Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

A composite Feature Selection Method to improve Classifying Imbalanced Big Data

Shaymaa Ahmed Razoqi[1, *] and Ghayda Abdulaziz Al-Talib[2]

[1]*Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul, Iraq*
[2]*Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, IRAQ*
Emails: shymaa.raazoqi@uomosul.edu.iq, ghaydabdulaziz@uomosul.edu.iq

Article information

Article history:
Received:24/4/2024
Accepted:5/8/2024
Available online:15/12/2024

Abstract

Feature selection is one of the methods used to improve the performance of machine learning algorithms, especially when classifying the big data. Fined of new method was be more needed when dealing with the big data is imbalance. An imbalance in the data appears when there is a discrepancy in the sampling distribution between the two data classes in the training set. To solve the imbalance problem, there are several methods used, some of which depend on redistributing the data and others of which depend on improving the classification algorithm itself. The feature selection can also affect the improvement of imbalanced data classification results when the features are chosen carefully. Therefore, this research proposed a composite feature selection method using the filter feature selection technique and permutation-based important features with the ensemble learning method. Three classifiers were used with three performance metrics to show the effect of proposed feature selection method with imbalanced big data. The results of using proposed method led to improved classification on five standard imbalanced data sets.

Keywords:
Imbalance data; Big data; Permutation-Based Features Importance; Information Gain; Ensemble learning

Correspondence:
Author: Shaymaa Ahmed Razoqi
Email: shymaa.raazoqi@uomosul.edu.iq,

I. Introduction

Big data is complex data that requires a lot of computation and storage resources to process. Big data may be data with a large size of samples and a limited number of features or it may be data with a large number of features with a small number of samples, or the big data can be a very extensive with high dimensions and a very large sample size[1]. In recent years, large amounts of data have been collected in various fields such as public administration, marketing, health care, and research in the fields of chemistry, physics, and applications of social communication and intrusion detection [2]. This availability of big data in several fields has provided a new opportunity for researchers to work on extracting important information in different ways and innovative new ways and benefit from it in the best way. Big data includes the amount of data which has no importance when analyzing and processing for decision production. The main problem when processing big data lies in the good selection of features that have a high importance in describing that data. [2-3]

If the data set that is collected has an uneven distribution of data in classes, then it suffers from the problem of data imbalance. The issue of data imbalance appears when there is a class that has a big number of samples compared to a much smaller number of samples in another class in the binary classification. The big class is the majority, while the small class is the minority [4]. Machine learning

algorithms constitute as one of the artificial intelligence branches and it depends on building computer systems that can be trained on available data or gain experience when working without having to be explicitly programmed to deal with new data. Machine learning algorithms can gain knowledge, adapt to variables, and improve performance when new data is present [5].

In the imbalanced data set, the machine learning training faces difficulty as the number of majority class samples is too large [6]. This makes machine learning tend to the majority class, which leads to less performance in classifying minority class, and this makes classification models form inappropriate decision boundaries and makes minority samples invisible [7]. To address this challenge, there are three levels of solutions, one is data level, the second is algorithm level. Resampling methods are the policies that are followed at data level solutions, which seek to reduce the degree of data imbalance that exists between the classes [8-9]. Selecting samples from the majority class to be deleted is one of the easiest forms used to solve the data imbalance. Also, producing a new data from the minority class is another form within the data level solutions. Algorithm-level methods are usually called internal approaches because this solution focuses on improving the capability of the existing classifier algorithms to learn from the minority classes. Creating an algorithmic solution needs the acquaintance of both the related classifier learning algorithm and the application domain. The third level is hybrid methods which are a mix of the data level and the algorithmic level methods. [10].
The success of machine learning performance depends on the nature of the data related to the training process; the dimension curse is one of the most important characteristics of the data set that hinders the training process [11]. Dimensional reduction was used as a pre-processing step to solve imbalance problem when dealing with big data, in which features that do not affect the quality of machine learning are eliminated the appropriate number of features that must be kept are determined [12].

The current research proposed a composite feature selection method to improve the classification results of imbalanced data. The proposed method included a method that combines the results of filter methods for feature selection with permutation methods based on ensemble learning. The Decision Tree (DT), K-Nearest Neighbor (KNN) and Gaussian Naïve Bayes(GNB) classifiers were chosen to measure the improvement resulting after applying the proposed feature selection method using standard imbalanced datasets. The research paper is organized to show previous studies in Paragraph 2 and the classification of imbalanced data in Paragraph 3, while Paragraph 4 mentions feature selection methods, Paragraph 5 mentions ensemble learning methods, and Paragraph 6 mentions the proposed method and methodology, followed by Paragraph 7 to state the results and discussion.

II. Related work

Based on synthetic features, the experimental results demonstrated the effectiveness of the suggested strategy in handling a high-dimensional and imbalanced class data problem. Liu et al. presented a new embedded feature selection method Gini Index Feature Selection (GIFS) to deal with the high dimensional and imbalanced data using a weighted Gini index (WGI), where the attributes are weighted based on their approximated probability density values. This approach begins with assessing the contribution of each attribute, and the highest weighted attributes are selected. The experimental results proved that GIFS is an effective method to overcome feature selection and imbalanced data problems in comparison with Chi2, F-statistic and Gini index methods [13].

Thaher et al. introduced a new feature selection and skewed dataset using wrapper feature selection and Synthetic Minority Oversampling Technique (SMOTE). The Binary Queuing Search Algorithm (QSA) is used as a search strategy in a wrapper FS method. Binary QSA shows superior efficacy in feature selection compared to other algorithms, also the combination of QSA and SMOTTE achieves acceptable AUC results using 14 real datasets [14].

Pirgazi et al. proposed a hybrid technique based on the Incremental Wrapper-based feature Subset Selection (IWSSr) method and the Shuffled Frog Leaping Algorithm (SFLA) for gene selection in high-dimensional data sets. The method is implemented in two phases: filtering and wrapping. The filter phase used the Relief method for weighting. The wrapping phase used SFLA and IWSSr algorithms for search-effective features. The proposed method was evaluated using standard gene expression datasets and achieved a more compact set of features with high accuracy compared to similar methods [15].
Abdulrauf et al. developed a hybrid filter–wrapper feature selection approach to overcome the high dimensional and imbalanced data sets by selecting optimal feature subsets representing both minority and majority classes. The proposed method is called Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm (rCBR -BGOA) by implementing an ensemble of multi-filters coupled with the correlation methods [16].

Namous et al. proposed a swarm-based wrapper method for feature selection on highly skewed data sets. The g-mean fitness function is recommended for imbalanced datasets in the swarm method, where the accuracy fitness function should be avoided as it can mislead feature identification. The paper evaluates the implementation of the Evolutionary search Algorithm and Particle Swarm Optimization (EA-PSO) as feature selection techniques, while the number of selected features may slightly increase when using the g-mean fitness function [17].

Fu et al. introduced the Hellinger distance-based stable sparse feature selection(sssHD) algorithm in class imbalanced data. The proposed sssHD algorithm performs well and is competitive against existing selection methods and can be

easily extended with different rebalance sampling, sparse regularization structures and classifiers. The HD-based selection outperforms AUC-based and ROC-based selection in terms of FDR and shows limited variation with an increasing class-imbalance ratio. Hellinger distance can be used directly for feature selection without depending on any classifiers [18].

Ebiaredoh-Mienye et al. proposed combining information gain filter-based feature selection and cost-sensitive AdaBoost for the CKD detection approach. The experimental results of the proposed approach show improved performance compared to other classifiers; this approach can potentially be applied for early detection of CKD through computer-aided diagnosis [19].

