theory technique for selecting and ranking the best features with dataset. Table 2 presents the outcomes of the applying Lasso model in selecting the best features within diabetes data which provides feature importance scores accordingly.

Table 2: Lasso model selected features with XGBoost model classification outcomes.

Data Feature	Importance Score	Rank
Glucose	0.346466	1
BMI	0.183006	2
Age	0.180830	3
DiabetePedigreeFunction	0.147554	4
Preganacies	0.142144	5

From Table 2, the index of the selected features are 'Pregnancies', 'Glucose', 'BMI', 'DiabetesPedigreeFunction', and 'Age'. Table 2. Lasso Model Feature Importance and ranking outcomes.

The graphical illustration of the Lasso model's feature importance of the sampled data is shown in Figure 2.

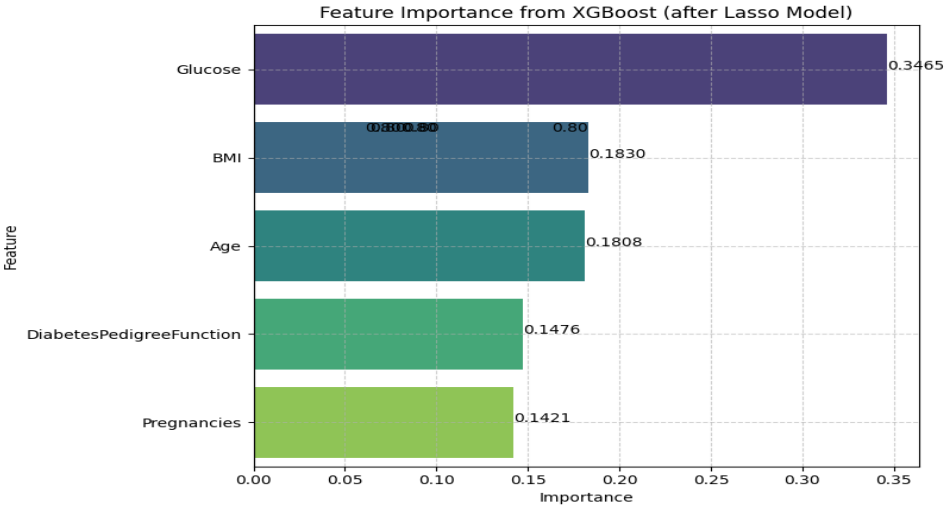

Figure 2: Lasso model feature importance and ranks representation.

In Figure 2, the feature importance as computed by the Lasso model utilised for XGBoost model-based decisions about risk of diabetes in women in Prima community of Arizona, USA indicated that, the first place is Glucose (34.65%), the 2nd placed important feature is BMI (18.30%), the 3rd placed Age (18.08%), the 4th placed feature is DiabetesPedigreeFunction (14.76%), and last placed important feature is Pregnancies (14.21%) respectively.

The XGBoost model with the selected features from Lasso model attained accuracy of 71.86%, recall of 71.25%, precision of 57.58%, and F1-score of 63.69%. Figure 3 illustrates the confusion matrix of the XGBoost model's classification outcomes with selected features through Lasso model. [109 42 23 57]

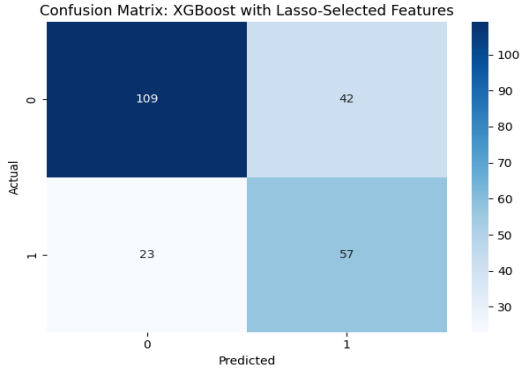

Figure 3: Confusion matrix of XGBoost model with Lasso model selected features.

