iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44962

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Prediction of Fraudulent or Genuine Transactions on Credit Card Fraud Detection Dataset Using Machine Learning Techniques

Sunday Oluwafemi Akinwamide[1], Fele Taiwo[2], Faluyi Bamidele Ibitayo[3]
[1, 2, 3]*Department of Computer Science, The Federal Polytechnic, Ado Ekiti, Nigeria*

Abstract: ***Credit card fraud is one of the important financial frauds that has caused a huge amount of financial losses more than the past. Several protection mechanisms such as Fraud Prevention Systems (FPSs) are used to combat the current credit card frauds, but these methods are not efficient enough to reduce the impact of these frauds. It is therefore important that credit card companies are able to recognize and prevent fraudulent credit card transactions so that customers are not charged for items that they did not purchase. In this research, we present our approach to predict legitimate or fraudulent transactions on Credit Card Fraud Detection Dataset from Kaggle.***

This work is based on analysis at two levels using performance evaluation metrics such as Precision, Recall, F1_Score and ROC_AUC. In the first stage of the research, the original imbalanced and skewed dataset was used to train, predict and evaluate the six supervised machine learning classifiers considered in this research including: Extreme Gradient Boosting (XGBoost), Random Forest (RF), K-Nearest Neighbour (KNN), Decision Tree (DT), Logistic Regression (LR), and Naïve Bayes (NB) while the same set of classifiers were also trained, predicted and evaluated with the same dataset but now resampled using SMOTE-Tomek, i.e., a combination of both under-sampling and over-sampling technique to eliminate the imbalanced nature of the dataset during the second stage. The results of the two stages are compared to select the best overall performance. However, the result of the second stage of the experiment where models are trained, tested and evaluated with resampled dataset gave the overall best results where XGBoost, RF and DT have 100% in Precision, Recall, F1_Score and ROC_AUC respectively. While comparing the overall results of our research with all the papers reviewed in Section 2 of this work, it is worth noting that our research achieved the best performance so far where 100% were recorded from three different classifiers in all the four metrics used to evaluate our work.

Keywords: Credit Card Fraud; Resampling Techniques; SMOTE-Tomek; Machine Learning; XGBoost; Random Forest; Decision Tree, Class Imbalance

I. INTRODUCTION

A credit card is a simple yet no-ordinary card that allows the owner to make purchases without bringing out any amount of cash. Instead, by using a credit card, the owner borrows funds from the issuing company, which is often a bank, to make purchases whether online or onsite [1].

Credit cards are becoming more and more indispensable in our everyday lives, being used for online shopping or at a supermarket. However, identity card data breach and fraudulent credit card transactions are also increasing with the expansion of credit card use.

Credit card fraud is the intentional procurement of goods, services, or monies with a stolen, lost, cancelled, or counterfeit credit card. Credit card fraud is one of the most frequently committed types of white collar crime [2]. It is perpetrated by individuals as well as crime rings. Credit card fraud is an ever-growing financial crime.

The five key types of credit card fraud, according to the Australian Payments Network [3], are: Card-not-present (CNP) fraud, Counterfeit and skimming fraud, Lost and stolen card fraud, Card-never arrived-fraud and False application fraud.

Data from Australian Payments Clearing Association (APCA) to June 2016 shows that 0.0279% ($530million) of all credit card and cheque transactions were fraudulent [4]. Over the last ten years, the amount of fraud has increased significantly due to a rise in online transactions (where the physical card is not present), and sophistication of the technology used by the criminals.

One American over ten has been a victim of credit card fraud (median amount of $399), according to the Statistic Brain Research Institute [5].

According to the latest European Central Bank (ECB) report [6], the total level of card fraud losses amounted to €1.8 billion in 2018 in the Single European Payment Area (SEPA). There are many issues and difficulties when it comes to detecting credit card fraud. This type of fraud detection relies heavily on studying data and much of this data is unavailable from banks and financial institutions due to its sensitive and personal nature. Imbalanced Data i.e., most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones. Also, due to high volume of transactions every day, the analysis poses significant challenges in terms of information technology and for researchers analysing the data. As fraud detection techniques develop and become more sophisticated, so are fraudsters also changing their methods over time to achieve their goals [7].

Credit card fraud detection can be carried out in two major ways, namely: Machine Learning (ML) credit card fraud detection and conventional fraud detection. ML credit card fraud detection is characterized by detecting fraud automatically, real-time streaming, requiring little time for verification methods and identifying hidden correlations in data; while the conventional fraud detection involves manually making a decision on determining schemes, taking enormous amount of time, requiring multiple verification methods which is inconvenient for the users and can only detect obvious fraudulent activities [8].