FAHRUDY et al. suggest paper aims to reduce data complexity and irrelevant features. The study applies random oversampling and feature selections to overcome class-imbalanced data. It developed classification models using experimental-based data mining [20].

III. Imbalanced Data Classification

The negative impact of the imbalanced data set is very clear on ML classifiers. These classifiers are designed to learn from the available data when trained in large quantities and create virtual boundaries to distinguish between classes. The default boundaries were built by classifiers following different strategies based on the parameters for each classifier. The classification decision is often based on the boundary areas between classes in the data space. The classifier was tested by the test data, often these test data are new data for the classification. The classifier's ability to make the best decision depends on the success of the training process to separate the test data [4]. However, the training process is highly dependent on the size and nature of the data being generated.

In the imbalanced data set, the ML classifier in the training phase faces difficulty as the number of majority class samples is too large. This makes the classifier tend to the majority class, which leads to less accuracy in classifying minority class samples. Data imbalance is present in most realistic data, and the greater the data imbalance rate, the more difficult the learning process from the data set. The imbalanced datasets are skewed and distributed among the data classes. The severity of the uneven distribution varies from a very small degree that does not affect the performance to a very large degree that hurts performance [6].

However, the traditional classification algorithms still have challenges of reducing the classification accuracy by imbalanced data problems, which leads to poor performance of the model in practical applications. It is a essential to rebalanced the data set to solve this problem by several ways such as:

- Resampling techniques: This involves over-sampling, synthetic sample generation, and under-sampling techniques, all those techniques used to improve the precision and resilience of classification algorithm.
- Data collection and preprocessing: The distribution of samples should be taken into account, and efforts should be made to guarantee that the dataset is balanced throughout the data-gathering stage. To eliminate superfluous or unnecessary features and enhance model performance, apply techniques like feature extraction and feature selection [21].

IV. Feature selection

The performance of ML classifier depends mainly on the volume and nature of the training data. The ML performance was increased when the dataset features are large, this is not absolutely, as it was often reached a certain limit after that it achieves a negative affect and will be a significant drop in performance. That is, reducing the dimensions leads to an improvement in machine learning performance, in addition to reducing the time complexity and storage problems, Algorithm1expleaned the process of feature selection. Reducing dimensions is one of the most important steps of pre-processing, in which features that do not affect the quality of machine learning are eliminated and the appropriate number of features that must be kept are determined, and thus we get rid of features that negatively affect the accuracy of performance [12] [22]. Feature selection was adapted to select features that guide more significant separability between classes [23]. The most reasonable features are chosen depends on their contribution to the increased performance of the model in the final results [12]. Feature selection techniques were aimed at selecting a subset of the original features that are more efficient to perform machine learning. The subset of features to be selected should be able to characterize all data elements and not lead to loss of information as in the full set or better. Methods for selecting features vary according to the nature of the data to be processed. Traditional methods for selecting features can be divided into three types, the first type includes filtering methods, which calculate the importance of each feature, the second type includes methods of wrapper approaches, and here a classifier is used to create the best set of features in an iterative procedure to reduce the number of original features, while the third type is embedded methods and it depends on selecting a specific classifier and making a selection of implicit features during its training process [5][10][11]. The combinations of those types in hybrid techniques are also available [24].

Algorithm 1 The process of feature selection

Input : X ,the original dataset

Output : *FS*

Start

While *the Termination condition is not met* ***do***

Generate FS the selected features subset from all features using selecting approaches

Evaluate the FS subset using evaluation measures;

End While

End

The process of feature selection was explained as a kind of learning approach aiming to collect the features and find the appropriate parameters as shown in Fig.1. The FS subset can be generated by selecting strategies such as the random search strategy, the stepwise addition or deletion of features, and heuristic search methods. In the next step, after FS subset obtained, the performance of it must be calculated [25].

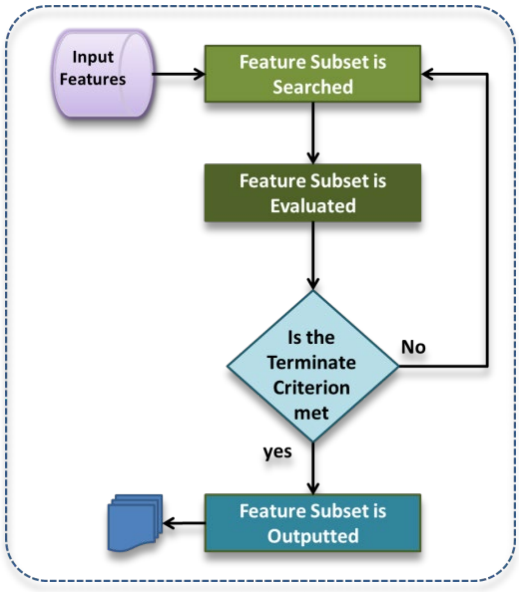

Fig. 1. Flowchart of feature selection [25]

The filter approach chooses the highest-ranking features based on a statistical or information measure, such as information gain and gain ratio, and evaluates the significance of features by focusing primarily on the intrinsic properties of the data. Filter-based selection has two shortcomings: it disregards the connections between features, and it is independent of the classifier, which can result in worse classification accuracy with certain classifiers. and second, it disregards the features' interdependencies [18].

The information dependency should be examined in performing variable selection, often powerful related features are similar. When selecting features, the related features could be crucial, particularly in high-dimensional environments. A wrapper technique like a genetic algorithm searches the space of all feature subsets by encircling the classification model with a search method. The wrapping method does have one clear problem, though: as the number of features increases, so does the number of subsets from the feature space, making it computationally costly. The embedding method is searching for the best key features while classifier building.

A. Information Gain (IG)

To create a superior prediction model, less beneficial qualities could be eliminated through effective feature selection. Aside from that, features unrelated to the target variable must be eliminated because they may raise the computational cost and hinder the model's ability to function at its best. The information gain (IG) technique is used in this study to extract the best features. IG is a filter-based feature selection that calculates the predictor variable's ability to classify the dependent variable.

The IG technique, which determines the statistical dependence between two variables, has its origins in information theory. The IG between two variables, X and Y, can be expressed mathematically as in (1).

$$IG(X|Y) = H(X) - H(X|Y), \quad (1)$$

where the conditional entropy for X given Y is represented by $H(X|Y)$, and the entropy for variable X is denoted by $H(X)$. The process of determining the IG value for an attribute is taking the target variable's entropy for the entire dataset and deducting the conditional entropies for each possible attribute value. Additionally, the conditional entropy $H(X|Y)$ and entropy $H(X)$ are calculated as in (2) and (3).

$$H(X) = -\sum_{x\in X} P(x)\log_2(x), \quad (2)$$

$$H(X|Y) = -\sum_{x\in X} P(x)\sum_{y\in Y} P(x|y)\log_2\big(P(x|y)\big), \quad (3)$$

Therefore, if $IG(X|Y) > IG(Z|Y)$, then given two variables, X and Z, a given variable Y is said to have a higher substantial correlation to X than Z. Moreover, IG evaluates each attribute separately, determines its gain information, and determines how relevant it is to the target variable [19]. An attribute selection method called information gain was used to rank each feature from highest to lowest value. Its value was determined by subtracting the entropy of each criterion from the overall entropy for all criteria on the feature. The better the diversity in the data, the greater the entropy value. The purpose of the value measurement was to determine whether or not to employ particular attributes. When assessing the individual features' gains to the goal variable, IG is a reliable algorithm. The classifiers are trained using the attributes whose IG values are above a specific threshold, and the attributes with the lowest IG values are eliminated [19].

