International Journal of Advanced Technology and Engineering Exploration, Vol 8(85)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.874615

Research Article

A two-phase feature selection technique using mutual information and XGB-RFE for credit card fraud detection

C. Victoria Priscilla[1] and D. Padma Prabha[2*]
PG Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women (Autonomous), Affiliated to University of Madras, Chennai, India[1]
Department of Computer Applications, Madras Christian College (Autonomous), Affiliated to University of Madras, Chennai, India[2]

Received: 17-August-2021; Revised: 18-December-2021; Accepted: 20-December-2021

Abstract
With the rapid increase in online transactions, credit card fraud has become a serious menace. Machine Learning (ML) algorithms are beneficial in building a good model to detect fraudulent transactions. Dealing with high-dimensional and imbalanced dataset becomes a hinder in real-world applications like credit card fraud detection. To overcome this issue, feature selection a pre-processing technique is adopted considering the classification performance and computational efficiency. This paper proposes a new two-phase feature selection approach that integrates filter and wrapper methods to identify the significant feature subsets. In the first phase, Mutual Information (MI) has been adopted due to its computational efficiency to rank the features based on their feature importance. However, they cannot drop the less important features. Thus, a second phase is added to eliminate the redundant features using Recursive Feature Elimination (RFE) a wrapper method employed by 5-fold cross-validation. eXtreme Gradient Boosting (XGBoost) is adopted as the estimator for RFE by adjusting the class weights. The optimal features obtained from the proposed method were used in four boosting algorithms such as XGBoost, Gradient Boosting Machine (GBM), Classic Gradient Boosting (CatBoost) and Light Gradient Boosting Machine (LGBM) to analyse the performance of classification. The proposed approach has been applied to the credit card fraud detection dataset obtained from the IEEE-CIS, which consists of imbalance in the binary class target. The experimental outcome shows promising results in terms of Geometric mean (G-Mean) for XGBoost (84.8%) and LGBM (83.7%), the Area Under a Receiver Operating Character (ROC) Curve (AUC) has increased from 79.8% to 85.5% for XGBoost and also the computation time are reduced in training the classifiers.

Keywords
Recursive feature elimination, Hyper-parameter optimization, Class imbalance, XGBoost, Binary classification.

1.Introduction
In 2025, Nilson report pointed out that the gross credit card fraud worldwide has been expected to be $35.31 billion [1]. It was found successful in fighting against criminals by Machine Learning (ML) models through analysing massive datasets generated, but still, the happening of fraud cannot be stopped [2]. The high volume of transactions needed to be processed to identify the fraud that does not happen frequently generating an imbalanced dataset [3]. The fraud considered as a legitimate transaction is the considerably higher cost than identifying a legitimate transaction as fraud [4]. As e-commerce widely grows, merchants are charged back for fraud loss generated [5].

ML models are developed to identify these anomalies in the credit card fraud detection problem, the supervised models are generally applied by many researchers to classify the binary targeted data.

Dealing with imbalanced data is a major challenge, focused on minimizing the error rate of the negative class while ignoring the positive class [6]. Facing the challenge of high dimensional imbalanced data in many real-world applications such as medical diagnostic, credit defaulters, fraud detection, etc. [7]. De Sá et al. [5] developed a customized classification algorithm that automatically generates the Bayesian network classifier to manage the class imbalance. Even though effective methods such as data level, algorithm level, hybrid and cost-sensitive learning is proposed by researchers to normalise the imbalanced

*Author for correspondence

dataset [3]. Oversampling methods influence the minority class while undersampling methods lose important information about the imbalanced distribution that is not remarkable on a high dimensional dataset [8].

Another big challenge in ML is the curse of dimensionality [9], as the dimensionality grows the search space also increases fast and data becomes scanty, resulting in space sparsity and overfitting. Therefore, feature selection attracted the research attention in dealing with high dimensional data to prevent overfitting. The quality of data is boosted by removing redundant and noisy features resulting in reducing the computational time by ignoring the irrelevant features [10]. The objective of feature selection is to minimize the number of features and maximize the performance of the classifier [11].

Feature selection is a pre-processing strategy to reduce the complexity of the high dimensional dataset without dropping the important features considered during the learning phase [12]. However, it is critical to identify which feature selection method is efficient in analyzing high-dimensional data [13]. Feature selection in ML can be classified as a filter, wrapper and embedded [7]. Filter methods compute the score for each feature and select the feature with the highest score. They are generally based on the properties of the dataset ignoring the learning models [14]. The best feature subset is based on the ranking scores of different statistical tests such as Correlation, Analysis of Variance (ANOVA), information gain and chi-square [11]. They cannot drop the irrelevant features as they are independent of each other, but evaluate the intrinsic behaviour of data [15]. Mutual Information (MI) is a popular filter method to estimate the association between individual features and mutual redundancy between two features [16].

Wrapper methods help to find solutions based on the fitness function. They interact with the learning algorithms to select the best features by training with different combinations of features based on the classification accuracy [11]. Algorithms namely best-first search, forward and backward elimination, recursive elimination can discover optimal features by enhancing or controlling the objective function however they have high computation costs [15]. Embedded methods integrate both filter and wrapper methods. It maintains the balance between computation time and precision. Methods like Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and ElasticNet have penalization methods to reduce the complexity and eliminate the feature with coefficients lesser than the threshold value [17].

To avoid the aforementioned limitations of the existing feature selection methods, a two-phase feature selection approach is being proposed in this study. The proposed hybrid model is a combination of filter and wrapper methods with their inherent benefits. The filter method, MI has been adopted in the first phase due to its computational efficiency to rank the features based on their feature importance. However, there is no decision boundary for the selected features in finding the optimal subset of features. Therefore a wrapper method Recursive Feature Elimination (RFE) is applied in the second phase to eliminate those irrelevant features. For the classification of binary targets four boosting algorithms eXtreme Gradient Boosting (XGBoost), Gradient Boosting Machine (GBM), Classic Gradient Boosting (CatBoost) and Light Gradient Boosting Machine (LGBM) respectively are selected due to their efficiency in preventing overfitting of data by regularizing the objective function. The optimal features obtained by the proposed two-phase approach are applied in the boosting models to train and evaluate the model performance.