Due to the imbalanced distribution of the classes in the dataset, the proposed approach highlights the imbalance class issue by using SMOTE-Tomek, i.e., a hybrid or combination of both under-sampling and over-sampling technique to normalize the dataset after choosing the best machine learning algorithms. This research proposed an advanced approach in terms of choosing the best machine learning classifiers in collaboration with the best resampling technique. This work is based on analysis at two levels using performance evaluation metrics. In the first stage of the research, the original imbalanced and skewed dataset was used to train, predict and evaluate the six classifiers considered in this research while the same set of classifiers were also trained, predicted and evaluated with the same dataset but now resampled to eliminate the imbalanced nature of the dataset in the second stage. The results of the two operations were compared to select the best overall performance.

The remaining parts of this paper are structured as follows: Related work were review in section two. In section three, materials and methodology were carried out. Sections four and five contain classification models and evaluation criteria respectively. Experimental setup and results discussion took place in section six and finally, conclusions and future work competed the work in section seven.

II. RELATED WORK

Much research has already been done using various artificial intelligence, data mining and machine learning techniques to predict credit card fraud detection. In this section, we presented various studies between the existing research papers related to the credit card fraud detection.

Development of effective approach using machine learning to prevent fraudulent credit card transactions in credit card fraud detection dataset of European cardholders was carried out by [9]. This paper studied a total of 66 machine learning models based on two stages of evaluation. In the first stage, nine machine learning algorithms were tested to detect fraudulent transactions. In the second stage, the best three algorithms were nominated to be used again with 19 resampling techniques used with each one of the best three algorithms. Imbalanced issue in the dataset was address in both stages using stratified K-fold cross-validation technique. Performance evaluation of the models were addressed through AUC, Accuracy, Recall, Precision and F1 Score. All K-Nearest Neighbors (AllKNN) under-sampling technique along with CatBoost (AllKNN-CatBoost) is considered to be the best proposed model having AUC value of 97.94%, a Recall value of 95.91%, and an F1-Score value of 87.40%. The research is computationally intensive as it took nearly a whole month to obtain outputs and results.

The work of [10] proposed an effective credit card fraud detection mechanism including a feedback system, dependent on machine learning methodology. Seven supervised machine learning classifiers were used to train a slightly skewed credit card fraud dataset. The efficiency of the classifiers was measured based on: precision, recall, F1-score, accuracy, and FPR percentage. Random Forest has the best performance accuracy of 95.00%, and precision of 95.99%. Because of the imbalance nature of the dataset, the performances recorded, most especially the accuracy might not be reliable.

Building of fraud detection systems using machine learning, deep learning, and data mining techniques to detect whether transactions are fraudulent or genuine based on IEEE-CIS Fraud Detection dataset was proposed by [11]. Random under sampling and SMOTE techniques were performed on the dataset to get it normalized. Immediately after that, two deep learning models, namely: Bidirectional Long short-term memory (BiLSTM) with max pooling layer and Bidirectional Gated Recurrent Units (BiGRU) with max pooling layer and six machine learning classifiers which are: Naïve base, Voting, Ada boosting, Random Forest, Decision Tree, and Logistic Regression were applied on the dataset. The metrics used to evaluate the performances of the classifiers include: AUC, precision, recall and f1 score. The results from machine learning classifiers show that the best AUC was 81% by hard

voting with under sampling and over sampling techniques while the result obtained by concatenating the two deep learning models achieved 91.37% of AUC. The performance of concatenating the two deep learning models was too encouraging because common machine learning models like decision tree and random forest can even perform better than that in some situations.

Proposing the design and development of fraud detection system that detects credit card frauds using supervised machine learning models was carried out by [12]. Here, different machine learning algorithms were implemented on an imbalanced dataset such as logistic regression, naïve bayes, random forest with ensemble classifiers using boosting technique. The dataset was divided into train – test in ratio 70:30. The model performances were evaluated on the basis of quantitative measurements such as accuracy, precision, recall, f1 score, support, confusion matrix. By comparing all the three methods, it was observed that random forest classifier with boosting technique performed better than the logistic regression and naïve bayes methods. Because of the imbalanced nature of the dataset used in the experiment, the results of the models might not be the true representation of the performances of the classifiers.