B. 4.2 Permutation-Based Features Importance

(PBFI)

The PBFI measures the importance of a feature based on permutation. The fundamental idea is to permute a particular feature from the prediction process at random, and then determine the degree to which this impacts the model's performance by computing the prediction error. A feature is significant if its exclusion causes the model's performance to suffer (an increase in model error). It should be highlighted that a feature's significance is not determined by depending just on it but rather by considering it in conjunction with the other elements. Breiman [26] proposed the PBFI measurement for random forests, and numerous researchers built on this concept to construct more sophisticated feature importance approaches in the years that followed [26-27].

One of the key issues of this method is that the error ratio which can be observed in algorithm 2 [28], is utilized to evaluate the significance of the feature rather than the error difference. Another crucial aspect of this strategy to keep in mind is that the existence of correlation features has a significant impact on it. Correlation or dependent features are equally important, and the performance of the model will be significantly impacted when one of these features is changed. Additionally, because the method depends on splitting the number of available samples into training data and validation data when determining the relevance of the features, PBFI produces unstable results when the amount of data is small [29]. To improve the findings' stability and lessen the random impacts of the permutation process, it is feasible to repeat the feature permutation process. However, this adds to the process' computational and temporal complexity [27].

Algorithm 2 The PBFI algorithm

Input : Training model f, feature matrix X, target vector Y, error measure E(y, f)
Output : Return sorting FI
Start :

Estimate the original model error $E_{original} = L(y, f)$

For *each feature* ***j*** *in a dataset M features* ***do***

Generate feature matrix X_{prem} *by permuting feature* ***j*** *in the data X.* …(this breaks the association between feature j and true outcome y)

Estimate error $E_{prem} = L(Y, f(X_{prem}))$ …("Based on the predictions of the permuted data")

Calculate permutation feature importance as quotient $FI_j = E_{prem} / E_{original}$ *or difference* $FI_j = E_{prem} - E_{original}$

End For

Sort features by descending FI

END

V. Ensembel Learning

Various techniques are used to design ensemble learning algorithms by thinking of a weak learning algorithm to create one stronger [30]. The essential concept of a learner in the ensemble classifier is to form multiple classifiers that deal with the original data and later collect the prediction results of those classifiers to classify unknown samples. An ensemble learning model considers several methods: Voting, Stacking, Boosting, and Bagging [31]. Voting is a straightforward process for classification or Averaging process for regression. A voting classifier is a method for predicting the final output class label based on the maximum vote majority [32-33]. Staking is an ensemble learning technique used to merge many models via a meta-model [34]. Staked generalization uses two different models: basic models at level 0 and meta-models at level 1. Meta models get learning from basic model outputs [35]. As a result, layered learning performs better in terms of prediction outcome accuracy than the best base model [36]
Boosting and bagging methods decrease the variance and provide more stability when a model is designed. Boosting is an integrated learning method that concatenates different weak classifiers sequentially. Depending on the results predicted by the previous classifier, the training of the next classifier is done and gets the results. Boosting does not consider the structure of the particular weak classifier models themselves but rather manages the training data set and the association method to receive smaller errors. Common Boosting methods are Adaptive Boost (AdaB), Gradient Boosting Decision Tree (GBDT), and Histogram Gradient Boosting (HistGB). Boosting needs bootstrapping, which means that some samples will be run more frequently than others. Bagging emanates from the words bootstrap aggregating and it was a technique that approach enables stabilizing classifiers by training a classifier on multiple bootstrap samples of the training data set. The multiple instances of the identical classifier only differ in their training sets. Depending on sampling with replacement operation, a bootstrap sample likely includes repeated samples, this improves the predictive power of the base learner involved with those samples. Bagging utilizes and integrates several learners using an averaging technique to reduce variation and bias and these approaches give fine work results. This may show that a low learner can transform into a powerful learner and also, this solution may be slightly better than guessing randomly [4][7][26]. Random forests (RF) and extra trees(ExtraC) are two examples of bagging ensembles [31].

VI. Methodology and discussion

This research paper applies a proposed composite feature selection approach to an imbalanced training dataset in a classification model. This feature selection approach is merging Flitter and Permutation based on Ensemble Learning feature selection methods. This approach selects the best features (to determine the importance of features) to improve the performance of machine learning classification.
The general structure of this proposed model is shown in Fig.

2.

The proposed composite feature selection method was merge filter and PBFI methods. These feature selection methods were chosen after implementing several studies on filter methods and applying multiple bagging and boosting ensemble classifiers with PBFI to choose the best methods which achieve high performance.

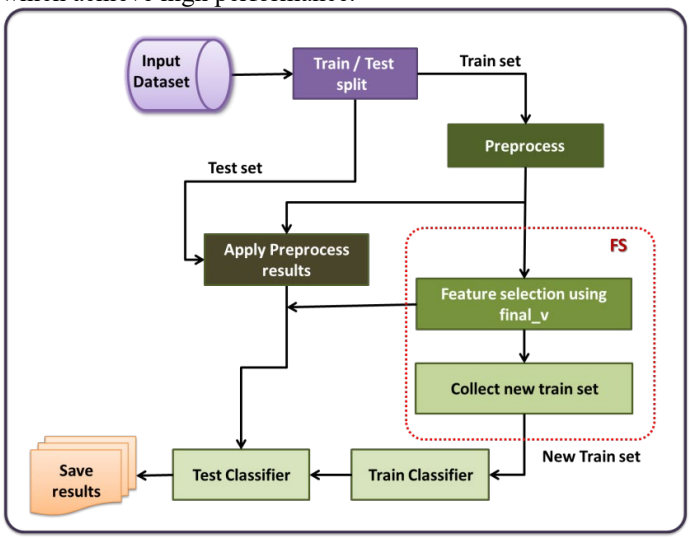

Fig. 2. Proposed Feature selection.

A. Dataset

This research applies a proposed feature selection approach to the imbalanced training dataset before classification. The best features were selected to improve the performance of machine learning classification. Five datasets are used from the dataset Kaggle repository, Table 1 explains the number of attributes, the size, and the imbalanced ratio (Ir) of each one of them.

Table 1. Datasets used.

Dataset	No. Attributes	No. samples	IR
Madelon[37]	500	1729	2.83
Covtype[38]	54	581012	32.5
Colon cancer[39]	2000	62	1.87
Creditcard[40]	30	284807	592.35
Higgs8[41]	28	6629122	7.29

B. Initialization steps

The first process in the initialization step is calculate the size of the dataset, and then calculate the imbalanced ratio Ir in the dataset as in (4).

$$Ir = All\_Maj_{size} / All\_Min_{size} \quad (4)$$

Where:

All_Maj_{size} = *number of majority class samples,*

All_Min_{size} = *number of minority class samples.*

Which are calculated from the label present with dataset.

