Journal of Theoretical and Applied Information Technology
30th June 2021. Vol.99. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

USING MACHINE LEARNING MODELS TO COMPARE VARIOUS RESAMPLING METHODS IN PREDICTING INSURANCE FRAUD

MOHAMED HANAFY[1,2], RUIXING MING[1]

[1] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China

[2] Department of Statistics, Mathematics, and Insurance, Faculty of commerce, Assuit University, Asyut, Egypt

E-mail: [1] mhanafy@commerce.aun.edu.eg, [2] ruixingming@aliyun.com

ABSTRACT

One of the most common types of fraudulent is insurance fraud. And in particular fraud in automobile insurance, the cost of automobile insurance fraud is substantial for property insurance companies and has a long-term impact on insurance firms' pricing strategies. And In order to minimize insurance rates, car insurance fraud detection has become necessary. Although predictive models for the detection of insurance fraud are in active use in practice, there are relatively few documented studies on the use of machine learning approaches to detect insurance fraud, likely due to the lack of available data. In this paper, by using real-life data, we evaluate 13 machine learning approaches. And Because of the imbalanced datasets in this area, predicting insurance fraud has become a significant challenge. Due to our data consist mostly of a "non-fraud claims " class with a small percentage of "fraud claims. " Thus that the prediction of fraud appears weakly with classification models; therefore, the present study seeks to suggest an approach that enhances machine learning algorithms' results by using resampling techniques, such as Random Over Sampler, Random under Sampler, and hybrid methods, to address the issue of unbalanced data. And we compare between them. This paper shows that after using resampling techniques, the efficiency of all ML classifiers is enhanced. Furthermore, the findings confirm that there is no one resampling method that overall outperforms. Besides, among all the other models, the Stochastic Gradient Boosting classifier obtained the best result when using the hybrid resampling technique.

Keywords: *Automobile Insurance; Insurance Fraud; Fraud Detection, Classification; Machine Learning; Imbalanced Data; Resampling Methods.*

1. INTRODUCTION

The insurance industry is crucial in ensuring that individuals, companies, and organizations are safe from financial risks. And several insurance firms promote the growth of different industries [1]. Yet, insurance fraud has become a major threat to the insurance industry's viability [2]. In this article; We'll deal with car insurance fraud. And Car insurance is a form of mobility insurance. And this kind of insurance has grown to be a significant industry linked to the world economy's expansion. And many citizens' livelihoods; With growing trust in the insurance industry's positive growth, more capital will enter the insurance sector. For that reason, there will be an extremely fierce rivalry between insurance companies. As a result, the focus of many insurance companies is on lowering rates and retaining a competitive edge. At the same time, insurance fraud constitutes a key proportion of insurance firms' costs because Insurance fraud decreases the profits of the insurance company and has a long-term effect on insurance firms' pricing strategies. Every year, several million dollars were lost due to insurance fraud. For instance, The Australian Insurance Fraud Bureau was in 2017 uncovered fraudulent claims totaling $280 million [3]. And in France, there are 44814 fraud claims found for a sum of EUR 214 million in 2013; in 2018, the amount increased to EUR 500 million [4]. According to figures from the United States Coalition Against Insurance Fraud, insurance fraud accounts for 17% of total compensation paid out by insurance firms, with an approximate value of 80 billion dollars a year [5] and in 2007, the Insurance Bureau of Canada (IBC) reports that car insurance

fraud exceeded 542 million Canadian dollars. In addition, insurance authorities in China announce that the share of insurance fraud is around 20 percent of overall insurance company payments, with an average of RMB 35 billion a year in 2011. lastly, the estimated cost of insurance fraud in developing countries is $600 million a year. Insurance fraud, therefore, is a global issue and has detrimental effects on the state and community.

As a result, how to accurately define risk factors and reducing damages incurred by fraud claims is a critical issue that insurance firms must resolve immediately. And the expert experience is essential in deciding whether a claim is fraudulent or not [6], while the number of experts is negligible as compared to the increasing number of claims. Therefore, it is difficult to sufficiently extract, interpret and assess the details of cases by a comparatively limited number of experts. Furthermore, a lack of experience can contribute to decision bias. Even when dealing with the same situation, different experts' opinions can vary significantly due to their varying points of view. Some practitioners and academic researchers, on the other hand, have made great efforts to detect car insurance fraud using machine learning algorithms [7].and by Compared the performance of these algorithms to the performance of experts in identifying fraud. Machine learning algorithms' performance is relatively more efficient on fraud detection.

Many of the applications of machine learning on insurance claims were studied but concentrated on medical insurance and not short-term insurance such as auto insurance. [8][9]. According to the papers published, motor insurance needs further effort and analysis to combat fraud. Non-fraudulent and fraudulent cases have a lot in common. And also, what makes detecting fraud even more complicated is the fact that the lack of a clear and accurate rule to characterize fraud cases. Consequently, using machine learning techniques to build automatic detection models has become a requirement for effectively combating fraud. This paper provides a comparative study of 13 machine-learning algorithms to predict fraud in automobile insurance. And one of the big issues with machine learning techniques is that they are influenced by unequal binary class distribution in the data set. In other words, when the data is unbalanced, some machine learning techniques will simply disregard little class and assign most of the cases into the common class since this will produce high overall model accuracy. Still, the prediction models' efficiency for the small class will significantly decline. And to deal with this problem, we will apply resampling techniques. Thus, the aim of this study is to suggest a method for improving the performance of machine learning algorithms to handle the imbalanced data issue by using various resampling techniques such as SMOTE, Random under Sampler, Random Over Sampler, and compare between them.

2. PROBLEM STATEMENT

Every year, insurance fraud costs the insurance industry billions of dollars, and this harms insurance company earnings, growth and also harms the national economic growth [10]. And insurance firms are moving the entire loss of fraud risk to the consumer by raising premium rates [11]. This gives an essential idea of how important it is to deal with fraud.