The work of [13] built a model which predict fraud and non-fraud transactions with respect to time and the amount involved in the transaction using machine learning algorithms and neural networks. Credit card fraud detection dataset was used to train – test logistic regression (LR), naïve bayes (NB), decision tree (DT) and artificial neural network (ANN) models. The models were evaluated based on accuracy, precision and recall. The result of the experiment showed that ANN has the best performance of 98.69% accuracy, 98.41% precision and 98.98% recall. The result of the experiment might not truly represent the performance of the models because the dataset was unbalanced and skewed.

Determining the potential fraudsters by referring to their previous mistakes and details using Machine Learning and Neural Networks was proposed by [14]. Machine Learning algorithms such as Multinomial Naive Bayes, Random Forest Regression, Logistic Regression, Support Vector Machine and a basic Neural Network were used with the credit card fraud detection dataset. The models were evaluated based on confusion matrix and classification reports using the sklearn library. From the results obtained, it was concluded that K-Nearest Neighbours outperformed the rest of the classifiers. It was also noted that the performance of neural network could have been improved using some other techniques as well. The result of the experiment might not truly represent the performance of the models because the dataset was not resampled.

The research of [15] developed credit card fraud detection system using machine learning models to classify both fraudulent and normal transactions in credit card fraud dataset. The algorithms used are random forest and the Adaboost. The two algorithms were compared to choose the algorithm that better detect credit card fraudulent transactions. The results of the two algorithms are based on accuracy, precision, recall, and F1-score.

The random forest algorithm has the highest precision, recall, and F1-score of 95%, 77% and 85% respectively. The result and outcome of this research can be misleading because the dataset was not balanced.

Article [16] applied artificial Intelligence and machine learning techniques to detect fraud transactions in credit card using credit card fraud dataset. Various machine learning techniques such as Artificial Neural Network (ANN), Decision Trees, Support Vector Machine (SVM), Logistic Regression and Random Forest were used to detect fraudulent transactions. The performance measurement of the techniques was measured using accuracy, precision and false alarm rate. The experiment showed that Radom Forest achieved an accuracy of 99.21%, Decision Tree 98.47%, Logistic Regression 95.55%, SVM 95.16% and ANN 99.92%. The imbalanced nature of the dataset was not taken care of and the results obtained from this experiment might be mis-leading.

Inspecting execution of, Support Vector Machine, Naive Bayes, Logistic Regression and K-Nearest Neighbour on exceptionally distorted data on credit card fraud was proposed by [17]. Different phases were engaged in the experiment. The learning phase is where the classifier's system was created and supplied with the extracted information. The execution of these techniques was assessed based on accuracy, sensitivity, precision, specificity. The outcomes show an accuracy for logistic regression, Naive Bayes, k-nearest neighbour and Support vector machine classifiers having 99.07%, 95.98%, 96.91%, and 97.53% respectively. Because of the skewed nature of this dataset, the result and outcome of this research can be incorrect and misleading.

The research of [18] analyzed various machine learning algorithms, such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB) and Multilayer Perceptron (MLP) in order to determine which algorithm is most suitable for credit card fraud detection. SMOTE technique was used for resampling the dataset. Furthermore, feature selection was performed on the dataset. The algorithms used in the experiment were Logistic Regression, Random Forest, Naive Bayes and Multilayer Perceptron. The performances of the classifiers were determined using: accuracy, recall and precision where RF has the best performance of 99.99% accuracy, 81.63% recall and 96.38% precision respectively. The result of this research would have been more generalized if under-sampling techniques were also considered.

Proposing the design and development of a novel fraud detection method for Streaming Transaction Data, with an objective to analyse the past transaction details of the customers and extract the behavioural patterns was performed by [19]. Cardholders were clustered into different groups based on their transaction amount. Sliding window strategy were used to aggregate the transaction made by the cardholders from different groups so that the behavioural pattern of the groups can be extracted respectively. Later different classifiers are trained over the groups separately. After pre-processing, different classifiers on each group was trained using the cardholder's behavioural patterns in that group and extract fraud features. SMOTE operation was performed on the dataset to get it balanced and finally, the classifier that is used for training the group is applied to each cardholder in that group. The classifier with highest rating score is considered as cardholder's recent behavioural pattern. Accuracy, precision and Matthews Correlation Coefficient (MCC) were used to evaluate the performances of the classifiers. The classifier with highest rating score is considered as cardholder's recent behavioural pattern. It was finally observed that Logistic regression, decision tree and random forest are the algorithms that gave better results. The limitation observed in this work is that over-sampling the dataset does not provide any good results.