Then the dataset was split into two groups, 70% was used for train data (TrainSet) and 30% was adopted for test data (TestSet). The correlation coefficient is about 0.9. was used to identify the correlation features in the TrainSet. The final process in this step is to separate the TrainSet according to the actual output value (Target) into two groups: the first is the majority data set (MajSet), which is the class that includes the largest number of data instances, and the minority class set (MinSet), which is the group that includes the minor number of data instances, as shown in (5) and (6).

$$MajSet = \{ instance \in TrainSet, if\ instance\ [Target] = 0\} \quad (5)$$

$$MinSet = \{ instance \in TrainSet, if\ instance\ [Target] = 1\} \quad (6)$$

C. Proposed Feature Selection

This stage consists of merging two types of feature selection techniques: filter feature selection and permutation-based feature importance (PBFI). The model as shown in Fig. 3. produces three vectors (V1, V2, and V3) representing the importance of the features in the dataset and computes the final result (final_v) by combining the results of the previous three vectors. The V1 vector is constructed by using a statistical method called information gain (IG) feature selection. The features are arranged in descending order relative to the rank value from highest to least important.

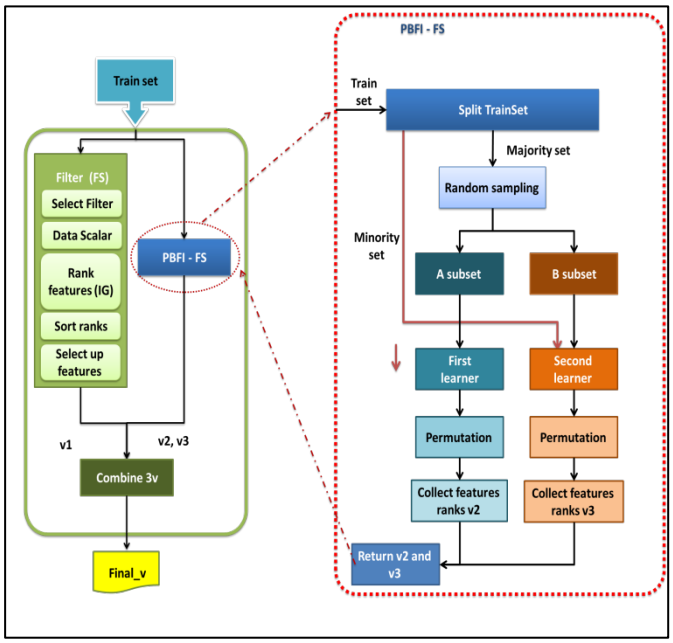

Fig. 3. The Composite Feature Selection steps.

The V2 and V3 vectors are constructed by using the PBFI feature selection method depending on two subsets of training data and two learners. The first learner is used to determine the importance of the features on the first subset and produce vector V2. The second learner is used to determine the importance of the features on the second subset and produce vector V3. In other words, the second and third vector production methods share a sequence of sub-processes and differ in the machine learning method used to estimate the importance of features and the subset used in each iteration. The F-score and AUC measurements were chosen to

determine the importance of features in the PBFI method. The size of important features in V1, V2, and V3 depends on the user specifies with the feature ratio variable *fr* (the set of features with the highest importance will be chosen to represent each vector). The value of fr is depending on the dataset size and the imbalanced ratio of it. The final vector final_v collected from the results of the three vectors (V1, V2, and V3) in addition to the *fr%* value of the actual number of total features as shown in (7).

$$final_v = (V1[bfr] \cap V2[bfr] \cap V3[bfr]) \quad (7)$$

Where:

$$bfr = number\ of\ features\ in\ TrainSet \times fr\%$$

Two different methods of composed feature selection were suggested to determine the importance of features in this research using the information gain filter method to select features to produce the first vector V1. The two suggested feature selection methods are different in producing V2 and V3: the first is called the bag2g method based on bagging learning (RF is used as the first learner, and ExtraC is used as the second learner). The second suggested method is called the boost2g method based on boosting learning (AdaB as a first learner, and HistGB as a second learner).

VII. Results and Discussion

After determining the final feature vector, feature selection is applied to the complete training data set to obtain only the data with important features. To compare the two methods (bag2g and boos2g) in classification of imbalanced big dataset, the classifier is trained using the original training data and tested using the original test data at first. Then, the same classifier is trained using the newTrainSet resulting from the two proposed methods and tested using the original test data. The classifier performance is calculated using AUC, F-score and g-mean metrics in all situations.

Three classifiers - Decision Tree (DT), K-nearest neighbors (KNN), and Gaussian Naïve-Bayes (GNB)- were selected in this research and collected results at each of the benchmark datasets. To obtain greater accuracy in the results for each experiment carried out on one of the classifiers, the classification experiments were re-applied five times using different random-state values in a dataset splitting and the average of the results that were implemented for each dataset was taken. The average results collected from the experiments when applying all classifiers are shown in the following figures.

Fig.4 to Fig. 8 show the performance results of applying the two proposed methods bag2g and boost2g on Madelon, covtype6, colon-cancer, creditcard, higgs8 datasets respectively. The classification results were compared to the original data (Base.data) when using DT classifier. Each figure represents DT classification performance using three performance metrics (ROC-AUC, F-scoe, and g-mean) with feature rates *fr* in [30%, 40%, 50%, 60%, or 70%].

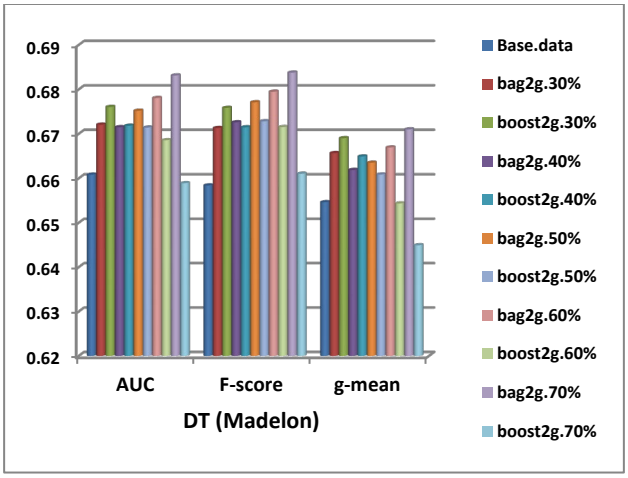

Fig 4. DT classification results on Madelon dataset.

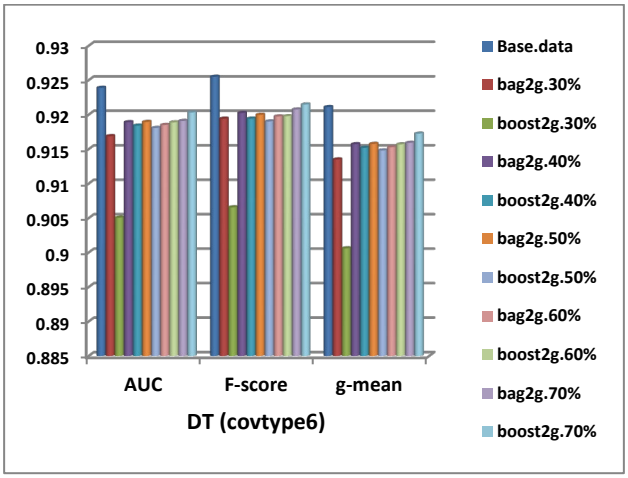

Fig 5. DT classification results on covtype6 dataset.