From Figure 3, the XGBoost model correctly identified 57 breast cancer class of 109 original cases. Again, the XGBoost model identified wrongly 23 breast cancer class of the 42 unidentified cases.

The Wrapper Technique based on Recursive Feature Elimination calculates the best five features from the 8 original features are given by the index: ['Pregnancies', 'Glucose', 'Insulin', 'BMI', 'Age] in descending order of magnitude.

Performance of the RFE technique in selecting the best features achieved accuracy of 70.56%, recall of 61.25%, precision of 56.98%, and F1-score of 59.04%. The feature selection scores with RFE technique are presented in Table 3.

Table 3: The RFE feature selection importance scores and ranks.

Feature	**Importance Score**	**Rank**
Glucose	0.332591	1
BMI	0.193819	2
Age	0.191871	3
Insulin	0.151204	4
Pregnancies	0.130515	5

The graphical illustration of the feature importance from XGBoost model is shown in Figure 4. As shown, the Glucose, BMI and Age are the top three features from the dataset. While, the Pregancies and Insulin fall at the top 5.

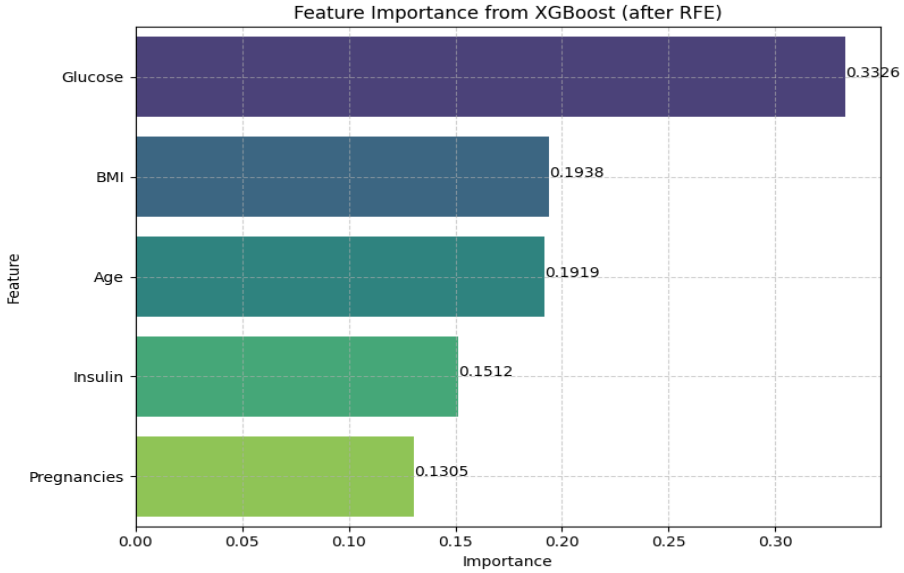

Figure 4: The Top5 important features after RFE with XGBoost model.

The confusion matrix of the reduced features from the diabetes as classified with XGBoost is given by Figure 5.

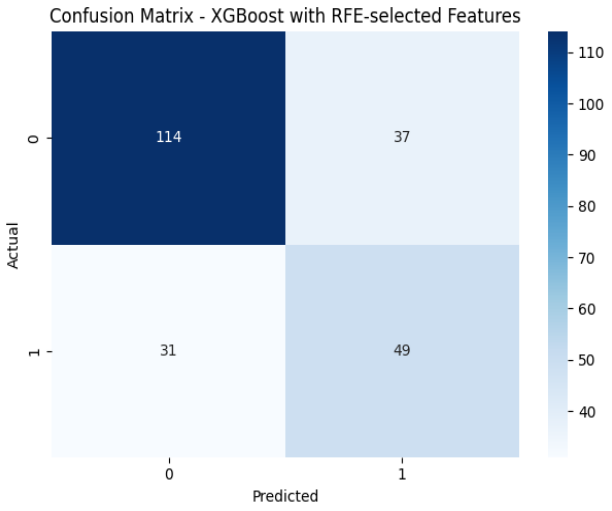

Figure 5: Confusion matrix of RFE and XGBoost model.