The work of [20] detected fraudulent transactions while minimizing the incorrect fraud classifications using supervised machine learning techniques. Analysis and pre-processing of credit card transaction dataset as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm was carried out. The model performances were evaluated on the basis of quantitative measurements such as accuracy, precision, recall and f1 score. While the algorithm reached over 99.6% accuracy, its precision remained at 28% when a tenth of the dataset is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rose to 33%. The result of this experiment where two classifiers were employed i.e., Local Outlier Factor and Isolation Forest Algorithm together with imbalanced nature of the dataset might not be the true reflection of the performance of the models.

In the work of [21] proposed implementing supervised machine learning algorithms for the classification of a credit card transaction as either fraudulent or non-fraudulent. To mitigate the imbalanced class size of the dataset, SMOTE sampling technique was applied to the dataset. Thereafter, the performance of three machine learning algorithms: Random Forest, Support Vector Machine and Logistic Regression were compared based on static and incremental learning to detect fraud on real-life data containing credit card transaction. The performance of the techniques is evaluated based on area under the ROC curve and Average Precision (AP). LR has the best result, followed by RF while SVM has the poorest performance. The only two performance metrics i.e., area under the ROC curve and average precision (AP) used to evaluate the classifiers in the experiment might be too small for the result to be generalized.

Investigating the performance of naïve bayes, k-nearest neighbour and logistic regression on highly skewed credit card fraud data was done by [22]. A hybrid technique of under-sampling and over-sampling was carried out on the skewed data while the three supervised machine learning classifiers mentioned above were applied on the original skewed, under-sampled and over-sampled dataset. The performance of the techniques was evaluated based on accuracy, sensitivity, specificity, precision, Matthew's correlation coefficient and balanced classification rate. The performance accuracy of the research was 97.92%, 97.69% and 54.86% for naïve bayes, k-nearest neighbour and logistic regression respectively. The feature selection and reduction in conjunction with under-sampling techniques carried out on the dataset during pre-processing may lead to high loss of relevant information, thereby affecting the overall result of the research.

All the reviews presented here above specified the need for further research based on highlighted limitations among others which affets the performance and classification of the models. Hence, the need for this study arises.

III. MATERIALS AND METHODOLOGY

In this research, we made use of Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests.

A. About Dataset

The dataset contains transactions made by credit cards in September 2013 by European cardholders, downloaded from Kaggle [23]. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced with no missing value, the positive class (frauds) account for 0.172% of all transactions.

The dataset contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the original features and more background information about the data could not be provided.

Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.
This dataset is widely used by many researchers as reflected in the immediate previous section, i.e., Related Work; hence, this dataset is chosen to compare the evaluation metric values of our proposed model with the previous researches.

B. Dataset Pre-Processing

Having a close examination with the dataset, it was discovered that the dataset has no missing value(s), no categorical data and features V1, V2, … V28 are the result obtained from PCA transformation. Hence, the dataset has already undergone pre-processing and is already prepared and set to be used for the training, testing and evaluating the performances of the classifiers considered for this research.

IV. CLASSIFICATION MODELS

There are six machine learning algorithms that are being used in this research. They are XGBoost, RF, KNN, DT, LR, and NB. Each algorithm is explained as follows:

A. XGBoost

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models [24]. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

B. RF

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems [25].

C. KNN

KNN algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K-Nearest Neighbour) suggests, it considers K Nearest Neighbours (Data points) to predict the class or continuous value for the new datapoint. The algorithm’s learning is: 1.
Instance-based learning: Here we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data. 2. Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance. 3. Non-Parametric: In KNN, there is no predefined form of the mapping function [26].

D. DT

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree [27]. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

E. LR

LR is one of the traditional machine learning algorithms that is still used today due to its quick analysis and simple method of processing the features of a class. The ability of LR to associate various factors, especially with strong ones, and its ability to adjust to different factors depend on predictor variables and the outcome. LR uses values that are greater than 1 and less than 0 to treat the anomalies in the dataset, and it is not limited to classifying and predicting binominal outcomes, but also multinomial outcomes as well, and it uses the sigmoid function to estimate the values of parameters' coefficients [28]. When LR examines the values of the attributes during an ongoing transaction, it tells us whether the transaction should continue or not, as it is used for clustering [29].