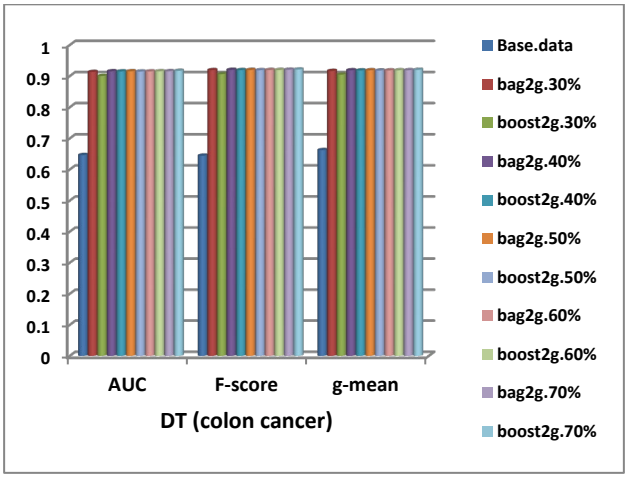

Fig 6. DT classification results on colon-cancer dataset.

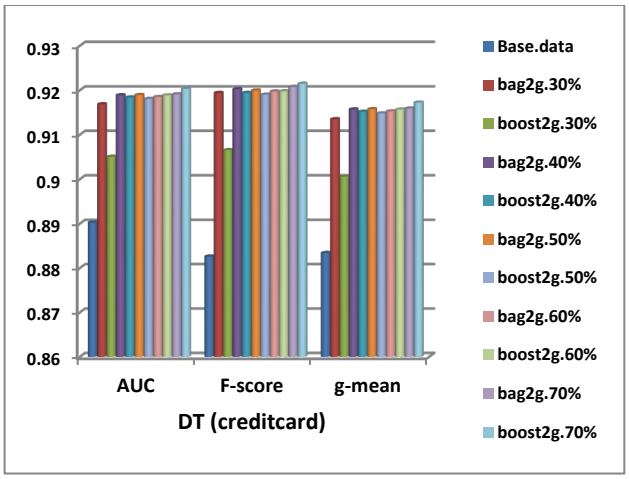

Fig 7. DT classification results on creditcard dataset.

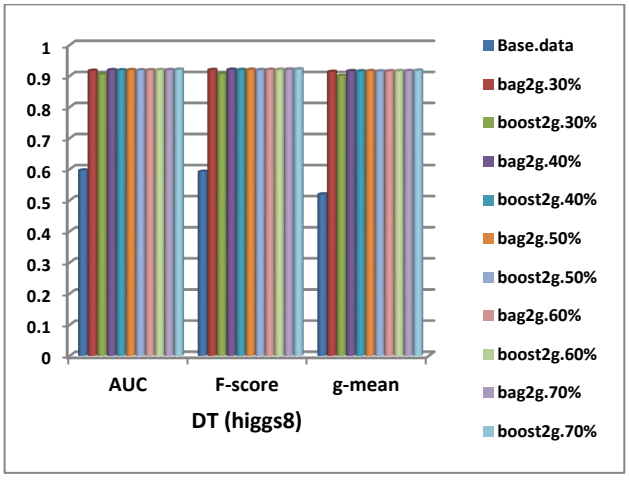

Fig 8. DT classification results on higgs8 dataset.

The DT results show that the performance was increased in the three metrics when using bag2g and boost2g methods over base data in four out of five datasets, the three metrics were increased in the same level except in covtype6 dataset where there is decreased. The two feature selection methods have the same effects in colon cancer, covtype6, creditcard, and higgs8 datasets. While they have different effects in Madelon dataset, that was shown when using bag2g has more positive effect in the three metrics.

In GNB classifier, Fig. 9 to Fig. 13, the results were affected in small amounts when using bag2g and boost2g feature selection methods before classification process in all datasets with the three metrics.

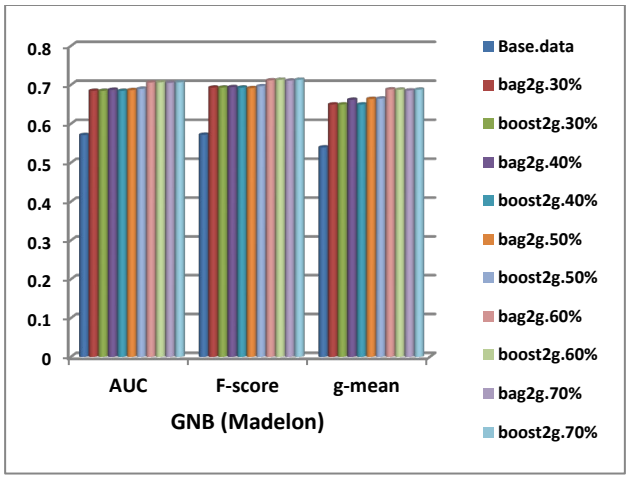

Fig 9. GNB classification results on Madelon dataset.

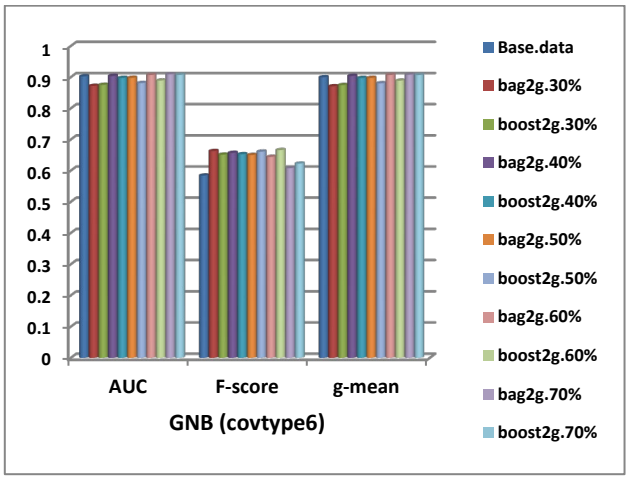

Fig 10. GNB classification results on covtype6 dataset.

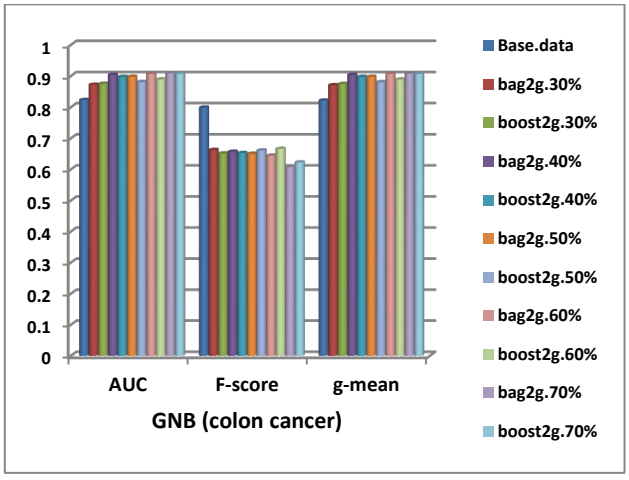

Fig 11. GNB classification results on colon-cancer dataset.

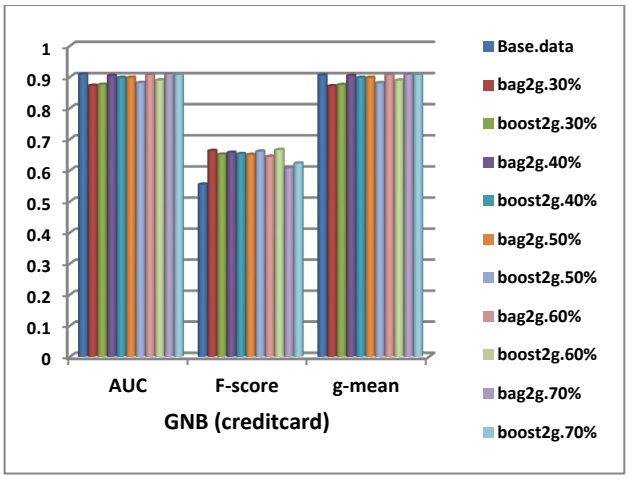

Fig 12. GNB classification results on creditcard dataset.