The paper is organized as follows. Section 2 provides the literature review of the related works. Section 3 illustrates the methodology applied to the proposed two-phase feature selection. Section 4 discusses the results obtained by the proposed MI-XGB(w)-RFE method. Section 5 discusses the impact of the study and its limitations. Finally, Section 6 concludes the study with future work.

2.Literature review

As the complexity of the imbalanced data is inherent, feature selection is required to transform a huge amount of data into an informative data distribution [18]. MI is one of the common feature selection techniques adopted widely in many studies. It is a statistical model to find the dependency between the two features by getting information of one feature by the other feature [19].

Hancer et al. [16] proposed MI feature selection algorithm based on single and multi-objective differential evolution that can be modified for the continuous dataset. Merging MI with the measurement of kernel canonical correlation analysis algorithm to maximize the features and target class

labels and minimizes the redundancy between features [20]. Wrapper methods are efficient in their performance, but computationally high expensive compared to filter methods [21]. The wrapper method RFE is widely applied in previous studies, Jeon and Oh combined an ensemble method of RFE with Support Vector Machine (SVM), Random Forest (RF) and GBM learning algorithms using different feature weights [22]. Balancing the dataset with Synthetic Minority Oversampling Technique (SMOTE) and hyper-parameter optimization using GridSearchCV methods improved the RF classifier to detect the frauds and SVM-RFE was introduced to predict the features [23]. As the wrapper methods are more efficient, a forecasting model was proposed using wrapper based feature selection method for handling multi-objective technique [24].

A new feature selection method introduced by Tales Matos et al. [25] was building graph to determine most important features using association rules and propositional logic and achieved f-score up to 76.88%. Overlapping of data should also be noted during feature selection. A sparse feature selection method was built to reduce overlapping in a binary classification with data balancing techniques [26].

Hybrid methods are generally adopted in feature selection techniques [12]. Researchers have developed new adaptive hybrid feature selection technique by combining different individual methods to obtain a generalized solution [27].

The hybrid approach suggested by El-Hasnony et al. [28] combines grey wolf and particle swarm optimization and also solve the local optima using tent chaotic map and achieved an average accuracy of 90% for the balanced dataset. An RF - RFE approach was developed to find the significant features after removing the redundant features by correlation method [15].

Lian et al. [29] proposed a novel stacking model based on RFE with cross-validation, calculating the sum of decision coefficient to retain the best features for intrusion detection.

A framework was proposed by Zhang et al. [30] for feature selection based on behavior analysis on a collection of homogenous historical transactions. A novel approach cumulative distribution function gradient was embedded to determine the features automatically identifying the rank of each feature, and these features are fed into the second phase to derive the best features using function perturbation ensemble [31]. Even though hybrid framework is effective, it has the limitation in selection of method [32].

Integration of feature selection techniques with metaheuristic models was also popularly done by researchers to improve the performance rate. A two-layer approach was proposed with genetic algorithm in the first layer to reduce the search space for finding the optimal features and ElasticNet model in the second layer adjusts the penalty and regularization and further reduces the predicted features from the first layer [17]. A hybrid approach using multi-filter and correlation based redundancy in the first stage to identify the best features and in the second stage grasshopper optimization algorithm was applied to predict the best features [7]. A correlation feature selection embedded with Bat Algorithm to remove the correlated features and used an ensemble approach by training with multiple classifiers using the voting method based on probabilities [33]. A novel filter-based approach was implemented to group features based on correlation coefficient and integrated with a cuttlefish algorithm to improve the performance [34]. A genetic crow search algorithm was developed for feature selection and used convolutional neural network for classification and obtained 95.34% of the selected features for heart disease prediction [35].

Boosting algorithms improve the accuracy of weak learners. The performance of different learning algorithms performs well with different predictors, even if the training set is the same [36]. In this experiment boosting algorithms were used as learning models, namely GBM, XGBoost, CatBoost and LGBM for training and validation of the models. Boosting algorithms integrate the weak learners created sequentially to provide a good accuracy. The residual errors produced by the previous learner are corrected in each iteration. Finally, the boosting weight predicts the performance [37]. GBM uses the boosting technique with the additive learning of weak learners, later new algorithms were introduced based on GBM. Information gain was obtained by LGBM and CatBoost to solve the problem of target leakage found in GBM. XGBoost uses the regularization term to bound the complexity of the model.

Although various studies on combining filter and wrapper methods have been proposed in the literature, the contribution of this study focused on the classification of fraudulent and legitimate

transactions from a high-dimensional imbalanced dataset. The proposed method combines MI the filter and RFE the wrapper methods to remove the irrelevant features that minimize the computation time and maximize the performance. SVM and RF are widely used learners that are embedded in the RFE, in this study XGBoost an efficient learner with class weight is embedded in the RFE and designed in such a way to tackle class imbalance.

3.Methods

This section elaborates the proposed two-phase feature selection model. In the first phase, a filter method MI approach is implemented to select the best feature subset. In the second phase, the features extracted from the first phase are given to the RFE with cross-validation further to reduce the remaining irrelevant features to increase the classification accuracy.

3.1Mutual Information (MI)

MI is an efficient filter method that measures the relationship between the variables. *Figure 1* shows the theoretical approach of MI between two discrete random variables X and Y for N observation is defined as Equation 1.

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{x\varepsilon X, y\varepsilon Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \quad (1)$$

In *Equation 1*, $H(Y)$ represents the entropy of Y which measures the degree of ambiguity in a discrete random variable Y and $H(Y|X)$ represents the conditional entropy of X given Y; $p(.)$ denote the probability mass function [20]. In the context of feature selection, MI provides a way to compute the relevance of the feature subset. If the MI value is greater between the two discrete random variables, then the two variables are highly dependent, when the MI value is zero then the variables X and Y are statistically independent. MI is linearly related to the entropies of the variables X and Y through the resulting *Equation 2.*

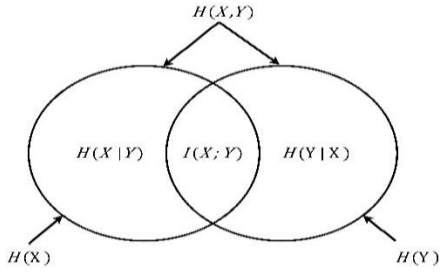


Figure 1 Venn diagram showing the relationship between mutual information and entropy

$$I(X;Y) = \begin{cases} H(X) - H(X|Y) \\ H(Y) - H(Y|X) \\ H(X) + H(Y) - H(X,Y). \end{cases} \quad (2)$$

The shortfall of this probabilistic method is testing the conditional independence for all features and estimating the probability density functions [38].