From Figure 5, the XGBoost model predicted correctly 49 breast cancer class of 114 original cases. Also, the XGBoost model predicted incorrectly 31 breast cancer class of the 37 unidentified cases.

The Filer Technique-based feature selection and ranking with Chi-Square scores computed for the features as shown in Table 4.

Table 4: The Chi-Square data feature importance score and rank computed.

Data Feature	Chi-Square Score	Rank
Age	8.205691	1
Glucose	7.094910	2
Pregnancies	6.559982	3
DiabetesPedigreeFunction	2.758584	4
Insulin	2.571590	5

From Table 4, the rankings of the data features with the Chi-Square scores are given as follows: Age (8.21), Glucose (7.09), Pregnancies (6.56), DiabetesPedigreeFunction (2.76), and Insulin (2.57). This shows that, Glucose is the topmost rank features when diagnosing diabetic patient. It implies that, Insulin and Age are the lowest and highest of the top-five features for diagnosing diabetic patients as shown in Figure 6.

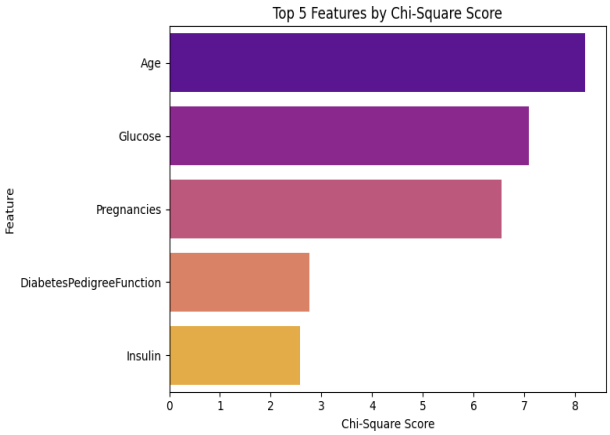

Figure 6: The 5 topmost features selected with Chi-Square Score.

The performance of the XGBoost model using the selected features from Chi-Square scores achieved accuracy of 75.76%, recall of 66.25%, precision of 64.63%, and F1-score of 65.43% respectively. The index of the selective features include: 'Pregnancies', 'Glucose', 'Insulin', 'DiabetesPedigreeFunction', and 'Age'. The confusion matrix of the Chi-Square Score with XGBoost model classification outcomes is shown in Figure 7.

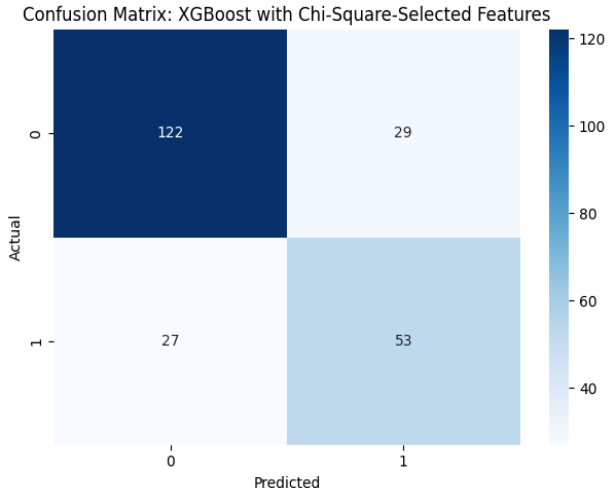

Figure 7: The confusion matrix of the Chi-Square scores with XGBoost model.

From Figure 7, the XGBoost model correctly predicted 53 breast cancer class of 122 original cases. Again, XGBoost model incorrectly predicted 27 breast cancer class of the 29 unidentified cases.

The feature importance scores computed by Random Forest algorithm and the corresponding rankings are provided in Table 5.