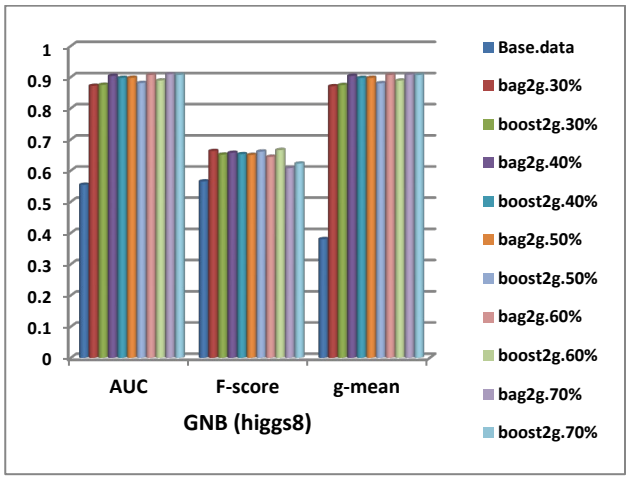

Fig 13. GNB classification results on higgs8 dataset.

Madelon, covtype6, and creditcard have the same increased amount level in the three metrics. In higgs8 dataset, the results show that was a high increased in AUC and g-mean, and small increased in F-score. The increased in AUC and g-mean is very small in colon-cancer, and it has a negative effect in F-score.

The Fig. 14 to Fig. 18 show KNN results when classifying Base data and the data after feature selection process for Madelon, covtype6, colon-cancer, creditcard, higgs8 datasets respectively. It was clear that the two proposed methods made a dataset more perfect in classification process. They increased the performance to a high level in Madelon, colon-cancer, creditcard, higgs8 datasets, only the covtype6 dataset has a negative effect when using feature selecting before classification by KNN.

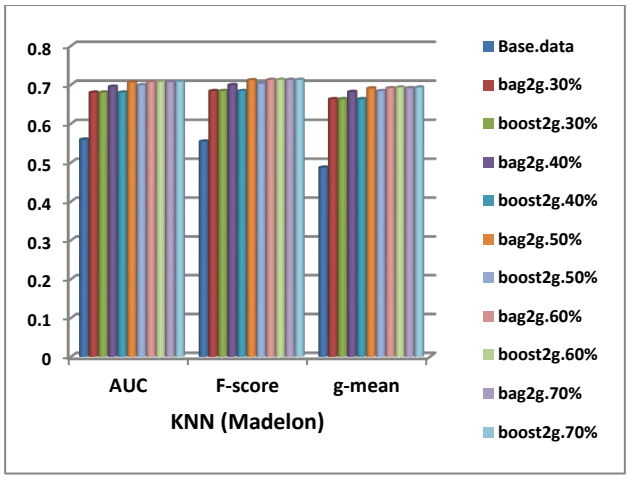

Fig 14. KNN classification results on Madelon dataset.

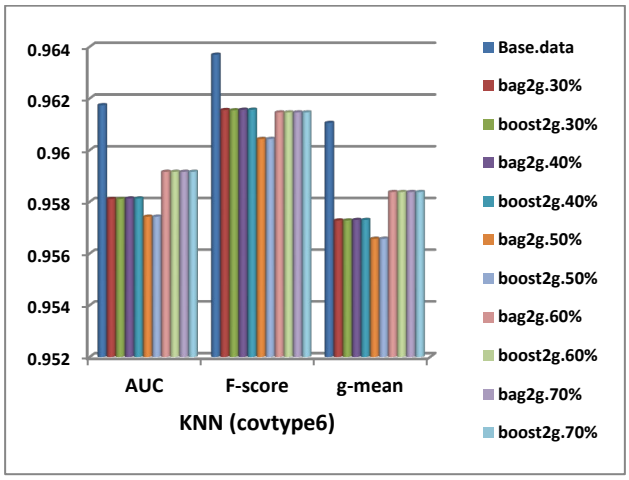

Fig 15. KNN classification results on covtype6 dataset.

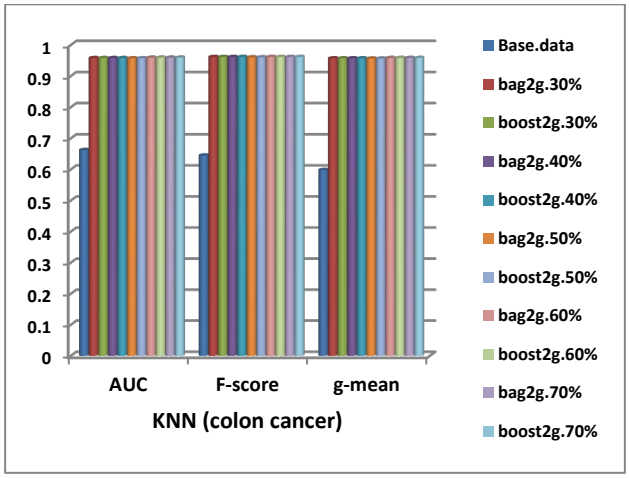

Fig 16. KNN classification results on colon-cancer dataset.

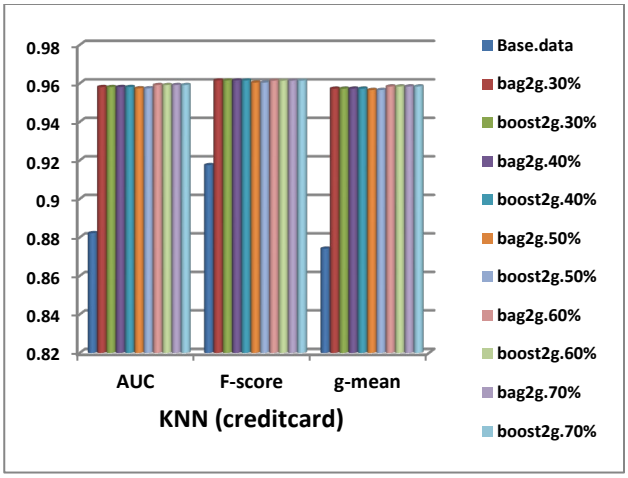

Fig 17. KNN classification results on creditcard dataset.

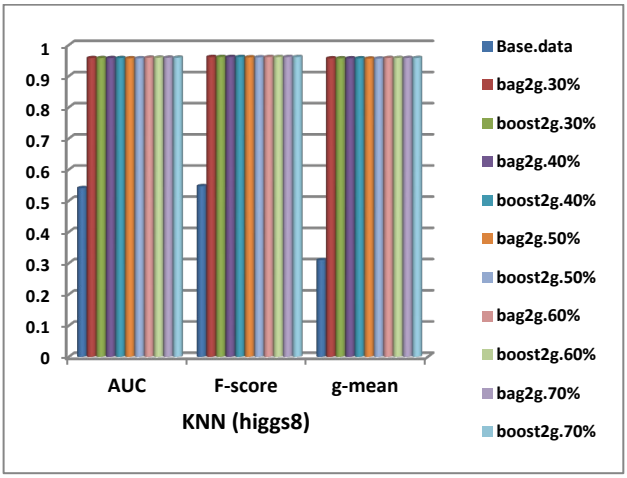

Fig 18. KNN classification results on higgs8 dataset

The three metrics were affected by the same level in all datasets, this mean that the feature selection process have stability effect in data while making data in small size by eliminate features which have minimum scores.