F. NB

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [30]. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at equation 1 below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

where:

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

G. Resampling Technique

A widely adopted technique for dealing with highly imbalanced datasets is called resampling technique [31]. Taking a look at the dataset used in this work, the total number of valid cases is 284,315 and the total number of fraud cases is 492. This implies that the valid cases are 99.827% while the fraud cases are only 0.173% of the total number of cases. Obviously, the dataset is highly imbalanced. Therefore, performing resampling technique on the dataset is necessary, because the imbalance class will negatively affect the performance of the algorithms. There are 3 main categories of resampling techniques: under-sampling, oversampling, and the hybrid or combination of both under-sampling and oversampling.

1) *Under-sampling:* Under-sampling techniques usually provide a compact balanced training set, and one of the merits of this kind of technique is that it reduces the cost of the learning phase [32]. One of the disadvantages of under-sampling techniques is the removal of a large volume of the training set, especially when the majority class instances are tremendously huge, which can lead to the loss of important cases that would, in turn, lead to problems in classification and prediction Examples of under-sampling technique are: Random Under-sampling, Condensed Nearest Neighbour, Tomek, One-Sided Selection, Edited Nearest Neighbours, All K-Nearest Neighbours, etc.
2) *Over-sampling:* Over-sampling methods aim to preserve the majority class instances and replicate the minority class instances in order to solve the problem of imbalanced training set. The challenge with this kind of technique is that it may lead to poor performance of the model in some cases because it may be hard to generate the minority data in the training set [33]. Examples of over-sampling technique are: Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic, etc.
3) *Hybrid of Combination of Both Under-sampling and Over-sampling:* This combination aims to use both under-sampling and over-sampling techniques at the same time. By merging these techniques, the imbalance class issue is addressed differently. Examples of Hybrid or Combination of Both Under-sampling and Oversampling are: SMOTE with Edited Nearest Neighbor (SMOTE-ENN), SMOTE-Tomek Links method (SMOTE-Tomek), etc. Here in this research, we decided to adopt hybrid or combination of both under-sampling and over-sampling technique. The specific example of hybrid or combination of both under-sampling and oversampling technique used here is SMOTE-Tomek.

This technique is briefly explained as follows:

a) *SMOTE-Tomek:* SMOTE-Tomek is a hybrid or combination of the two most powerful algorithms used for over-sampling and under-sampling imbalanced datasets, namely: SMOTE and Tomek Links. Introduced first by [34], this method combines the SMOTE ability to generate synthetic data for minority class and Tomek Links ability to remove the data that are identified as Tomek links from the majority class as mentioned in sections 4.7.1 and 4.7.2 respectively thereby making samples of data from the majority class that is closest with the minority class data.

V. EVALUATION CRITERIA

When there is a case of class imbalance just as we are having it in the dataset we used in this research, i.e., Credit Card Fraud Detection dataset, accuracy can become an unreliable metric for measuring the performance of our models. On this note, all the six models considered in this work is evaluated based on: Precision, Recall, F1-Score and ROC_AUC_Score. All the evaluation metrics used in this proposed approach depend on a confusion matrix in one way or another [35].

A. Confusion Matrix

A confusion matrix is a table that is used to describe the performance of a classification model, or a classifier, on a set of observations for which the true values are known (supervised) [36]. Each row of the matrix represents the instances in the actual class while each column represents the instances in the predicted class (or vice versa). For example, here is a dummy confusion matrix for a binary classification problem predicting yes or no (1 or 0) from a classifier:

Table 1: Confusion matrix

		Predicted	
		NO (0)	YES (1)
Actual	NO (0)	TN	FP
	YES (1)	FN	TP

From Table 1, the following terms could be defined:

True Positives (TP): These are cases that was predicted Yes (1) and were actually labelled Yes (1).

True Negatives (TN): These cases were predicted No (0), and they were actually labelled No (0).

False Positives (FP): These cases were predicted Yes (1), but they were actually labelled as No (0). This is known as Type I error.

False Negatives (FN): These cases were predicted No (0), but they were actually labelled Yes (1). This is known as Type II error.

B. Precision

Precision is the ratio of true positives and total positives predicted. It is calculated as:

$$P = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

C. Recall (Sensitivity)

This is the proportion of actual positive cases which are correctly identified and it is calculated thus:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

D. F1-Score

The F1-Score is the harmonic mean of the precision and recall, where an F1-Score reaches its best value at 1 (perfect precision and recall) and worst at 0. F1-Score is calculated as:

$$F1 = 2 \text{ x } \frac{Precision \text{ X } Recall}{Precision + Recall} \quad (4)$$

E. ROC_AUC_Score

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve [37]. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

VI. EXPERIMENTAL SETUP AND RESULTS DISCUSSION

Tables 2 – 5 together with Figures 1 and 2 show the results of this proposed research work. This work is based on analysis at two levels using performance evaluation metrics. In the first stage of the research, the original imbalanced and skewed dataset was used to train, predict and evaluate the six classifiers, namely: XGBoost, RF, DT, LR, NB and KNN considered in this work. From Table 4, it is obvious that XGBoost has the best performance of 98.73% Precision, 79.56% Recall, 88.14% F1-Score and 89.80% ROC_AUC, followed by RF and DT while KNN has the highest Precision of 100% but with the poorest overall performance.