3. RELATED WORK.

Machine learning methods have been commonly applied for detection fraud purposes since the advent of artificial intelligence theory, such as.

[12] created a new approach for improving the accuracy of fraud prediction. Ten machine learning algorithms for fraud prediction were tested for efficacy and verifiability. They used car insurance data claims. According to this study, Random Forest outperforms all other algorithms in terms of fraud prediction. And by using data mining techniques, [13] predicts fraudulent claims and estimates insurance premium amounts for a range of customers depending on their personal and financial data. This study aids in the claims analysis screening process, which saves time and resources. Since the dataset for insurance and premium analysis is not accessible easily, this study creates a synthetic dataset. And the synthetic dataset's nature and the number of it is attributes are depending on field and case studies related to car insurance fraud. Ans this synthetic Insurance dataset is used to develop classification models that aid in the detection of fraudulent claims. Moreover [14], Using a Genetic Algorithm-based Fuzzy C-Means clustering and various supervised classifier models, this paper proposes a novel hybrid method for detecting frauds in car insurance claims. A test sample is first extracted from the original insurance dataset, the remaining data will be the training set, and the under-sampler applied to the training set using the clustering approach. And the test instances are classified as legitimate, malicious, or

suspicious after being exposed to the clusters. Genuine reports and fraud charges are ruled out, while questionable (suspicious) cases are investigated further using four different models, including Group Method of Data Handling, SVM, MLP, and DT. Finally, when SVM is used as the classifier, the proposed model has the highest Sensitivity and Specificity values. And [15] With the unbalanced data distribution, this study proposes a groundbreaking insurance fraud detection process. The idea is to create insurance fraud detection classifiers using ML models such as DT, SVM, and ANN using data partitions derived from under-sampling the majority class and combining it with the minority class. Finally, since the DT classifier performs better than other ML models, the DT classifier was used to compare various partitioning-under-sampling methods. The outcomes also show that when using the under-sampling process, the ML models perform better than when using the original imbalanced data.

And the study of [16] proposes OCSVM-based under-sampling. And in order to demonstrate the effectiveness of the proposed methodology, they used a dataset of vehicle insurance fraud and a dataset of credit card consumer turnover from the literature. They used DT, SVM, LR, PNN, and GMDH for classification (Group Method of Data Handling). Finally, they recommend using DT over other classifiers because it generates "if-then" rules while retaining a high AUC. And also [17], By combining k Reverse Nearest Neighbor and One Class support vector machines, propose a novel hybrid approach for correcting data imbalance. They used a car Insurance Fraud detection dataset and also a consumer Credit Card Churn prediction dataset. And to show the viability of the proposed classifier, they used various models such as Probabilistic Neural Network, Group Method of Data Handling, MLP, DT, SVM, and LR models. And in this study, they preferred the DT classifier in the Insurance Fraud Detection dataset because it generates "if-then" rules. And [18] also present a comparative analysis of prediction the fraud. They contrasted the efficiency of decision trees, survival analysis, and artificial neural networks (ANNs) by using a real-life car insurance fraud dataset from the United States.

Also [19], This study proposes an insurance fraud detection approach based on a random rough subspace neural network ensemble. This approach starts with a rough set reduction to produce a set of reductions that can keep data information stable. Second, the reductions are chosen at random to create a subset of reductions. Thirdly, using the insurance data, each of the selected reductions is used to train a neural network classifier. Finally, ensemble strategies are used to combine the qualified neural network classifiers. In addition, a real car insurance case is used to test the effectiveness and efficiency of the proposed method. The results of this paper show that detecting fraudulent insurance claims using a random rough subspace dependent neural network ensemble approach can be faster and more accurate, making it a promising tool for detecting insurance fraud. Moreover, some detection models have been developed that combine intelligent techniques with a variety of traditional statistical approaches, such as Bayesian networks, to improve prediction accuracy. For detecting insurance fraud, by [20]. they have proposed a Bayesian learning neural network. The explanatory capabilities of neural network classifiers with automatic significance determination weight regularization are investigated in this paper, as well as the effects of using these networks to detect fraud in personal injury protection auto insurance claims. To determine which inputs are the most insightful to the trained neural network model, they used the automated relevance determination objective function scheme. An implementation of [21,22] evidence framework approach to Bayesian learning is proposed as a practical way of training such networks. The value of predictors calculated by common logistic regression and decision tree classifiers was compared to the results of the neural network.

Additionally,[23], In this empirical study using Kohonen's Self-Organizing Feature Map to classify automotive bodily injury statements based on the degree of fraud suspicion. by Using feed-forward neural networks and a backpropagation algorithm, the validity of the Feature Map approach is investigated. Comparative experiments show the potential efficacy of the proposed technique. They show that this method outperforms both an insurance adjuster's and an insurance investigator's fraud evaluation in terms of precision and reliability. And [24]. used a multi-layer perceptron model for classification purposes in the field of medical insurance fraud and achieved high detection performance. The aim of this paper is to present the results of a study in which an MLP neural network was used to classify the practice profiles of a community of general practitioners.

Table 1 offers a brief list of articles on using machine learning to detect insurance fraud.

Table 1. Review Of Articles On Using Machine Learning To Detect Insurance Fraud.

The study	ML models														Imbalance data problem			
	Decision tree	Neural Networks	Logistic Regression	Multi-Layer Perceptron	Support Vector Machine	Random Forest	AdaBoost	CART	J48	NAÏVE BAYIES	k-Nearest Neighbors	C5.0	Stochastic Gradient	XGBOOST	Random Over Sampler	Random under Sampler	SMOTE	HYBRID
[12]	√		√	√	√	√	√	√	√	√								
[13]						√			√	√								
[14]	√			√	√											√		
[15]	√	√			√											√		
[16]	√	√		√	√											√		
[17]	√	√	√	√	√											√		
[18]	√	√																
[19]		√																
[20]	√	√	√															
[23]		√																
[24]				√														
[25]					√	√												
Present study	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Although the real-life insurance data always be heavily unbalanced. The table above shows a lack of using the resampling methods to solve the unbalanced data problem on previous studies that using machine learning models to detect insurance fraud except for few studies that applied the random Under sampler as table 1 show. So in this study, we will try to handle the unbalanced data problem using four different resampling methods that have never been applied before in the insurance field and compare them to find the best resampling method.