From other side, calculating the average improvements for all fr% values in the two proposed feature selection methods and compared them in each metrics to explain the effect of them in each classifier are shown in the Tables 2 to 4.

From Table 2, it appears that the improvement average of AUC obtained using bag2g is greater than using the boost2g method, and this appears clearly when using the three classification algorithms reaches (0.125, 0.105, 0.185). But in the same time, this greatest in very small in all datasets.

On Table 3, which gives the average of improvement in F-score that was obtained by the three classifiers for all datasets. The difference between using the two-feature selection method before classification is very small also. The average of all datasets F-score shows that DT and KNN classifiers have greater improvement when using bag2g, while GNB classifier has greater improvement when using boost2g method.

Table 2. Classification improvement in auc results.

Dataset	FS + DT		FS + GNB		FS + KNN	
	bag2g	*boost2g*	*bag2g*	*boost2g*	*bag2g*	*boost2g*
Madelon	0.0152	0.0085	0.1229	0.1234	0.1397	0.1356
Covtype6	-0.0054	-0.0077	-0.0045	-0.0124	-0.0033	-0.0033
Colon cancer	0.2687	0.2662	0.0748	0.0669	0.2958	0.2958
Creditcard	0.0282	0.0259	-0.0102	-0.0181	0.0762	0.0762
Higgs8	0.3225	0.3201	0.3436	0.3357	0.4177	0.4177
The average rate	**0.1258**	0.1226	**0.1053**	0.0991	**0.1852**	0.1844

Table 3. Classification improvement in f-score results.

Dataset	FS + DT		FS + GNB		FS + KNN	
	bag2g	*boost2g*	*bag2g*	*boost2g*	*bag2g*	*boost2g*
Madelon	0.0184	0.0121	0.1284	0.1299	0.1496	0.1449
Covtype6	-0.0055	-0.0082	0.0602	0.0665	-0.0024	-0.0024
Colon cancer	0.2756	0.2728	-0.1544	-0.1481	0.3162	0.3162
Creditcard	0.0374	0.0346	0.0894	0.0958	0.0437	0.0437
Higgs8	0.3284	0.3256	0.0790	0.0854	0.4143	0.4143
The average rate	**0.1309**	0.1274	0.0405	**0.0459**	**0.1843**	0.1833

Table 4 shows the average classification results in g-mean metrics, and it clearly shows the superiority of the bag2g method, as it gave a greater improvement when using the three classifiers. Furthermore, the different in improvement was small in all datasets when using the GNB classifier.

Table 4. Classification improvement in g-mean results.

Dataset	FS + DT		FS + GNB		FS + KNN	
	bag2g	*boost2g*	*bag2g*	*boost2g*	*bag2g*	*boost2g*
Madelon	0.0112	0.0042	0.1305	0.1286	0.1963	0.1918
Covtype6	-0.0059	-0.0084	-0.0023	-0.0101	-0.0035	-0.0035
Colon cancer	0.2560	0.2537	0.0760	0.0682	0.3591	0.3591
Creditcard	0.0318	0.0293	-0.0080	-0.0158	0.0834	0.0834
Higgs8	0.3964	0.3939	0.5172	0.5094	0.6475	0.6475
The average rate	**0.1379**	0.1345	**0.1427**	0.1360	**0.2566**	0.2557

A negative value appeared in Tables 2-4 indicates that there was inverse effect on the classification performance when using the proposed feature selection methods before

classification process, and also those value is very little in respect to the decreased in data size when eliminate unimportant features. This most reduce the computation cost in training process when using DT and GNB, and testing process when using KNN.

To assert the results obtained previously, the performance of the proposed method was compared with the methods mentioned in [15] and [16] using the same metrics, when classifying the Colon-cancer dataset by KNN, as shown in Table 5. The comparison shows that the proposed model (using composite feature selection) achieves high performance in G-mean, Accuracy, and AUC over the two previous studies, and it has small decreased in sensitivity against the two studies by 0.05% and 1.28%.

Table 5. Comparison with previous studies.

Model	G-mean	Sensitivity	Accuracy	AUC
[15] IWSSr-SFLA	-	95.87%	94.50%	-
[16] rCBR	83.9%	97.1%	-	84.9
Proposed models	95.76%	95.82%	99.65%	95.84

VIII. Conclusion

Using feature selection methods improves the performance of machine learning classifiers at different feature selection ratios when classifying imbalanced data. The reduction in data size as a result of feature selection can be considered as an improvement even if there is no change in the performance metrics values.

In the research, a composite feature selection method was proposed that combining filtering methods and PBFI methods to produce a combined method to improve the performance of classification process on imbalanced data. The proposed idea adopted the use of two learning methods when selecting features using PBFI method, first method is bag2g which adopted the bagging machine learning methods, and second method used boosting machine learning methods. Applying the proposed feature selection methods before classification process show that, the results were still in stability level when eliminate features with minimum score. Furthermore, it is noted that the two methods bag2g and boost2g have the same effect with very little variance at different feature selection ratios and on different datasets with different specifications in terms of degree of imbalance, size and number of features, at different standards for measuring performance. For the most datasets used in this research, the bag2g feature selection method was gave more improvement amount in classification results over boost2g method. To make further improvements to the proposed algorithm in the future, other bagging and boosting learners can be proposed in PBFI and the difference with the current method can be studied. Filter methods can also be used to select new features instead of information gain.

Acknowledgement

We extend our thanks to the University of Mosul, as well as our Colleges, the College of Computer Science and Mathematics and the College of Education for Pure Science to support this report.

References

[1] K. C. Maurya, "Anomaly Detection in Big Data," Ph.D. Thesis, India Institute of Technology Roorkee, 2016.

[2] R. Pereira and M. Pereira, "Challenges, Open Research issues and Tools in Big Data Analytics Covid-19", International Journal for Research in Applied Science & Engineering Technology. 2022, pp. 2649- 2659. https://doi.org/10.22214/ijraset.2022.41820

[3] J.Wang, C. Xu, J. Zhang and R. Zhong, "Big data analytics for intelligent manufacturing systems: A review", Journal of Manufacturing Systems. 2022, pp. 738-752 .https://doi.org/10.1016/j.jmsy.2021.03.005

[4] BEJ, Saptarshi, "Improved imbalanced classification through convex space learning," PhD Thesis, Rostock University, 2021.

[5] Meoni, Marco. "Mining Predictive Models for Big Data Placement," PhD Thesis, Pisa University, 2018. https://cds.cern.ch/record/2647981/files/TS2018_030_2.pdf

[6] A. Desai and S. Chaudhary, "Distributed adaboost extensions for cost-sensitive classification problems", Int. J. Computer. 2018, pp.1-8. http://dx.doi.org/10.5120/ijca2019919531

[7] Leevy, Joffrey. "Machine Learning Algorithms for Predicting Botnet Attacks in IoT Networks," PhD Thesis, Florida Atlantic University, 2022.

[8] Gao, Yang, et al., "Enhancing Classification and Retrieval Performance by Mining Semantic Similarity Relation from Data," PhD Thesis, University of Texas at Dalla, 2021.