In the second phase of the research, the original imbalanced and skewed dataset was resampled using SMOTE-Tomek, i.e., a hybrid or combination of both under-sampling and over-sampling technique to normalize and make the dataset balanced. The six classifiers were also trained, tested and evaluated with the resampled dataset. According to Table 5, XGBoost, RF and DT attained the best and highest performance of 100% in Precision, Recall, F1-Score and ROC_AUC respectively while NB has the lowest overall performance. It was discovered that using resampling techniques such as SMOTE-Tomek makes the overall performance very reasonable as compared with when the resampling techniques were not used on the imbalanced dataset. This is represented in Tables 4 and 5. However, obtaining the highest score of 100% results in Precision, Recall, F1-Score and ROC_AUC most especially with XGBoost, RF and DT show a great improvement in credit card fraud detection without compromising on the detection of fraudulent cases as much as possible in credit card transactions. While comparing the overall results of our research with all the previous papers reviewed in Section 2 of this work, it is worth noting that our research achieved the best overall performance so far where the highest score of 100% were recorded from three of the classifiers in all the four metrics used to evaluate this work!

Table 2: Confusion Matrix of Original Imbalanced Dataset

	XGBoost	RF	DT	LR	NB	KNN
TN	56863	56862	56835	56829	56502	56864
TP	78	76	78	55	62	5
FN	20	22	20	43	36	93
FP	1	2	29	35	362	0

Table 3: Confusion Matrix of SMOTE-Tomek Resampled Dataset

	XGBoost	RF	DT	KNN	LR	NB
TN	56864	56864	56864	55075	55883	56448
TP	98	98	98	98	89	69
FN	0	0	0	0	9	29
FP	0	0	0	1789	981	416

Table 4: Results of Original Imbalanced Dataset

	Precision (%)	Recall (%)	F1-Score (%)	ROC_AUC (%)
XGBoost	98.73	79.59	88.14	89.80
RF	97.44	77.55	86.36	88.77
DT	72.90	79.56	76.10	89.77
LR	61.11	56.12	58.51	78.03
NB	14.62	63.27	23.75	81.31
KNN	100	5.10	9.71	52.55

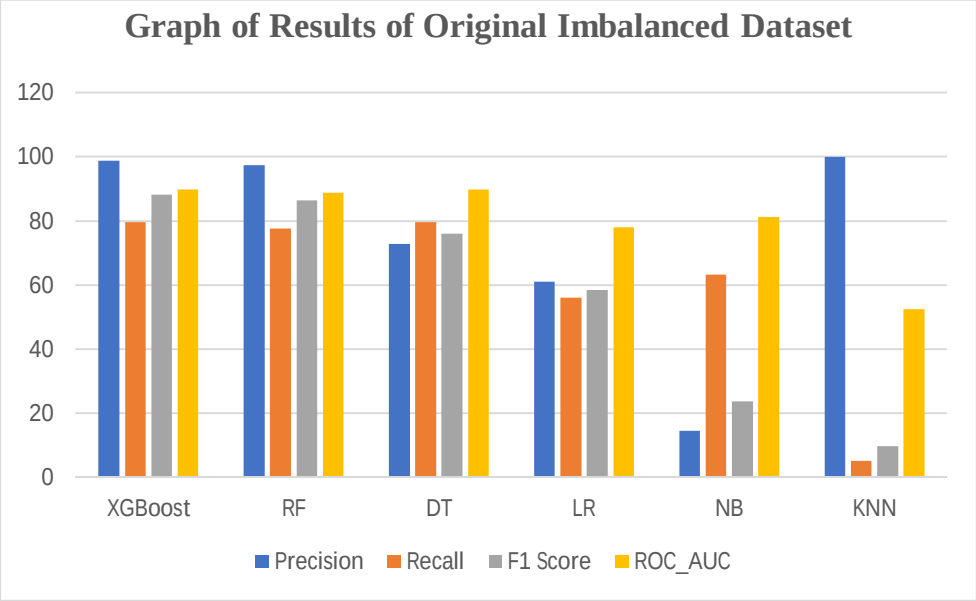


Figure 1: Graphical Representation of Results of Original Imbalanced Dataset

Table 5: Results of SMOTE-Tomek Resampled Dataset

	Precision (%)	Recall (%)	F1-Score (%)	ROC_AUC (%)
XGBoost	100	100	100	100
RF	100	100	100	100
DT	100	100	100	100
KNN	5.19	100	9.87	98.43
LR	8.32	90.82	15.24	94.55
NB	14.23	70.41	23.67	84.84