4. THE CONTRIBUTION OF OUR WORK

Real-world datasets for insurance are typically unbalanced and largely made up of one class and a small amount of the other class, so the prediction of this imbalanced variable appears so poorly with classification models because machine learning models will simply disregard the marginal class and assign most of the cases into the majority class, The present study, therefore, seeks to suggest an approach that enhances the results of machine learning models and handle the imbalanced data in the classification of insurance premium defaulting prediction problems by using resampling methods such as SMOTE, Random Over Sampler and Random under Sampler and compared them. Moreover, there is not any study in the insurance field applied or compared these approaches with each other.

In short, because of The significance of the data imbalance problem and the lack of implementation of resampling techniques as a way to deal with it in the insurance industry. The aim of this study is to investigate the effect of the imbalanced data issue on the output of ML algorithms, as well as how to overcome this issue using four different resampling methods, and we compare these different resampling methods by using different ML models to fill the gap in the previous studies.

As compared to similar studies, the following are the novel developments and essential processes of this study:

- Comparing and implementing four different resampling approaches.
- applied 13 machine learning classifiers to compare the efficiency of resampling methods used in this study.
- Illustrating the effects of resampling techniques on the efficiency of machine learning models.
- The use of four different resampling methods, such as over-sampling, under-sampling, combining between the over and under, and SMOTE approaches, and compare them for selecting the best makes this study unique.

5. PROPOSED MODEL

The proposed model's main phases are data collection, pre-processing, handle the unbalanced datasets, implementing classification models, and assessing the performance. Each stage of the model proposed is essential and has a beneficial impact on its efficiency. Figure 1 shows the proposed model of the detection of insurance fraud in this study.

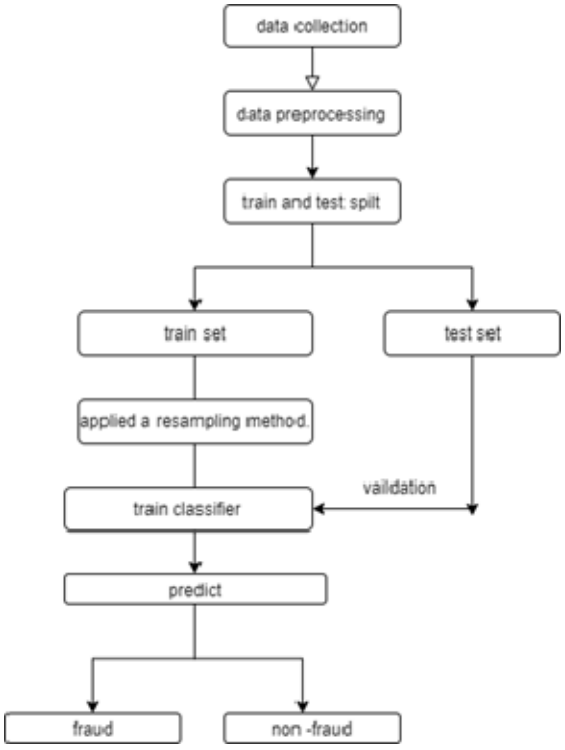

Figure 1: Proposed Model Of Fraud Insurance Detection.

5.1 Data Collection.

The data used in this study is real-life data obtained from an Egyptian car insurance firm, and fraud claims are verified by the competent department of the insurance company. We end up with 37082 claims in the dataset, each of which is a car insurance claim. In total, there are 2595 fraud claims and34487 non-fraud claims, suggesting that the data is unbalanced. And as we mention, ***t***he performance of classification algorithms is greatly affected by imbalanced data. So we use various resampling techniques employed to address the issue of data imbalance. SMOTE, Random Over Sampler, Random Under Sampler, and SMOTE-ENN are resampling approaches we use ***t***o build a balanced dataset. Lastly, ***e***ach claim comprises 22 features and one fraudulent label showing whether or not the claim is fraudulent. Table 2 provides a summary of the data.

Table 2. Attributes of the Data

No.	Description
1	Represent the time in months that the insured individual was a client.
2	Representing the age of the insured ranges between 19 and 64 years.
3	Represent The insured gender.
4	Represent the insured education level.
5	Represent the insured's work
6	Represent the amount that the insured can bear in the event of an accident before the insurance provider covers any expenses.
7	Represent the fee charged to the insurance provider for a one-year insurance policy.
8	Represent The maximum amount the insurance company can pay in the event of an accident.
9	Represent the incident type, including four categories, Multi- Collision, Single Collision, and Theft.
10	Represent the collision type, including four categories: Not Applicable, Front, Rear, and Side.
11	Represent the incident severity, including four categories: Major, Minor, Total Loss, and Trivial Damage.
12	Represent the authorities contacted, including four categories: Ambulance, Fire, None, Police and Other.

13	Represent number of vehicles involved in the accident.
14	Represent if there property damage, include three categories: Not Applicable, Yes and No.
15	Represent the number of persons injured in the accident if there is.
16	Represent the number of witnesses if there is.
17	Represent if there is a police report available, include Three categories: Not Applicable, Yes and No.
18	Representing the total claims amount.
19	Representing the total injury claim amount.
20	Representing the total property claim amount.
21	Representing the total vehicle claim amount.
22	Representing the brand of the car insured.
23	Fraudulent or not.

5.2 Data Pre-Processing.

Data preprocessing is one of the most critical steps in machine learning. This phase translates the raw data into a format that the machine learning models can understand. Datasets may include multiple errors; thus, this phase move will remove the errors, making datasets easier to manage [26]; for instance, numeric or binary values are allocated to categorical variables. For example, instead of male or female as the gender of the insured, the "Male" component would be (1), and "female" would be (0). After this point, we can apply this knowledge to all ML models used in this analysis. Fortunately, as a data preprocessing step, handling the missing data is not required since there is no missing data in the datasets used in this analysis.