3.2Recursive feature elimination (RFE) with cross-validation

RFE is a wrapper feature selection approach that fits a learning model and eliminates the less important features. From the score obtained by the learning model, the features are ranked and eliminated recursively based on the iterations. RFE removes the dependency and collinearity among features. To obtain the optimal features a cross-validation technique was implemented in RFE to find the best scoring features by fitting the estimator several times in each iteration to remove the weakest features. The cross-validation approach combines different combinations of features using RFE. Based on the decision coefficient, the significant features are ranked by the score obtained from the classifier such that those features are retained. The feature subset does not remain the same for different classifiers [39]. Cross-validation helps to train the model to avoid overfitting of data, therefore it is used in the RFE technique to make the estimator learn and produce the best scoring weights of features. The Pseudo-code shows the process of RFE for any classifier. The decision coefficients are produced by training to extract the important features that are stored for the next iteration of training. RFE algorithm recursively eliminates the size of the feature in each recurrence. RFE with cross-validation is better compared with RFE to select the optimal features from the first n features by feature ranking [29].

3.3Proposed MI-XGB$_{(w)}$-RFE feature selection approach

The architecture of the proposed feature selection model is given in *Figure 2*. After the data preprocessing step, the training data is fed to the first phase of the proposed model. Even though MI is an influential statistical method used in many research, this filter method can able to rank the features by finding the dependencies between the pair of features [16]. Therefore the most important features ranked by the MI have been selected. The MI searches the irrelevant features based on the dependency and sorts them according to their feature importance. From the feature ranking, the best predictors are selected for

the next stage of processing by RFE with cross-validation. The best feature subset identified from the first phase is given as the input for the second phase to filter the features with high relevance and correlation. The RFE method eliminates the less important features recursively for each iteration based on the decision scores returned by the classifier.

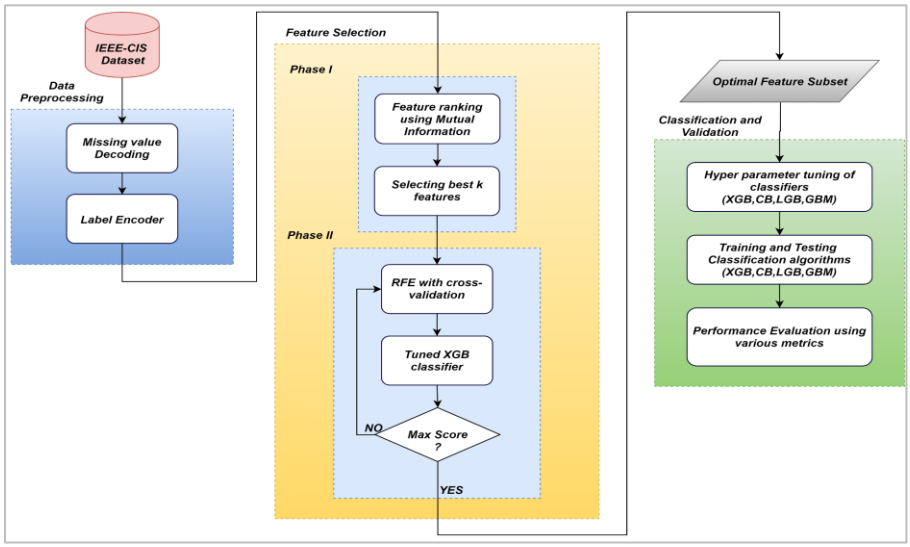


Figure 2 Flowchart of the proposed MI-XGB$_{(w)}$-RFE approach

As the focus of the study is to prevent overfitting of the model. Hence selected XGBoost a powerful model, which adds a regularization term to limit the complexity of the model. The reason for choosing XGBoost as the learner for RFE is that the class weight of the dataset is balanced by the learner to estimate the decision score. Thus, the probability of having irrelevant features is decreased, resulting in preventing the chance of bias in the prediction model. However, the performance of the model significantly depends on the hyper-parameters, hence they are tuned to find the optimal parameter for the model. The dataset is highly imbalanced the sample weights are adjusted to equalize the ratio of samples to solve the pitfall of overfitting. Therefore, in each iteration, the best scoring features are stored to get the final optimal features for training and validating the hyper-parameter tuned models for prediction.

The theoretical learning of feature selection problem can be formalized as: For a given training dataset T={X, F, C} consist of a training dataset for n samples where $X = [x_1, x_2, \dots, x_n]^T$ and feature set $F = [f_1, f_2, \dots, f_p]$ to denote the features for the samples in X. Each instance in X belongs to a binary class label C = {1,0}. The proposed feature selection model consists of two phases. In the first phase, the MI filter method has been implemented to find the best subset of features. Hence few subsets of the less important features have been removed to increase the computational efficiency and classification accuracy. In the second phase, RFE with a cross-validation approach is adopted to eliminate the remaining redundant and correlated features existing in the first phase. In cross-validation, the dataset is divided into k folds where k-1 folds are considered as training set and 1 fold as the testing set to find the score values for N observations. This process is repeated k times such that each fold will be considered once for validation. Hence, it outperforms other methods to reduce the high dimensional features to avoid overfitting of the model. The feature ranking attained by the proposed model is presented in *Figure 3*.

Pseudo-code for recursive feature elimination

Input:

$X_0 = [x_1, x_2, \dots, x_n]^T$- Training Dataset for samples.

$F = [f_1, f_2, \dots, f_p]$ - Set of p features

$S = [1,2, \dots, D]$ - Subset of features for each sample x_i to rank them

Ranking Method M(X,F)

Output:

The ranked feature set R.