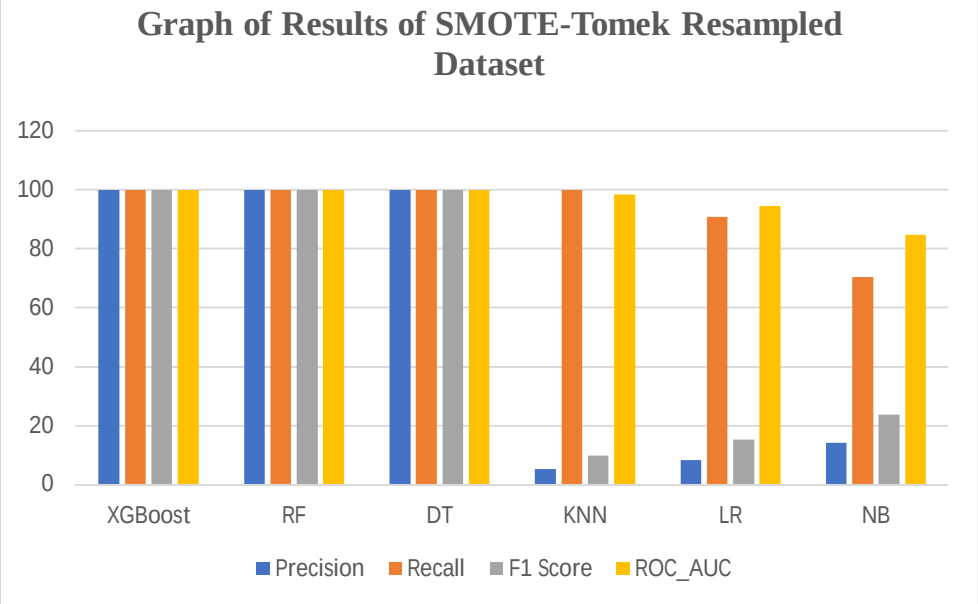


Figure 2: Graphical Representation of Results of SMOTE-Tomek Resampled Dataset

VII. CONCLUSIONS AND FUTURE WORK

With increased dependency on online, electronic and credit cards transactions, criminals and fraudsters are developing various ways to steal and defraud other people's money. However, a proactive measure must be considered by harnessing data mining, artificial intelligence and machine learning tools to immediately tackle this challenge, regardless of sophisticated operations of these fraudsters and criminals.