5.3 Imbalanced Data.

The imbalanced data issue exists in many datasets; as a result, classifiers models are be biased against the minority class and are unable to predict it accurately [27]. In contrast, most machine learning models perform better when applied with balanced datasets. Analysis of the databases introduced shows that they are extremely imbalanced, and the two forms of insurance fraud are not balanced, with 2595 fraudulent claims and 34487 non-fraudulent claims. As a consequence, the imbalanced data issue must be addressed.

Several techniques have been developed to solve the problem of unbalanced data. The most successful approach for handling unbalanced data is the use of SMOTE as well as a sampling-based approach, either Random Over Sampler [28], Random Under Sampler [29].

We will use the ROSE package for over-sampling for the minority class and also for the majority least sampled class in our dataset, combining the over and under methods as a hybrid method. Also, we will use DMwR package to implement SMOTE as a reconfiguration method.

5.3.1 OVER-SAMPLING TECHNIQUE

This technique increases the weight of the minority class. It's important to note that the technique of over-sampling is typically used more than other methods.

- Random Over Sampler

Random Over-Sampling is a technique based on bootstrap that supports the binary classification task in the presence of unbalanced classes by generating synthetic examples from a conditional density estimation of the two classes [30]. It handles both continuous and categorical data. By repeating the original samples, as a result of this process, the dataset grows in size. The argument is that no new samples are generated by a random over-sampler, and the variety of samples remains constant [31].

- SMOTE

SMOTE is an effective method for re-balancing training data, which has been shown to be successful in solving the problems of unbalanced datasets, and it improved the output of the models in many recent studies such as in [[32],[33],[34],[35]]. SMOTE is similar to random oversampling. However, it does not regenerate the same instance. It creates a new instance by appropriately combining existing instances, thus making it possible to avoid the disadvantage of overfitting to a certain degree. Moreover, SMOTE is an oversampling technique that produces new minority samples by combining two minorities and one of their K nearest neighbors [36]. This approach is a statistical technique for creating new instances to increase the number of minority samples in a dataset. This algorithm takes characteristic features for the target class and its closest

neighbors, then produces new samples by combining the characteristics of a specific case with those of its neighbors. Often, new cases are not duplicates of minority samples that already exist.

5.3.2 Random Under Sampler.

Under-sampling is one of the simplest techniques to dealing with the issue of unbalanced data. It balances the majority and minority classes. The process of under-sampling includes arbitrarily deleting examples from the majority class in the training dataset, referred to as random under-sampling [37].

6. MACHINE LEARNING CLASSIFIERS.

Different machine learning classifiers are carried out in this paper, including Artificial Neural Network (ANN) [38], Multi-Layer Perceptron(MLP) [[39],[40]], Random Forest (RF) [41], K-nearest-neighbor (KNN) [[42],[43]], XG-boost (XG) [[44],[45]], AdaBoost [[46],[47],[48]], Support Vector Machine (SVM) [[49],[50]], Decision Tree [[51],[52]], and Naïve Bayes (NB)[53], Stochastic gradient boosting(SGB)[54], Logistic Regression (LR)[[55],[56][59]]. Table 3 lists all of the machine learning algorithms used in this paper, along with their parameter settings. And to make the comparisons as fair as possible, we tune all machine learning models to achieve their best performance. And the parameters of all machine learning used in this study are selected by 10-fold cross-validation. And by using these hyper-parameters, the best results are obtained for each machine learning model.

Table 3. Machine Learning models with settings for their parameters.

METHODS	parameters	METHODS	parameters
ANN	Hidden layer Neuron Count =20,50,100.200 ,500. Activation Function=RrLU, softmax. Optimizer=adam. Learning rate=0.001. Max Epochs=1000.	XG-boost	Eta=0.4. max_depth=2. gamma = 0. colsample_bytree=0.8. min_child_weight = 1. subsample =1. nrounds=50.
K-NN	K=9	CART	Cp=0.03932584
RF	Mtry= 93	AdaBoost	nIter=150 method= Adaboost.M1,
LR	no tuning parameters.	SVM	C=1
MLP	Activation Function=sine Scaling Function=Tanh Learning rate=0.1 Momentum=0	C50	Model=rules. winnow =FALSE. trials=20
NB	Laplace=0 Adjust=1 Usekernel=TRUE.	SGB	n.trees=150, interaction.depth =2, shrinkage=0.1 n.minobsinnode = 10 .
J48	C=0.010 M=1		

7. MODEL VALIDATION.

This paper employs a common cross-validation technique known as 10-fold cross-validation. And The dataset is divided into two sections, the first of which is referred to as the training data and the second as the test data. The training data accounts for around 80% of the overall data used, with the remainder being test data. With the training data, these models are trained and tested with the test data. The resampling approach should only be applied to the training data. As a consequence, only the training set is subjected to all of the resampling approaches, while the test classes must still be unbalanced.

8. EVALUATION METHODS.

Evaluation methods are essential in comparing and selecting the best model Because they are assessing the efficiency of classifiers [57]. Accuracy alone cannot always be reliable for a classification problem, as it can provide bias for a majority class, especially in the case of imbalanced data [58],[59]. And since the majority of policyholders do not commit fraud, car insurance claims are an excellent example of unbalanced data. Therefore, there will be a prejudice against a fraud class if accuracy is only used. So various measurement methods are used, including accuracy, sensitivity, precision, and F1-score, as well as the region under the curve (AUC). If results need to balance sensitivity and Specificity, AUC may be a better metric to use, particularly when there is an unbalanced class distribution.

$$Accuracy=(TP+TN)/(TP+FP+TN+FN). \quad (1)$$

$$Sensitivity=TP/(TP+FN). \quad (2)$$

$$Specificity=TN/(FP+TN). \quad (3)$$

$$Precision =TP/(TP+FP). \quad (4)$$

$$F\text{-}measure= (2 * Precision * Recall) / (Precision + Recall). \quad (5)$$

Where TP is the number of true positives, the number of false positives is FP, the number of true negatives is TN, and the number of false negatives is FN. Accuracy measures the proportion of predictions that are accurate, and a higher Accuracy value means a better overall performance of the forecast. Sensitivity relates to the ability to detect fraud claims correctly. Specificity refers to the ability to detect legal claims correctly. The significance of the predicted positives is referred to as precision. And The harmonic average of precision and sensitivity is the F1 score. AUC is the global classifier performance metric [60]. It is used to compare the overall performance of the model.