[9] M. S. Mahmud, , J. Z. Huang, .S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis", Big Data Min and Analytics. 2020, pp. 85-101. https://doi.org/10.26599/BDMA.2019.9020015

[10] Spelmen, Vimalraj S., and R. Porkodi. "A review on handling imbalanced data." 2018 international conference on current trends towards converging technologies (ICCTCT). IEEE, 2018. https://www.researchgate.net/publication/329584489_A_Review_on_Handling_Imbalanced_Data

[11] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," Pattern Recognition, vol. 124, pp. 108511, 2022. DOI.org/10.1016/j.patcog.2021.108511

[12] M. Fattahi, M. H. Moattar, and Y. Forghani, "Improved cost-sensitive representation of data for solving the imbalanced big data classification problem," Journal of Big Data, vol. 9.1, pp. 1-24, 2022. DOI.org/10.1186/s40537-022-00617-z

[13] Liu, Haoyue, MengChu Zhou, and Qing Liu. "An embedded feature selection method for imbalanced data classification." IEEE/CAA Journal of Automatica Sinica , vol. 6, no. 3 ,p. 703-715, 2019. http://dx.doi.org/10.1109/JAS.2019.1911447

[14] Thaher, T., Mafarja, M., Abdalhaq, B., & Chantar, H. "Wrapper-based feature selection for imbalanced data using binary queuing search algorithm". In 2019 2nd international conference on new trends in computing sciences (ICTCS) (pp. 1-6). IEEE.2019. http://dx.doi.org/10.1109/ICTCS.2019.8923039

[15] Pirgazi, J., Alimoradi, M., Esmaeili Abharian, T., and Olyaee, M. H."An Efficient hybrid filter-wrapper metaheuristic-based gene selection

method for high dimensional datasets". Scientific reports, vol. 9, no. 1, p. 18580, 2019. DOI: 10.1038/s41598-019-54987-1

[16] Abdulrauf Sharifai, G., and Zainol, Z. ,"Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm". Genes, vol.11, no. 7, p. 717, 2020. http://dx.doi.org/10.3390/genes11070717

[17] Namous, F., Faris, H., Heidari, A. A., Khalafat, M., Alkhawaldeh, R. S., & Ghatasheh, N. ,"Evolutionary and swarm-based feature selection for imbalanced data classification". Evolutionary Machine Learning Techniques: Algorithms and Applications, p. 231-250, 2020. http://dx.doi.org/10.1007/978-981-32-9990-0_11

[18] Fu, G. H., Wu, Y. J., Zong, M. J., and Pan, J. "Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data". BMC bioinformatics, vol. 21, no. 1, p. 1-14. 2020.https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3411-3

[19] Ebiaredoh-Mienye, S. A., Swart, T. G., Esenogho, E., & Mienye, I. D. "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease". Bioengineering, vol. 9, no. 8, p. 350, 2022. http://dx.doi.org/10.3390/bioengineering9080350

[20] Fahrudy, Dony, Uyun, Shofwatul. "Classification of Student Graduation using Naïve Bayes by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection". JOIV: International Journal on Informatics Visualization, vol. 6, no. 4, p. 798-808, 2022. https://doi.org/10.30630/joiv.6.4.982

[21] Quan, Li; Gong, Tao; Jiang, Kaida. "Denying Evolution Resampling: An Improved Method for Feature Selection on Imbalanced Data". Electronics, vol. 12, no.15, p. 3212, 2023. http://dx.doi.org/10.3390/electronics12153212

[22] , Hu, Zhigang. "Development of a Machine Learning-Based Financial Risk Control System". PhD Thesis. Utah State University, 2022.

[23] Bekkar, Mohamed; Alitouche, Taklit Akrouf. "Imbalanced data learning approaches review". International Journal of Data Mining & Knowledge Management Process, vol. 3.4, no. 15, 2013.

[24] Ebenuwa, S. H., Sharif, M. S., Al-Nemrat, A., Al-Bayatti, A. H., Alalwan, N., Alzahrani, A. I., and Alfarraj, O. "Variance ranking for multi-classed imbalanced datasets: A case study of One-Versus-All". Symmetry, vol. 11, no. 12, p. 1504, 2019. http://dx.doi.org/10.3390/sym11121504

[25] Rong, Miao; Gong, Dunwei; Gao, Xiaozhi. "Feature selection and its use in big data: challenges, methods, and trends". IEEE Access, vol. 7, p. 19709-19725, 2019. http://dx.doi.org/10.1109/ACCESS.2019.2894366

[26] Lalchand, Vidhi. "A meta-algorithm for classification using random recursive tree ensembles: A high energy physics application". 2001, no.06880, 2020.

[27] Pathan, Muhammad Salman, Nag, Avishek, Pathan, Muhammad Mohisn, Dev, Soumyabrata." Analyzing the impact of feature selection on the accuracy of heart disease prediction". Healthcare Analytics, vol. 2, p. 100060, 2022. https://doi.org/10.1016/j.health.2022.100060

[28] Fisher, Aaron; Rudin, Cynthia; Dominci, Francesca. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". J. Mach. Learn. Res., vol. 20, no.177, p. 1-81,2019.

[29] Kaneko, Hiromasa. "Cross-validated permutation feature importance considering correlation between features". Analytical Science Advances, vol. 3, no. 9-10, p. 278-287, 2022. https://doi.org/10.1002/ansa.202200018

[30] Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., and Mushtaq, M. F. "Imbalance class problems in data mining: A review". Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no. 3, p. 1560-1571, 2019. http://dx.doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

[31] Meshoul, S., Batouche, A., Shaiba, H., and AlBinali, S.. "Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping". Mathematics, vol. 10, no. 22, p. 4271, 2022. https://doi.org/10.3390/math10224271

[32] Shen, Y., Zheng, K., Yang, Y., Liu, S., and Huang, M. "CBA-CLSVE: A Class-Level Soft-Voting Ensemble Based on the Chaos Bat Algorithm for Intrusion Detection". Applied Sciences, vol. 12, no. 21, p. 11298. 2022. https://doi.org/10.3390/app122111298

[33] Chatterjee, Subhajit; Byun, Yung-Cheol. "Voting Ensemble Approach for Enhancing Alzheimer's Disease Classification". Sensors, vol. 22, no.19, p. 7661, 2022. https://doi.org/10.3390/s22197661

[34] Z. Ma, P. Wang, Z. Gao, R. Wang, and K. Khalighi, "Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose," PLoS ONE, vol. 13, no. 10, pp. 1–12, 2018. https://doi.org/10.1371/journal.pone.0205872

[35] Sikora R. and Al-laymoun O.H., "A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms", Journal of International Technology and Information Management. vol.23, No.1, pp.1-12, 2014. https://doi.org/10.4018/978-1-4666-7272-7.ch004

[36] Chanamarn, Nipaporn; Tamee, Kreangsak; Sittidech, Punnee. "Stacking technique for academic achievement prediction". Int. Work. Smart Info-Media Syst. Asia (SISA 2016), no. Sisa, 2016, 2016: 14-17.

[37] Kaggle Machine Learning and Data-Science community(Madelon dataset). Online Available: https://www.kaggle.com/datasets/sayroy1997/madelon

[38] Kaggle Machine Learning and Data-Science community(Covtype-data). Online Available: https://www.kaggle.com/datasets/ilginkarakas/covtype

[39] Kaggle Machine Learning and Data-Science community(colon-cancer-gene-expression). Online Available:https://www.kaggle.com/datasets/masudur/colon-cancer-gene-expression-data

[40] Kaggle Machine Learning and Data-Science community(Credit-card Dataset). Online Available: https://www.kaggle.com/datasets/jacklizhi/creditcard

[41] Kaggle Machine Learning and Data-Science community(HIGGS UCI DATASET). Online Available:https://www.kaggle.com/datasets/erikbiswas/higgs-uci-dataset