Begin:

Set R =[ ]

While S ≠ [ ] do

Repeat for I in [1 : p]

Rank the feature set using M(X,F)

$S(f^*) \leftarrow$ last ranked feature in F
$R(p - i + 1) \leftarrow S(f^*)$
$S(R) \leftarrow S(R) - S(f^*)$
End while
End

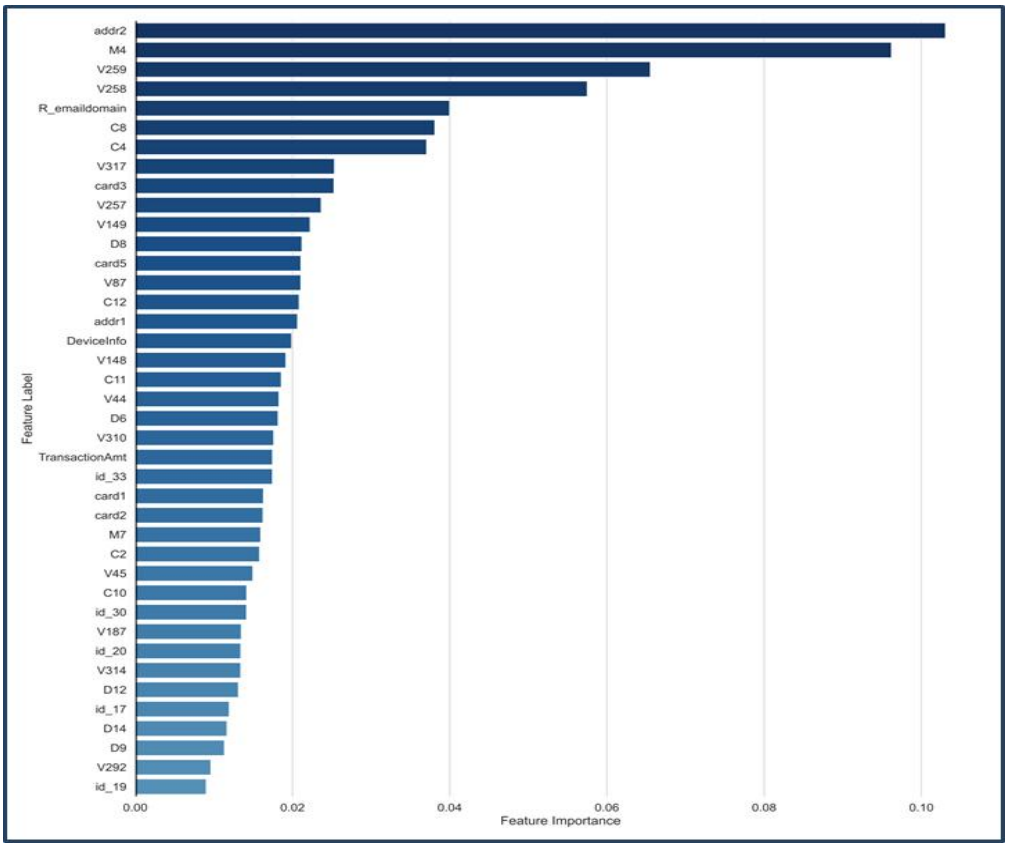


Figure 3 Feature importance of the proposed MI-XGB$_{(w)}$-RFE

3.4 Data description and pre-processing

The IEEE-CIS fraud detection dataset provided by Kaggle has two tables, namely transaction and identity [40]. In the transaction table, there are 394 features comprising 372 numerical features and 22 categorical features. Most of the numeric features are anonymous. The identity table consists of 41 features including categorical and numeric information. These two tables are merged with the primary key column TransactionID. The target feature isFraud is the binary target with a high imbalance in 1's and 0's. The dataset is highly imbalanced with 3.4% of fraudulent transactions. Hence the overall training features used in the experiment was 433. However, the dataset has been preprocessed as there are lots of Not a Number (NaN) in the numerical and categorical columns is replaced by -1 and 'NA' respectively. All the categorical features are transformed to numerical form using the label encoder technique.

4.Experimental results and analysis

This section briefs the objective of the proposed two-phase feature selection method MI-XGB$_{(w)}$-RFE over the existing four popular boosting algorithms in dealing with an imbalanced and high dimensional dataset. Moreover, hyper-parameter tuning was carried out for the ML models since it avoids overfitting and improves the performance of the model. The effectiveness of the proposed model is compared with other feature selection techniques and with the original features. This experiment was done on an Intel(R) Core(TM) i3-7100U CPU and 12 GB RAM in windows 10 platform. The algorithm was implemented using Python.

4.1Evaluation metrics

Determining suitable evaluation metrics is a foremost issue in the binary classification problem having an imbalanced dataset [41]. Measuring the quality of fraud detection is to increase the detection rate and reduce false positives.

To evaluate the performance of the classifier, the following evaluation metrics have been adopted for this study. Accuracy (ACC) is the commonly used metric for evaluation, but cannot be a prominent metric to evaluate the performance while the minority class is few [8]. Hence other metrics F1-score, G-mean, Area Under a Receiver Operating Character (ROC) Curve AUC, etc. are also considered. Firstly the classification metrics are derived from the confusion matrix given in *Table 1*. The True Positives (TP): samples that are correctly predicted as fraudulent. The True Negatives (TN): samples that are correctly predicted as legitimate. The False Positives (FP): samples that are predicted as fraudulent but actually legitimate. The False Negatives (FN): samples that are predicted as legitimate but actually fraudulent.

Table 1 Confusion matrix

Actual	**Predicted**	
	Fraudulent	**Legitimate**
Fraudulent	TP	FN (Type 2 Error)
Legitimate	FP (Type 1 Error)	TN

Precision

Precision measures the proportion of correctly predicted positive targets that are positive (Equation 3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

Recall

Recall or Sensitivity measures the positive targets that are correctly predicted as positive (Equation 4).