Since the class imbalance may result in high accuracy and high specificity rating but very low sensitivity. And since there is a weakness in accuracy, precision, and recall, because we cannot Reliance on them in the case of unbalanced data, unlike the AUC value; which a popularly used ranking evaluation technique, otherwise known as the receiver operating characteristic (ROC) or the global classifier performance metric, since all different classification schemes are measured to compare overall performance [60]. If the test set were to change its distribution of positive and negative instances, the previous metrics might not perform as well as when they were previously tested. However, the ROC curve is insensitive to the change in the proportion of positive and negative instances and class distribution [61]. So in this study, the AUC is weighted to emphasize overall model performance.

9. RESULTS&DISCUSSION.

This paper aims to use different resampling techniques to address the imbalanced data problem, and we applied 13 machine learning models. And to show the difference between the ability of machine learning models to predict the insurance fraud before and after handling the unbalanced data problem, we compared all applied models on the unbalanced data and also on the balanced data created by resampling techniques. And to evaluate the performance of the machine learning algorithms on fraud discrimination, we randomly split the data and use 80% as training data and the rest as testing data. We train the machine learning methods on the training data and predict whether the cases in the test data are fraud or not using the trained model. We report the performance measures of models on the testing data using six evaluation methods: accuracy, sensitivity, specificity, AUC, precision, and F1-score. And to make the comparisons as fair as possible, we tune all machine learning models to achieve their best performance. As we show in table 4 and the parameters are selected by 10-fold cross-validation. For this study, R x64 4.0.2 is used for implementing the models and techniques.

The results of the various classifiers on the imbalanced datasets and balanced datasets using four resampling techniques are shown in Table 4.

Table 4. Performance Of The Classifier.

MODEL	Unbalanced data					
	Accuracy	Sensitivity /recall	Specificity	AUC	Precision	F1
LR	0.7992	0.7536	0.8167	0.785	0.8963	0.8187
RF	**0.8394**	**0.7681**	**0.8667**	**0.817**	**0.9069**	**0.8317**
ANN	0.8032	0.6232	0.8722	0.748	0.8579	0.7219
MLP	0.6779	0.1731	0.8462	0.51	0.7543	0.2815
SVM	0.7992	0.5797	0.8833	0.732	0.8457	0.6878
NB	0.755	0.7681	0.75	0.759	0.894	0.8262
J84	0.8112	0.4783	0.9389	0.709	0.8244	0.6053
C50	0.8514	0.6232	0.9389	0.781	0.8667	0.7250
XG-boost	0.8032	0.6087	0.8778	0.743	0.8541	0.7108
KNN	0.6787	0.0724	0.91111	0.492	0.7193	0.1315
SGB	0.7871	0.3913	0.9389	0.665	0.80095	0.5257
CART	0.8353	0.6667	0.9	0.783	0.8757	0.7570
AdaBoost	0.8153	0.5942	0.9	0.747	0.8526	0.7003
	Random Over Sampler					
LR	0.7952	0.8841	0.7611	0.823	0.9448	0.9134
RF	0.8353	0.7681	0.8611	0.815	0.9064	0.8315
ANN	0.8313	0.7246	0.8722	0.798	0.892	0.7996
MLP	0.7928	0.7596	0.8055	0.782	0.8967	0.8224
SVM	**0.8635**	**0.8986**	**0.85**	**0.874**	**0.9561**	**0.9264**
NB	0.6305	0.8406	0.55	0.695	0.9	0.8692
J84	0.8072	0.7101	0.8444	0.777	0.8837	0.7874
C50	0.8394	0.7391	0.8778	0.808	0.8977	0.8107
XG-boost	0.8153	0.6087	0.8944	0.752	0.8564	0.7116
KNN	0.5462	0.4493	0.5833	0.516	0.7343	0.5574
SGB	0.8514	0.8406	0.8556	0.848	0.9333	0.8845
CART	0.8594	0.8261	0.8722	0.849	0.9289	0.8744
AdaBoost	0.8394	0.8261	0.8444	0.835	0.9268	0.8735
	Random under Sampler					
LR	0.7068	0.8841	0.6389	0.761	0.9349	0.9087
RF	0.8233	0.942	0.7778	0.86	0.9722	0.9568
ANN	0.8112	0.8696	0.7889	0.829	0.9404	0.9036
MLP	0.7566	0.8659	0.7148	0.79	0.9281	0.8959
SVM	0.8273	0.8551	0.8167	0.836	0.9363	0.8938
NB	0.5823	0.8986	0.4611	0.68	0.9222	0.9102
J84	0.7871	0.9275	0.7333	0.83	0.9635	0.9451
C5.0	0.8394	0.8696	0.8278	0.849	0.943	0.9048
XG-boost	0.747	0.7826	0.7333	0.758	0.8979	0.8362
KNN	0.49	0.6812	0.4167	0.549	0.7732	0.7242
SGB	0.8032	0.913	0.7611	0.837	0.958	0.9349