Table 5: The Random Forest algorithm's data features scores and ranks.

Data Feature	**Importance Score**	**Rank**
Glucose	0.282089	1
BMI	0.158120	2
Age	0.142116	3
DiabetesPedigreeFunction	0.113127	4
BloodPressure	0.084052	5

From Table 5, the Random Forest algorithm-based feature importance scores showed that, Glucose attribute was most important (28.21%), the 2nd placed important feature is BMI (15.81%), the 3rd placed important feature is Age (14.21%), the 4th placed important feature is DiabetesPedigreeFunction (11.31%), the least placed important feature is BloodPressure (8.41%) as depicted by Figure 8. The selected feature index includes: 'Glucose' 'BMI' 'Age' 'DiabetesPedigreeFunction' 'BloodPressure'

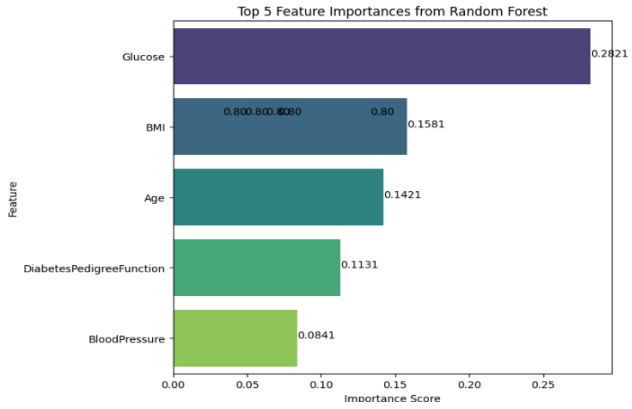

Figure 8: The Random Forest algorithm feature importance scores and ranks.

The outcomes of the classification tasks of the XGBoost model using importance features generated by Random Forest algorithm offered accuracy of 70.23%, F1-Score of 59.17%, recall of 62.50%, and precision of 56.18%. The confusion matrix of XGBoost mode after features selection with Random Forest algorithm is shown in Figure 9.

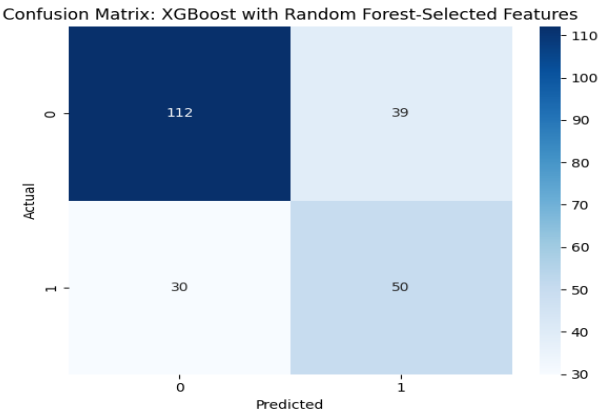

Figure 9. The confusion matrix of XGBoost model with Random Forest algorithm features.

From Figure 9, the XGBoost model rightly predicted 50 breast cancer class of 112 original cases. Again, XGBoost model incorrectly predicted 30 breast cancer class of the 39 unidentified cases.

4.1 Models comparisons

The comparisons of both approaches of features selection using the feature importance weights computed and classification metrics are presented in Table 6.

Table 6. Comparisons of feature selection models rankings.

Performance measure	Mutual Information Theory or Embedded Technique (Lasso model)	Wrapper (Recursive Feature Elimination)	Filter (Chi-Square)	Machine Learning (Random Forest)
Accuracy	71.86%	70.56%	75.76%	70.23%
Recall	71.25%	61.25%	66.25%	62.50%
Precision	57.58%	56.98%	64.63%	56.18%
F1-score	63.69%	59.04%	65.43%	59.17%
Confusion matrix	[109 42 23 57]	[114 37 31 49]	[122 29 27 53]	[112 39 30 50]

From Table 6, the best feature selection technique is the filter technique the XGBoost model at accuracy of 75.76%, precision of 64.63%, and F1-score of 65.43%. The mutual information theory or embedded technique offers the best recall score of 71.25% trailed by the filter technique. The mutual information provided the least false-positive of 23 followed by the filter technique at 27. The capacity of the filter and mutual information techniques in raising the classification outcomes of the XGBoost model when applied for computing the most importance factors for reaching the diagnosis diabetes among women in Pima India community in Phoenix, Arizona, USA. The reasons being that, the feature selection techniques reduced the dimensionality, redundancy in the variables, and retain the data variance.