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

F1-score

F1-score (Equation 5) is an accuracy metric that measures the harmonic mean of precision and recall from *Equation 3* and *4*.

$$\text{F1} - \text{score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision+ Recall}} \quad (5)$$

Area under ROC (AUC)

AUC is the capability of distinguishing between positive and negative classes. When AUC is higher the capability of prediction is better, the value bounds between 0 and 1.

Root mean squared error (RMSE)

RMSE is the standard deviation of prediction errors used to measure the difference between the values predicted by the model. It is formulated as Equation 6.

$$\text{RMSE} = \sqrt{\sum_{j=1}^{n} \frac{(y_j - y'_j)^2}{n}} \quad (6)$$

In *Equation 6*, n is the number of observations, y_j is the actual value and y'_j is the corresponding predicted value.

G-mean

G-mean measures the balance between the classifications of binary classes at a specific threshold.

$$\text{GM} = \sqrt{\text{Sensitivity X Specificity}} \quad (7)$$

Mathew's correlation coefficient (MCC)

MCC measures (Equation 8) the correlation between the observed and predicted classifications where the value is between -1to1. If the coefficient is +1 indicates a perfect prediction, -1 represents the prediction is exactly wrong.

$$\text{MCC} = \frac{\text{TP X TN} - \text{FP X FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \quad (8)$$

Brier score (BS)

BS measures (Equation 9) the mean squared error between predicted probability and their respective positive class. It checks the goodness of the predicted probability value.

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2 \quad (9)$$

In *Equation 9*, f_t represents the forecast probability and o_t the actual outcome of the event at time t.

Cohen's kappa(K)

Cohen Kappa ranges from -1 to +1, it measures the inter-rater reliability. It is calculated by the formula Equation 10.

$$K = \frac{p_o - p_e}{1 - p_e} \quad (10)$$

In *Equation 10*, p_o is the overall accuracy of the classifier and p_e is the agreement between the predictions.

4.2Hyper-parameter tuning

The performance of the algorithm depends on the hyper-parameters for complex models [23]. Default parameters can affect the performance of an algorithm, therefore parameter tuning is done for the specific problem. Parameters are set in such a way to attain the highest efficiency. Hence weights w of the target class is set to balance the class weights of XGBoost, an estimator of RFE to avoid overfitting of data. At the same time, the tuning process is done for boosting algorithms to improve the efficiency of classification. A random search approach has been devised to select the best parameters. Different values are considered for each parameter in the random search subset to find the best possible value. *Table 2* summarizes the tuned parameters of the classifiers for the original features and feature subset obtained from

different feature selection methods such as 80 features from MI , 122 features from XGB-RFE, 194 features from Correlation based Filter Stage (CFS) and 50 features from the proposed method MI-XGB(w)-RFE are used in this study.

Table 2 Hyper-parameter selection and optimal values of boosting classifiers for different feature selection methods

XGBoost			**Optimal value**		
parameters	**original**	**MI**	**XGB-RFE**	**CFS**	**MI-XGB$_{(w)}$-RFE**
n_estimators	400	400	350	500	250
min_child_weight	10	20	15	30	10
max_depth	100	20	120	120	180
learning_rate	0.1	0.75	0.1	0.5	0.25
gamma	0.1	0.1	0.05	0.01	0.005
colsample_bytree	0.1	0.1	0.05	0.1	0.1
GBM			**Optimal value**		
parameters	**original**	**MI**	**XGB-RFE**	**CFS**	**MI-XGB$_{(w)}$-RFE**
subsample	1	1	0.9	1	0.8
n_estimators	400	450	500	500	350
min_samples_split	0.1	0.005	0.01	0.05	0.005
min_samples_leaf	0.01	0.005	0.01	0.05	0.005
max_features	auto	auto	log2	sqrt	sqrt
max_depth	140	60	20	80	140
learning_rate	0.25	0.25	0.5	0.25	0.1
CatBoost			**Optimal value**		
parameters	**original**	**MI**	**XGB-RFE**	**CFS**	**MI-XGB$_{(w)}$-RFE**
od_type	Iter	Iter	IncToDec	IncToDec	Iter
learning_rate	0.25	0.05	0.1	0.5	0.1
iterations	300	500	150	450	400
LGBM			**Optimal value**		
parameters	**original**	**MI**	**XGB-RFE**	**CFS**	**MI-XGB$_{(w)}$-RFE**
n_estimators	150	500	250	500	500
min_child_weight	35	25	30	50	10
learning_rate	0.1	0.25	0.05	0.05	0.05
colsample_bytree	0.25	0.75	0.5	0.25	1

4.3Model validation

The experimental result comparing the proposed two-phase feature selection method with the two benchmarks MI a filter method and RFE with XGB$_{(w)}$ a wrapper method. Combining the two benchmarks to a single phase feature selection method, MI is an efficient filter method in selecting the irrelevant features furthermore XGBoost was chosen as the estimator for RFE due to its speed and performance. A 5-fold cross-validation was performed in RFE for the selection of optimal features. The estimator XGB $_{(w)}$ of the proposed feature selection method is tuned with hyper-parameters and the performance are evaluated by the ROC-AUC score attributes of RFE. For validating the proposed feature selection method, four boosting algorithms are constructed with tuned hyper-parameters. The classification performance is measured for all the models with original features and features obtained from other feature selection methods. The proposed model was tested with 3 different feature subsets 40, 50 and 80 to find the best performing method. Finally, the proposed method MI-XGB$_{(w)}$-RFE with 50 feature subsets was selected for other evaluation metric comparisons. The performance measures obtained through the proposed feature selection process are better than the other feature selection methods used in the experiment. *Table 3* shows the performance comparison of evaluation metrics such as precision, recall, f1-score, balanced accuracy, MCC, Cohen's Kappa and AUC for the proposed MI-XGB$_{(w)}$-RFE with respect to features. *Figure 4* illustrates the classification performance in terms of G-mean of boosting algorithms with original features and features obtained by the filter methods MI and CFS as well as the wrapper method XGB-RFE followed by the proposed MI-XGB$_{(w)}$-RFE . The number of features selected in each feature selection method varies, the

proposed method has the minimum number of features compared with other methods. We can see that the G-mean of XGBoost, CatBoost and LGBM attained above 80% with the optimal number of features obtained by the proposed MI-XGB$_{(w)}$-RFE where XGBoost and LGBM achieved 84.8% and 83.7% out of the three best performed algorithms.