CART	0.8273	0.8551	0.8167	0.836	0.9363	0.8938
AdaBoost	0.8353	0.913	0.8056	0.859	0.9603	0.9360
	SMOTE					
LR	0.7189	0.7826	0.6944	0.7826	0.739	0.7601
RF	0.8233	0.942	0.7778	0.942	0.86	0.8991
ANN	0.8353	0.8986	0.8111	0.8986	0.855	0.8762
MLP	0.766	0.826	0.743	0.826	0.784	0.8044
SVM	0.7671	0.8551	0.7333	0.8551	0.794	0.8234
NB	0.6145	0.8406	0.5278	0.8406	0.684	0.7542
J84	0.751	0.9275	0.6833	0.9275	0.805	0.8619
C50	0.8474	0.942	0.8111	0.942	0.877	0.9083
XG-boost	0.8193	0.8116	0.8222	0.8116	0.817	0.8142
KNN	0.5743	0.4058	0.6389	0.4058	0.522	0.4566
SGB	0.8032	0.7391	0.8278	0.7391	0.783	0.7604
CART	0.8313	0.913	0.8	0.913	0.857	0.8841
AdaBoost	0.8072	0.8551	0.7889	0.8551	0.822	0.8382
	HYBRID					
LR	0.7349	0.8261	0.7	0.763	0.913	0.8673
RF	0.8554	0.8696	0.85	0.86	0.9444	0.9054
ANN	0.8434	0.7971	0.8611	0.829	0.91716	0.8529
MLP	0.7	0.7182	0.56122	0.703	0.7587	0.7378
SVM	0.8635	0.8986	0.85	0.874	0.9563	0.9265
NB	0.6145	0.8696	0.5167	0.693	0.91176	0.8901
J84	0.7831	0.6812	0.8222	0.752	0.8706	0.7643
C50	0.8514	0.9275	0.8222	0.875	0.9673	0.9469
XG-boost	0.8153	0.7971	0.8222	0.81	0.9136	0.8513
KNN	0.5141	0.4928	0.5222	0.507	0.7287	0.5879
SGB	**0.8715**	**0.913**	**0.8556**	**0.884**	**0.9625**	**0.9370**
CART	0.7871	0.7246	0.8111	0.768	0.8848	0.7967
AdaBoost	0.8273	0.8261	0.8278	0.827	0.9255	0.8729

Table 4 shows the Accuracy of each machine learning technique on unbalanced data as well as balanced datasets generated by four different resampling models. And we should note that only if the data is balanced will Accuracy be a valuable metric, while When a collection of samples is unbalanced, the Accuracy would be meaningless since the model forecasts the majority class's value for most outcomes. Furthermore, from table 4, we can see that the accuracy results using the different balanced dataset are not substantially enhanced, which is understandable given that most models predict with poorer Accuracy on the balanced data since they consider all classes at the same time. And after resampling techniques are used to address the issue of imbalanced data, we can trust the Accuracy. In general, Accuracy is one of the most common assessment methods to calculate the efficiency of a classifier. While it is simple to understand, but it overlooks various important factors that must be considered in evaluating a classifier's output. And the Stochastic Gradient Boosting classifier achieved 87.15 percent accuracy by using the hybrid method technique, which is the highest of all other classifiers, while by using Random under Sampler as a resampling technique, the

lowest accuracy outcome goes to the K-nearest-neighbor with 49%.

Sensitivity refers to the ability to detect fraud correctly. We can note that the Sensitivity for all models with the unbalanced data is lowest than the Sensitivity for balanced data created by different resampling methods; this refers to the effectiveness of using the resampling methods for handle the unbalanced data problem in the insurance industry. The highest Sensitivity in the dataset belongs to RF and C50 classifiers with 94.20% using the Random under Sampler and SMOTE respectively, and the lowest one goes to K-nearest-neighbor with 7.246% using the unbalance data among all other classifiers.

Specificity relates to the ability to detect non-fraud claims correctly. We can note that the Specificity for all models with the unbalanced data is higher than the Specificity for balanced data created by different resampling methods, this because our data largely made up of non-fraud classes and a small amount of fraud. So machine learning models will simply disregard the minority class and assign most of the cases to the majority class. And the highest Specificity goes to C50 and Stochastic Gradient Boosting classifiers with 93.89% using the unbalanced data, while the lowest one goes to the K-NN model with 41.67% using Random under Sampler among all other classifiers.

Precision refers to the relevance of the predicted positives, and it is the portion of the relevant outcomes. We note that the Precision values are improved after using the resampling methods. And The maximum precision in the dataset goes to the RF model by using Random under Sampler and also with SMOTE with 97.33%, and the lowest one goes to K-nearest-neighbor with 71.93% using the unbalance data.

The F1-measure is the harmonic average of sensitivity and precision, and it contains valuable information about the performance of classifiers in each class. The F1-score is also more useful when used in conjunction with the Sensitivity and Precision measurements. Because it measures the gap between the Sensitivity and Precision. Table 4 shows that When using imbalanced data, the classifiers do not achieve a good outcome with the F1-measure and thus do not perform well with all classes. This is a crucial issue that needs to be tackled as part of the overall data imbalance issue. After applying various resampling techniques and addressed the imbalanced data issue; The outcomes show that models do not overlook any classes. And this is one of the most important motivations to use resampling approaches. The highest F1-measure goes to RF classifier with 95.68% using the under-sampler method, and the lowest one goes to K-nearest-neighbor with 13.15% using the unbalance data.

AUC is the global classifier performance metric. It is used to compare the overall performance of the model. We note that the values of AUC for machine learning models except the naive Bayes model improved after using different resampling methods.

The highest AUC goes to Stochastic Gradient Boosting classifier with88.4% using the hybrid method, while the lowest goes to the K-NN model using the unbalanced data.