In this way, overfitting can be minimized while increasing the degree of freedom of the base model (XGBoost) during classification tasks. In particular, the process of determining the best feature importance is repeated until expected features are identified and eventually chosen.

The significance (2-tailed) test conducted on the outcomes obtained for the Filter technique with test value of 0.7576 as the accuracy score which is $p<0.05$ (that, $p = 0.059$). Therefore, the filter technique outcomes are statistically significant at confidence value of 0.95.

5 Conclusion

This paper compared the feature selection performances of the different approaches including wrapper (Lasso, Recursive feature elimination), filter (Chi-Square), and machine learning algorithm (Random Forest). The base model was XGBoost because of its capacity to aggregate weak classifiers' outcomes to reach an improved decision. The different feature selection techniques were applied on the 8 features in the diabetes dataset. Majority of the feature selection techniques identified the following attributes has best for diagnosis heart diseases including: Glucose, BMI, Age, Pregnancies, BloodPressure, Insulim, and DiabetesPedigreeFunction.

The filter technique improves the accuracy of the XGBoost model by 75.76%; precision of 64.63%, and F1-score of 65.43%. Similarly, the mutual information theory or embedded technique gives the best recall score of 71.25% followed by the filter technique at 66.25%. the same trend was observed in that the mutual information provided the least false-positive of 23 superiors to the filter technique at 27. The filter technique outcomes two-tailed significance test score of $p = 0.059$), that is, $p<0.05$. It implies outcomes obtained are statistically significant at confidence value of 95%.

These feature selection techniques reduced the dimensionality, redundancy in the variables, and retain the data variance. Also, base model's overfitting is minimized while increasing its degree of freedom during classification tasks. Both feature selection methods undertake extended processes of determining the best feature importance before eventually choosing the best attributes within the data to explain the variance and patterns. Future work can consider use of mixture of the feature selection techniques, base models, and more complex datasets to ascertain the effectiveness of the approaches.

References

[1] D. A. Elmanakhly, M. M. Saleh, and E. A. Rashed, "An improved equilibrium optimizer algorithm for features selection: methods and analysis," *IEEE Access*, vol. 9, pp. 120309–120327, 2021, doi: 10.1109/access.2021.3108097.

[2] J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," *Electronic Markets*, vol. 32, no. 4, pp. 2159–2184, 2022, doi: 10.1007/s12525-022-00608-1.

[3] N. Jameel and H. S Abdullah, "Intelligent feature selection methods: A survey," *Engineering and Technology Journal*, vol. 39, no. 1B, pp. 175–183, 2021, doi: 10.30684/etj.v39i1b.1623.

[4] Y.-E. Kim, Y.-S. Kim, and H. Kim, "Effective feature selection methods to detect IoT DDoS attack in 5G core network," *Sensors*, vol. 22, no. 10, p. 3819, 2022, doi: 10.3390/s22103819.

[5] A. K. Hamoud *et al.*, "A comparative study of supervised/unsupervised machine learning algorithms with feature selection approaches to predict student performance," *International Journal of Data Mining, Modelling and Management*, vol. 15, no. 4, pp. 393–409, 2023, doi: 10.1504/ijdmmm.2023.134590.

[6] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," *Eng Appl Artif Intell*, vol. 100, p. 104210, 2021, doi: 10.1016/j.engappai.2021.104210.