Table 3 Classification performance of the proposed MI-XGB$_{(w)}$-RFE in comparison with different feature subsets

Classifiers	**Original features (433 features)**						
	Precision	**Recall**	**F1-score**	**Bal-Accuracy**	**MCC**	**Kappa**	**AUC**
XGBoost	0.903	0.598	0.720	0.798	0.729	0.713	0.798
GBM	0.954	0.506	0.661	0.753	0.689	0.655	0.753
CatBoost	0.577	0.724	0.642	0.854	0.635	0.631	0.854
LightGBM	0.353	0.773	0.485	0.866	0.504	0.465	0.866
Classifiers	**MI-XGB$_{(w)}$-RFE (80 features)**						
	Precision	**Recall**	**F1-score**	**Bal-Accuracy**	**MCC**	**Kappa**	**AUC**
XGBoost	0.602	0.696	0.646	0.842	0.637	0.635	0.842
GBM	0.883	0.301	0.449	0.650	0.508	0.441	0.650
CatBoost	0.535	0.709	0.609	0.846	0.603	0.597	0.846
LightGBM	0.429	0.730	0.540	0.851	0.544	0.524	0.834
Classifiers	**MI-XGB$_{(w)}$-RFE (50 features)**						
	Precision	**Recall**	**F1-score**	**Bal-Accuracy**	**MCC**	**Kappa**	**AUC**
XGBoost	0.401	0.742	0.520	0.855	0.529	0.503	0.855
GBM	0.892	0.506	0.646	0.752	0.665	0.639	0.752
CatBoost	0.626	0.663	0.644	0.826	0.634	0.633	0.826
LightGBM	0.363	0.727	0.485	0.845	0.496	0.465	0.845
Classifiers	**MI-XGB$_{(w)}$-RFE (40 features)**						
	Precision	**Recall**	**F1-score**	**Bal-Accuracy**	**MCC**	**Kappa**	**AUC**
XGBoost	0.728	0.672	0.699	0.832	0.691	0.690	0.832
GBM	0.929	0.482	0.634	0.740	0.663	0.627	0.740
CatBoost	0.341	0.724	0.463	0.842	0.477	0.442	0.842
LightGBM	0.505	0.721	0.594	0.850	0.590	0.581	0.850

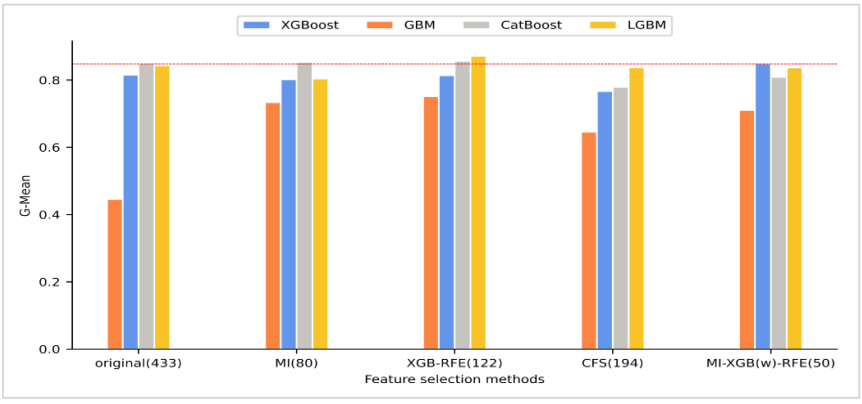
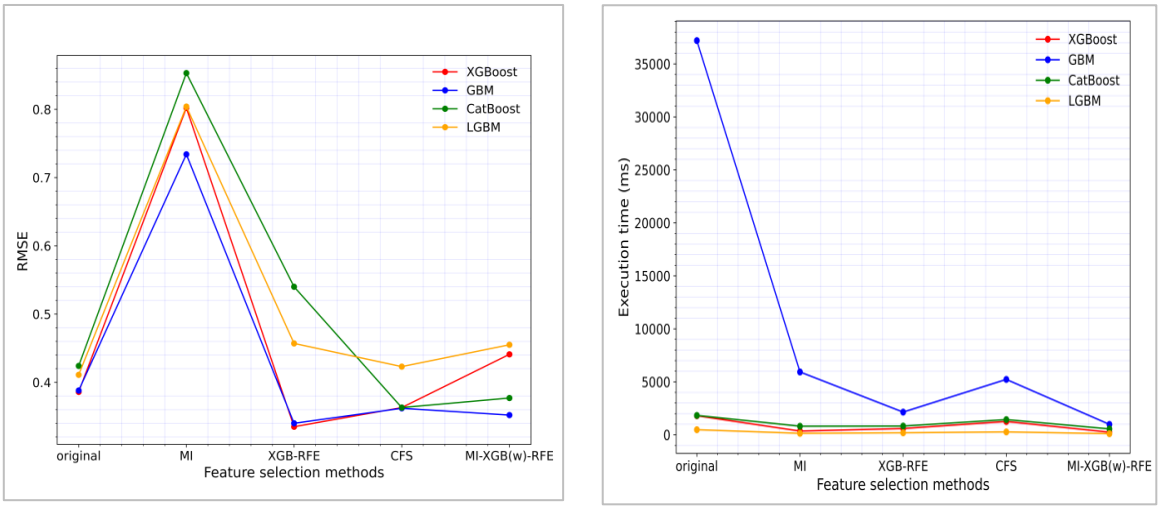


Figure 4 Comparison of proposed MI-XGB$_{(w)}$-RFE with other existing feature selection methods for different boosting classifiers with respect to G-mean

Figure 5(a) represents the RMSE that measures the standard deviation of the predicted errors of the boosting algorithms. As the values range between 0.2 and 0.5 of original features, CFS, MI-XGB$_{(w)}$-RFE for all the classifiers indicating that the model can proportionately predict the data accurately. But CatBoost model does not perform well on XGB-RFE and for MI as the RMSE values are above 0.7 for all

the models. *Figure 5(b)* compares the execution time for the different feature selection methods. The execution time is reduced for all the boosting algorithms performed by MI-XGB$_{(w)}$-RFE when compared with other feature selection algorithms and with the original features. These outstanding results show that highly important features selected by the proposed MI-XGB$_{(w)}$-RFE method alone can help to increase the performance of the model by reducing the complexity due to the high dimension of data.