In general, As previously stated, this paper focuses on a dataset in a classification problem, with the goal of determining the impact of imbalanced data and deciding the best resampling technique and also the best model to predict insurance fraud. The results show that the models do not produce exact outcomes when dealing with imbalanced datasets and that most models are unable to predict the fraud classes. As a result, resolving the issue of unbalanced data is critical. In addition, the results show that after solving the imbalanced data problem, the performance of the majority of the models improves, and no classes are overlooked. And one of the most compelling reasons to use resampling methods to create balanced data is that it aids in the making of informed decisions. The findings confirm that there is no one resampling method that overall outperforms. For example, the best model after using the Random Over Sampler method is SVM, the best model after using the Random Under Sampler method is C5.0, the best model after using SMOTE method is C5.0, and the best model after using hybrid method is the Stochastic Gradient Boosting. Furthermore, among all the other

models, the Stochastic Gradient Boosting classifier obtained the best result when using the hybrid resampling technique, with high accuracy of 87.15%, which achieves the best output among all ML models. It also has the highest AUC scores with 88.4% that achieves a strong balance between sensitivity and specificity, in the sense that this model has the smallest gap between the sensitivity and the specificity as the important performance measure.

In the following table, we will compare our study with state-of-the-art works applied to detect insurance fraud.

Table 5. Comparison Of New Approach Performance Against State-Of-The-Art Works Which Applied Resampling Methods To Detect Insurance Fraud.

Articles	Accuracy	Sensitivity	Specificity	AUC	Precision	F1
[14]	87.02	83.21	88.45	-	-	-
[15]	-	95.8	-	73.0	69.4	-
[16]	58.92	95.52	56.58	-	-	-
[17]	60.40	91.89	58.39	-	-	-
Present study	**87.15**	**91.30**	**85.56**	**88.4**	**96.25**	**0.9370**

The aim of present the above table is to compare the previous studies that applied the resampling technique with our study. We found that all studies applied only the under-sampler method. But in our study, we applied four different resampling method.

After applying the under-resampling method in the previous studies shown in table 5, these studies achieve high sensitivity, which means they achieve high accuracy in predicting fraud. But at the same time, they achieved very low Specificity, which means they lose the ability to correctly predict non-fraud claims. On the other hand, our approach achieves high overall accuracy; also, it achieves a strong balance between sensitivity and Specificity with most models. Finally, our best model is achieved by the hybrid method, and it outperforms the previous studies. Because this model has the smallest gap between the sensitivity and the Specificity, with high accuracy for predict fraud and non-fraud at the same time as an important performance measure.

10. CONCLUSION.

Insurance Data mining is an analytic power tool that can uncover substantial and practical information about the insurance industry; however, it could face some difficulties., such as predicting fraud using imbalanced insurance data.

This paper aims to demonstrate the impact of the imbalanced data issue and identify the best resampling technique among the various techniques for dealing with this issue, including Random Over Sampler, Random Under Sampler, and SMOTE as individual resampling techniques and also a hybrid resampling technique. And several classifiers are used to evaluate the different resampling methods. The results show that classifiers cannot make appropriate predictions by using imbalanced data. On the other hand, when we applied machine learning models on the different balanced data created by different resampling methods, we can notice that many classifiers' results have improved, and all classes can be predicted, indicating that the classifiers' performance is satisfactory. And also, the results show that classifiers work differently on the different balanced data, so it difficult to choose the best resampling method. And The findings confirm that there is no one resampling method that overall outperforms. For example, the best model after using the Random Over Sampler method is SVM, the best model after using the Random Under Sampler method is C5.0, the best model after using SMOTE method is C5.0, and the best model after using the hybrid method is the Stochastic Gradient Boosting. Furthermore, when using the hybrid resampling technique, the Stochastic Gradient Boosting model outperformed all other classifiers.

11. FUTURE WORK.

Future work may be done in the next directions: Using hybrid classifiers to improve comparison and performance. Furthermore, feature selection approaches may be used to enhance model

results and gain a deeper understanding of the important features. It will also be worthwhile to conduct this research for another insurance branch, whether to predict claim occurrences or to predict fraud because these kinds of data always are very heavily unbalanced.

REFRENCEs:

[1] Yusof, N. H. M., & Abd Razak, A. Z. A. (2018). Customer Intention to Commit Motor Insurance Fraud: A Literature Review. *International Business Education Journal*, *11*(1), 40-48.

[2] Yusuf, T. O., & Babalola, A. R. (2009). Control of insurance fraud in Nigeria: an exploratory study (case study). Journal of Financial Crime.

[3] Insurance Fraud Bureau of Australia Homepage, https://ifba.org.au, last accessed 15/02/2021.

[4] French Agency for the Fight against Insurance Fraud – ALFA. https://www.alfa.asso.fr/, last accessed 15/02/2021.

[5] S. Jordon, "Insurance fraud: 'its all over the place' and you should care about it officials say", Omaha World-Herald, 2016.

[6] Knapp, C. A., & Knapp, M. C. (2001). The effects of experience and explicit fraud risk assessment in detecting fraud with analytical procedures. Accounting, Organizations and Society, 26(1), 25-37.

[7] Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance, 69(3), 373-421.

[8] Shi, Y., Sun, C., Li, Q., Cui, L., Yu, H., & Miao, C. (2016, March). A fraud resilient medical insurance claim system. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

[9] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 1-10.

[10] Tseng, L. M., Kang, Y. M., & Chung, C. E. (2014). Understanding the roles of loss-premium comparisons and insurance coverage in customer acceptance of insurance claim frauds. Journal of Financial Crime.

[11] Mohamed, M. (2013). *Countering Fraud in the Insurance Industry: A Case Study of Malaysia* (Doctoral dissertation, University of Portsmouth).

[12].Itri, Bouzgarne, Youssfi Mohamed, Qbadou Mohammed, and Bouattane Omar. Year. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. Paper presented at the 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS).

[13].Kowshalya, G, and M Nandhini. Year. Predicting fraudulent claims in automobile insurance. Paper presented at the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).

[14]. Subudhi, Sharmila, and Suvasini Panigrahi. 2017. Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. Journal of King Saud University-Computer and Information Sciences.

[15]. Hassan, Amira Kamil Ibrahim, and Ajith Abraham. 2016. Modeling insurance fraud detection using imbalanced data classification. In Advances in nature and biologically inspired computing. Springer, pp. 117-27.