[7] Y. Chen *et al.*, "Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests," *Ecol Indic*, vol. 135, p. 108545, 2022, doi: 10.1016/j.ecolind.2022.108545.

[8] F. Saberi-Movahed *et al.*, "Decoding clinical biomarker space of COVID-19: Exploring matrix factorization-based feature selection methods," *Comput Biol Med*, vol. 146, p. 105426, 2022, doi: 10.1016/j.compbiomed.2022.105426.

[9] Y.-N. Sun, W. Qin, J.-H. Hu, H.-W. Xu, and P. Z. H. Sun, "A Causal Model-Inspired Automatic Feature-Selection Method for Developing Data-Driven Soft Sensors in Complex Industrial Processes," *Engineering*, vol. 22, no. 3, pp. 82–93, 2023, doi: 10.1016/j.eng.2022.06.019.

[10] Y. Zhang, H. Zhang, and B. Zhang, "An effective ensemble automatic feature selection method for network intrusion detection," *Information*, vol. 13, no. 7, p. 314, 2022, doi: 10.3390/info13070314.

[11] H. Almazrua and H. Alshamlan, "A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data," *IEEE Access*, vol. 10, pp. 71427–71449, 2022, doi: 10.1109/access.2022.3185226.

[12] O. Elsherbiny, Y. Fan, L. Zhou, and Z. Qiu, "Fusion of feature selection methods and regression algorithms for predicting the canopy water content of rice based on hyperspectral data," *Agriculture*, vol. 11, no. 1, p. 51, 2021, doi: 10.3390/agriculture11010051.

[13] H. N. Alshaer, M. A. Otair, L. Abualigah, M. Alshinwan, and A. M. Khasawneh, "Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application," *Multimed Tools Appl*, vol. 80, pp. 10373–10390, 2021, doi: 10.1007/s11042-020-10074-6.

[14] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022, doi: 10.3389/fbinf.2022.927312.

[15] F. Bagherzadeh, M.-J. Mehrani, M. Basirifard, and J. Roostaei, "Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance," *Journal of Water Process Engineering*, vol. 41, p. 102033, 2021, doi: 10.1016/j.jwpe.2021.102033.

[16] S. Jain and A. Saha, "Rank-based univariate feature selection methods on machine learning classifiers for code smell detection," *Evol Intell*, vol. 15, no. 1, pp. 609–638, 2022, doi: 10.1007/s12065-020-00536-z.

[17] J. Too and S. Mirjalili, "General learning equilibrium optimizer: a new feature selection method for biological data classification," *Applied Artificial Intelligence*, vol. 35, no. 3, pp. 247–263, 2021, doi: 10.1080/08839514.2020.1861407.

[18] G. Ge and J. Zhang, "Feature selection methods and predictive models in CT lung cancer radiomics," *J Appl Clin Med Phys*, vol. 24, no. 1, p. e13869, 2023, doi: 10.1002/acm2.13869.

[19] A. Al Marouf, M. M. Mottalib, R. Alhajj, J. Rokne, and O. Jafarullah, "An efficient approach to predict eye diseases from symptoms using machine learning and ranker-based feature selection methods," *Bioengineering*, vol. 10, no. 1, p. 25, 2022, doi: 10.3390/bioengineering10010025.

[20] F. Kamalov, H. Sulieman, S. Moussa, J. A. Reyes, and M. Safaraliev, "Nested ensemble selection: An effective hybrid feature selection method," *Heliyon*, vol. 9, no. 9, 2023, doi: 10.1016/j.heliyon. 2023.e19686.

[21] Y. Zhang, J. Liu, W. Li, and S. Liang, "A proposed ensemble feature selection method for estimating forest aboveground biomass from multiple satellite data," *Remote Sens (Basel)*, vol. 15, no. 4, p. 1096, 2023, doi: 10.3390/rs15041096.

[22] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft comput*, vol. 27, no. 16, pp. 11259–11274, 2023, doi: 10.1007/s00500-023-08587-x.