(a) (b)

Figure 5 Comparison of classifiers for different feature selection methods with respect to (a) RMSE (b) Execution time

5.Discussion

The results are promising due to the elimination of features by the combined feature selection method. The features obtained by different feature selection approaches MI, XGB-RFE, CFS and the original features are also implemented to compare the performance of the proposed MI-XGB$_{(w)}$-RFE with 50 features. Hyper-parameter tuning by random search method identifies the optimal parameters for the estimator of RFE and the boosting model. The class weights of the estimator XGB are adjusted to balance the dataset. The limitations of the existing MI filter method can only be able to rank the feature importance. Therefore, dataset with low dimensions can be performed well, but for the dataset with high dimensions, MI filter method is impractical. The existing wrapper method RFE is more expensive on computation time for the dataset with high dimensions. To overcome these limitations this study was performed by combining MI and RFE. As the highly ranked features are extracted from the first phase and fed as the input for RFE in the second phase can able to reduce the computation time of RFE. Since the dataset is highly imbalanced, XGBoost was opted to be the estimator for RFE.

The optimal values are obtained from the hyper-parameter tuning for the four classifiers. The common parameters like *n_estimators, max_depth, learning_rate*, etc. are estimated. The important parameter *n_estimators* of XGBoost and GBM for the features attained by the proposed method was reduced to 250 and 350 respectively were the estimators for original features are 400. The features *M4, Card3, V156, addr2, V249, V258* are found to be highly dominating features that are listed in *Figure 3*. The experiment to find the best feature subset of the proposed method was done with 80, 50 and 40. The AUC values of XGboost, GBM, CatBoost, LGBM are 85.5%, 75.2%, 82.6% and 84.5% respectively are substantial for the proposed method with 50 features. Hence, decided to choose the proposed method with 50 features. The G-mean measures the balance between the majority and minority class, the XGBoost, CatBoost and LGBM attained above 80% by the proposed MI-XGB$_{(w)}$-RFE. G-mean achieved for XGBoost and LGBM are 84.8% and 83.7%, respectively out of the three best performed classifiers.

The comparison of performance metrics for the four boosting models is given in *Table 3*. From the outcome of the results, the impact of the study is

- The computation time of RFE is reduced by eliminating the less important feature subset as the ranking is done in the first phase by MI.

- The dataset is balanced by optimizing the parameters of the classifiers to improve the performance.
- As the dimensionality is reduced by the proposed method the execution time for all the classifiers is also reduced. It was noted that the execution time for XGBoost was reduced from 1794 ms to 241 ms.

5.1Limitations

Even though we have the advantages of the proposed method, there are some limitations in our method. It is difficult to identify the limit in choosing the number of ranked features from the MI in the first phase. As we used recursive elimination the time complexity is slightly higher when compared with other methods. The performance of parameters depends on the method used for hyper-parameter tuning.

6.Conclusion and future work

In this study, a novel two-phase feature selection approach MI-XGB(w)-RFE is proposed to select the optimal subset of features to improve the prediction rate of credit card fraud detection. In the first phase, MI a filter method was employed to remove the related and irrelevant features. In the second phase, the highly ranked features are used as input to RFE a wrapper method to eliminate the less important features using weighted XGBoost as an estimator. The most important features are obtained as the best feature subset for training and testing to compare the performance of four boosting algorithms XGBoost, GBM, CatBoost and LGBM respectively. The parameter optimization of the boosting algorithms is tuned to increase the efficiency of the model and to avoid overfitting. A real-time credit card fraud detection dataset from IEEE-CIS are used for the current study. To validate the effect of the proposed feature selection method, a comparison is done with other feature selection methods for the same boosting algorithms. According to the results obtained, the proposed method reduced the dimensionality of the dataset by considering only the most important features, thereby, enhancing the efficiency of the classifier by decreasing the computation time. The experimental results are promising in terms of G-mean by attaining 84.8% for XGBoost and 83.7% for LGBM with the subset of the 50 features. The comparative analysis on execution time between the original features and the proposed method for the classifiers was done, it was noted that our proposed method reduces the execution time from 1794 ms to 241ms for XGBoost by achieving AUC from 79.8% to 85.5%.

Our future research will focus on testing the generalization ability of our proposed method by employing different dataset. Furthermore, other advanced optimization techniques can be implemented to reduce the time complexity of RFE.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1187. Accessed 22 July 2021.
[2] Bagga S, Goyal A, Gupta N, Goyal A. Credit card fraud detection using pipeling and ensemble learning. Procedia Computer Science. 2020; 173:104-12.
[3] Liu Y, Wang Y, Ren X, Zhou H, Diao X. A classification method based on feature selection for imbalanced data. IEEE Access. 2019; 7:81794-807.
[4] Mahmoudi N, Duman E. Detecting credit card fraud by modified fisher discriminant analysis. Expert Systems with Applications. 2015; 42(5):2510-6.
[5] De SAG, Pereira AC, Pappa GL. A customized classification algorithm for credit card fraud detection. Engineering Applications of Artificial Intelligence. 2018; 72:21-9.
[6] El HS, Malki J, Bouju A, Berrada M. A machine learning based approach to reduce behavioral noise problem in an imbalanced data: application to a fraud detection. In international conference on intelligent data science technologies and applications 2020 (pp. 11-20). IEEE.
[7] Abdulrauf SG, Zainol Z. Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. Genes. 2020; 11(7):1-26.
[8] Chen H, Li T, Fan X, Luo C. Feature selection for imbalanced data based on neighborhood rough sets. Information Sciences. 2019; 483:1-20.
[9] Pilnenskiy N, Smetannikov I. Feature selection algorithms as one of the python data analytical tools. Future Internet. 2020; 12(3):1-14.
[10] Liu H, Zhou M, Liu Q. An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of Automatica Sinica. 2019; 6(3):703-15.
[11] Abdel-basset M, El-shahat D, El-henawy I, De AVH, Mirjalili S. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Systems with Applications. 2020.
[12] Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: a review. Egyptian Informatics Journal. 2018; 19(3):179-89.
[13] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection

in high-dimensional classification data. Computational Statistics & Data Analysis. 2020.
[14] Albashish D, Hammouri AI, Braik M, Atwan J, Sahran S. Binary biogeography-based optimization based SVM-RFE for feature selection. Applied Soft Computing. 2021.
[15] Elavarasan D, Vincent PM DR, Srinivasan K, Chang CY. A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agriculture. 2020; 10(9):1-27.
[16] Hancer E, Xue B, Zhang M. Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-Based Systems. 2018; 140:103-19.
[17] Amini F, Hu G. A two-layer feature selection method using genetic algorithm and elastic net. Expert Systems with Applications. 2021.
[18] Fu GH, Wu YJ, Zong MJ, Yi LZ. Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics. Chemometrics and Intelligent Laboratory Systems. 2020.
[19] Barraza N, Moro S, Ferreyra M, De LPA. Mutual information and sensitivity analysis for feature selection in customer targeting: a comparative study. Journal of Information Science. 2019; 45(1):53-67.
[20] Wang Y, Cang S, Yu H. Mutual information inspired feature selection using kernel canonical correlation analysis. Expert Systems with Applications: X. 2019.
[21] Zhang J, Xiong Y, Min S. A new hybrid filter/wrapper algorithm for feature selection in classification. Analytica Chimica Acta. 2019; 1080:43-54.
[22] Jeon H, Oh S. Hybrid-recursive feature elimination for efficient feature selection. Applied Sciences. 2020; 10(9):1-8.
[23] Rtayli N, Enneya N. Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. Journal of Information Security and Applications. 2020.
[24] Karasu S, Altan A, Bekiros S, Ahmad W. A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy. 2020.
[25] Matos T, Macedo JA, Lettich F, Monteiro JM, Renso C, Perego R, et al. Leveraging feature selection to detect potential tax fraudsters. Expert Systems with Applications. 2020.
[26] Omar B, Rustam F, Mehmood A, Choi GS. Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: application to fraud detection. IEEE Access. 2021; 9:28101-10.
[27] Viharos ZJ, Kis KB, Fodor Á, Büki MI. Adaptive, hybrid feature selection (AHFS). Pattern Recognition. 2021.
[28] El-hasnony IM, Barakat SI, Elhoseny M, Mostafa RR. Improved feature selection model for big data analytics. IEEE Access. 2020; 8:66989-7004.
[29] Lian W, Nie G, Jia B, Shi D, Fan Q, Liang Y. An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning. Mathematical Problems in Engineering. 2020.
[30] Zhang X, Han Y, Xu W, Wang Q. HOBA: a novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences. 2021; 557:302-16.
[31] Chiew KL, Tan CL, Wong K, Yong KS, Tiong WK. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences. 2019; 484:153-66.
[32] Singh N, Singh P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. Chemometrics and Intelligent Laboratory Systems. 2021.
[33] Zhou Y, Cheng G, Jiang S, Dai M. Building an efficient intrusion detection system based on feature selection and ensemble classifier. Computer Networks. 2020.
[34] Mohammadi S, Mirvaziri H, Ghazizadeh-ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. Journal of Information Security and Applications. 2019; 44:80-8.
[35] Nagarajan SM, Muthukumaran V, Murugesan R, Joseph RB, Meram M, Prathik A. Innovative feature selection and classification model for heart disease prediction. Journal of Reliable Intelligent Environments. 2021:1-11.
[36] Das S. Filters, wrappers and a boosting-based hybrid for feature selection. In ICML 2001 (pp. 74-81).
[37] Priscilla CV, Prabha DP. Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection. In third international conference on smart systems and inventive technology 2020 (pp. 1309-15). IEEE.
[38] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Computing and Applications. 2014; 24(1):175-86.
[39] Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics. 2019; 9(4):1-21.
[40] https://www.kaggle.com/c/ieee-fraud-detection/data. Accessed 11 July 2020.
[41] Luque A, Carrasco A, Martín A, De LHA. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019; 91:216-31.

Dr. C. Victoria Priscilla is the Head and Associate Professor in the PG Department of Computer Science at SDNB Vaishnav College for Women, Chennai. She is a Ph.D. holder in the field of Image Processing. She has presented many research papers at several international conferences and also published research articles in peer-reviewed journals. She has been a member at various levels of academic

events in other institutions. Her research interest includes Image Processing, Data Mining and Machine Learning.
Email: victoriapriscilla.c@sdnbvc.edu.in

D. Padma Prabha is an Assistant Professor in the Department of Computer Applications at Madras Christian College, Chennai. She has completed her master's in Computer Applications. She is currently pursuing Ph.D. from the Department of Computer Science, SDNB Vaishnav College for Women, affiliated to University of Madras, Chennai. Her research interests are Machine Learning, Artificial Intelligence and Data Science.
Email: padmaprabha@mcc.edu.in

Appendix I

S. No.	Abbreviation	Description
1	AUC	Area Under a Receiver Operating Character Curve
2	BS	Brier Score
3	CatBoost	Classic Gradient Boosting Machine
4	CFS	Correlation Based Filter Stage
5	FN	False Negatives
6	FP	False Positives
7	GBM	Gradient Boosting Machine
8	G-Mean	Geometric Mean
9	K	Cohen's Kappa
10	LASSO	Least Absolute Shrinkage and Selection Operator
11	LGBM	Light Gradient Boosting Machine
12	MCC	Mathews Correlation Coefficient
13	MI	Mutual Information
14	ML	Machine Learning
15	NaN	Not a Number
16	RF	Random Forest
17	RFE	Recursive Feature Elimination
18	RMSE	Root Mean Squared Error
19	SMOTE	Synthetic Minority Oversampling Technique
20	SVM	Support Vector Machine
21	TN	True Negatives
22	TP	True Positives
23	XGBoost	eXtreme Gradient Boosting