[16].Sundarkumar, G Ganesh, Vadlamani Ravi, and V Siddeshwar. Year. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. Paper presented at the 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[17].Sundarkumar, G Ganesh, and Vadlamani Ravi. 2015. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence 37: 368-77.

[18].Gepp, A., Wilson, J.H., Kumar, K., & Bhattacharya, S. (2012). A Comparative Analysis of Decision Trees Vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. Journal of data science, 10, 537-561.

[19].Xu, Wei, Shengnan Wang, Dailing Zhang, and Bo Yang. Year. Random rough subspace based neural network ensemble for insurance fraud detection. Paper presented at the 2011 Fourth International Joint

Conference on Computational Sciences and Optimization.
[20].Viaene, Stijn, Guido Dedene, and Richard A Derrig. 2005. Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications 29: 653-66.
[21].MacKay, David JC. 1992 a. The evidence framework applied to classification networks. Neural computation 4: 720-36.
[22].MacKay, David JC. 1992 b. A practical Bayesian framework for backpropagation networks. Neural computation 4: 448-72.
[23].Brockett, Patrick L, Xiaohua Xia, and Richard A Derrig. 1998. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance: 245-74.
[24].He, Hongxing, Jincheng Wang, Warwick Graco, and Simon Hawkins. 1997. Application of neural networks to detection of medical fraud. Expert systems with applications 13: 329-36.
[25]. Nian, Ke, Haofan Zhang, Aditya Tayal, Thomas Coleman, and Yuying Li. 2016. A hybrid under-sampling approach for mining unbalanced datasets:spectral ranking for anomaly. *The Journal of Finance and Data Science* 2: 58-75.
[26].Kotsiantis, Sotiris B, Dimitris Kanellopoulos, and Panagiotis E Pintelas. 2006. Data preprocessing for supervised leaning. International Journal of Computer Science 1: 111-17.
[27].Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. 2006. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30: 25-36.
[28].Yap, Bee Wah, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. Year. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. Paper presented at the Proceedings of the first international conference on advanced data and information engineering (DaEng-2013).
[29].Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. 2008. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39: 539-50.
[30].Menardi, Giovanna, and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery 28: 92-122.
[31].He, Haibo, and Edwardo A Garcia. 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21: 1263-84.
[32].Fahrudin, Tora, Joko Lianto Buliali, and Chastine Fatichah. 2019. Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set. Int J Innov Comput Inf Control 15: 423-44.
[33].Ghorbani, Ramin, and Rouzbeh Ghousi. 2020. Comparing different resampling methods in predicting Students' performance using machine learning techniques. IEEE Access 8: 67899-911.
[34].Scrutinio, Domenico, Carlo Ricciardi, Leandro Donisi, Ernesto Losavio, Petronilla Battista, Pietro Guida, Mario Cesarelli, Gaetano Pagano, and Giovanni D'Addio. 2020. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Scientific reports 10: 1-10.
[35].Hussain, Lal, Kashif Javed Lone, Imtiaz Ahmed Awan, Adeel Ahmed Abbasi, and Jawad-ur-Rehman Pirzada. 2020. Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. Waves in Random and Complex Media: 1-24.
[36].Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. Year. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.
[37].Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. 2014. ROSE: A Package for Binary Imbalanced Learning. R journal 6.
38. Jain, Anil K, Jianchang Mao, and K Moidin Mohiuddin. 1996. Artificial neural networks: A tutorial. Computer 29: 31-44.
[39].de Oliveira, Fagner A, Cristiane N Nobre, and Luis E Zárate. 2013. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4,

Petrobras, Brazil. Expert Systems with Applications 40: 7596-606.
[40].Heaton, Jeff. 2008. Introduction to neural networks with Java. Heaton Research, Inc.
[41].Zhang, Guoqiang, B Eddy Patuwo, and Michael Y Hu. 1998. Forecasting with artificial neural networks:: The state of the art. International journal of forecasting 14: 35-62.
[42].Cunningham, Padraig, and Sarah Jane Delany. 2020. k-Nearest Neighbour Classifiers. arXiv preprint arXiv:2004.04523.
[43].Wang, Jigang, Predrag Neskovic, and Leon N Cooper. 2007. Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recognition Letters 28: 207-13.
[44].Zięba, Maciej, Sebastian K Tomczak, and Jakub M Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications 58: 93-101.
[45].Freund, Yoav, Robert Schapire, and Naoki Abe. 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14: 1612.
[46].Xiao, Ling, Yunxuan Dong, and Yao Dong. 2018. An improved combination approach based on Adaboost algorithm for wind speed time series forecasting. Energy Conversion and Management 160: 273-88.
[47].Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals of statistics 28: 337-407.
[48].Duffy, Nigel, and David Helmbold. 2002. Boosting methods for regression. Machine Learning 47: 153-200.
[49].Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. 2003. "A practical guide to support vector classification." Taipei.
[50].Suykens, Johan AK, and Joos Vandewalle. 1999. Least squares support vector machine classifiers. Neural processing letters 9: 293-300.
[51].Safavian, S Rasoul, and David Landgrebe. 1991. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 21: 660-74.
[52].Du, Wenliang, and Zhijun Zhan. 2002. Building decision tree classifier on private data.
[53].Rish, Irina. Year. An empirical study of the naive Bayes classifier. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.
[54].Friedman, Jerome H. 2002. Stochastic gradient boosting. Computational statistics & data analysis 38: 367-78.
[55].Cox, David R. 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) 20: 215-32.
[56].Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. Applied logistic regression. Vol. 398, John Wiley & Sons.
[57].Hossin, Mohammad, and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5: 1.
[58].Ganganwar, Vaishali. 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering 2: 42-47.
[59].Hanafy, Mohamed, and Ruixing Ming. 2021. Machine learning approaches for auto insurance big data. Risks 9: 42.
[60].Wu, Shaomin, and Peter Flach. Year. A scored AUC metric for classifier evaluation and selection. Paper presented at the Second Workshop on ROC Analysis in ML, Bonn, Germany.
[61].Bradley, Andrew P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30: 1145-59.