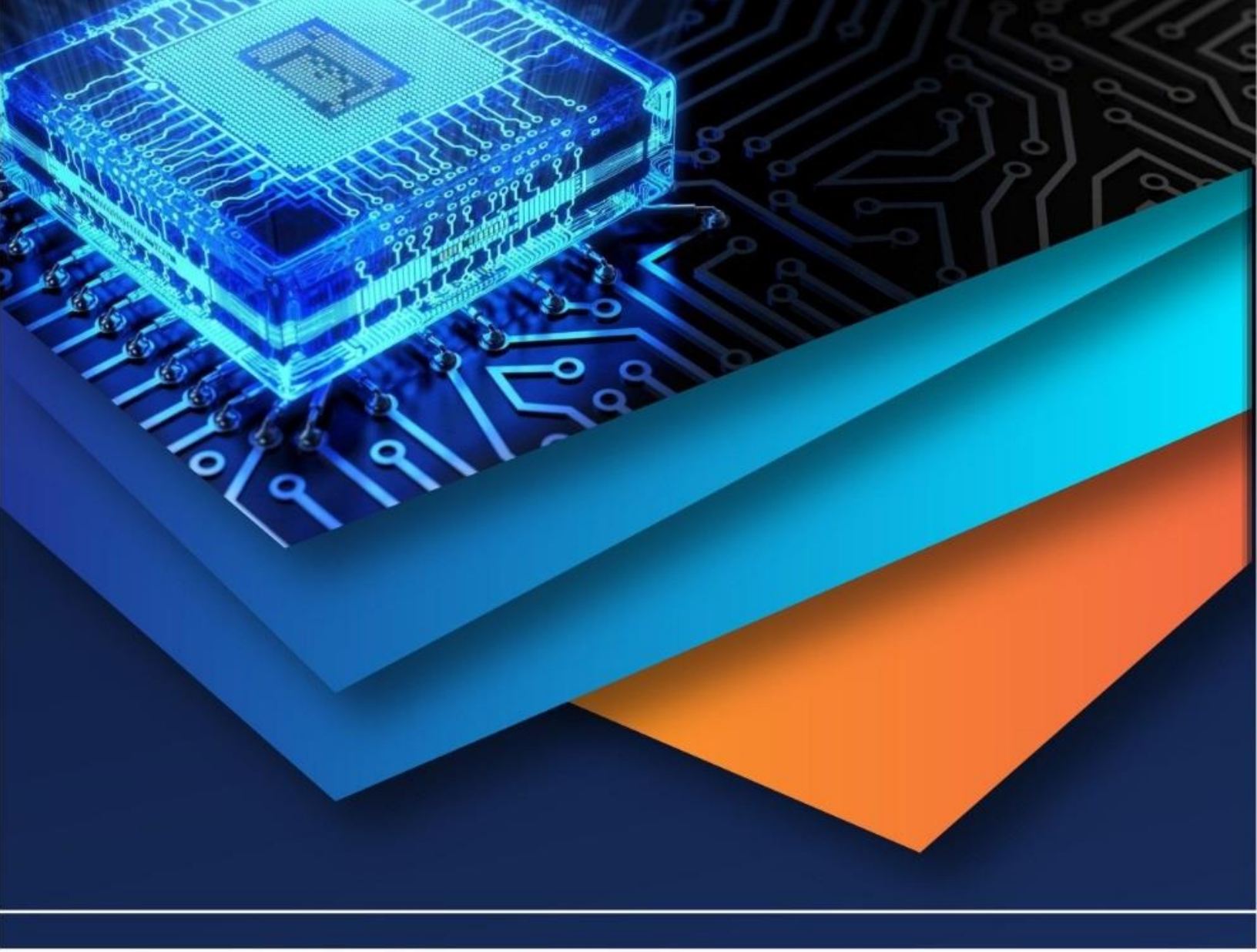
The proposed approach was divided into two phases. The results of the first stage where the original imbalanced and skewed dataset was used to train, predict and evaluate the six models namely: XGBoost, RF, DT, KNN, LR and NB were compared with the results of the second stage where the same dataset, after going through resampling process was used to train, predict and evaluate the same set of models. The results of the second stage where the models were trained, predicted and evaluated with resampled dataset outperformed the outcome of the first phase where XGBoost, RF and DT achieved the highest score of 100% performances in Precision, Recall, F1-Score and ROC_AUC as shown in Tables 3, 5 and Figure 2 respectively.

Future work may consider the inclusion of the results of under-sampling and over-sampling techniques separately with SMOTE-Tomek, i.e., a hybrid or combination of both under-sampling and over-sampling technique using the same dataset.

REFERENCES

[1] CFI (2020) Credit Card. A card that allows the owner to make cash-less purchases. https://corporatefinanceinstitute.com/resources/knowledge/other/credit-card/
[2] Impact Law (2022). Credit Card Fraud. https://www.impactlaw.com/criminal-law/white-collar/credit-card-fraud
[3] William Jolly (2019). Common credit card frauds and how to avoid them. https://www.savings.com.au/credit-cards/credit-card-fraud
[4] Avers Cloud Solution (2022). Credit Card Security - Merchant Responsibilities. https://avers.com.au/Bookkeeping/Blog/Credit-Card-Security---Merchant-Responsibilities/
[5] Statistic Brain Research Institute. Credit card fraud statistics. April 2018. [Online; Last consulted 30-March-2021]. https://www.statisticbrain.com/credit-card-fraud-statistics/
[6] European Central Bank. 6th report on card fraud. August 2020. (Online; Last consulted 09-October-2020). https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202008~521edb602b.en.html#toc2
[7] National Home Security Alliance. Credit Card Fraud Detection Techniques. https://staysafe.org/credit-card-fraud-detection-techniques/
[8] Roman Chuprina and Olena Kovalenko (2022). SPD-Group. Credit Card Fraud Detection: Top ML Solutions in 2021. https://spd.group/machine-learning/credit-card-fraud-detection/
[9] Noor Saleh Alfaiz and Suliman Mohamed Fati (2022). Enhanced Credit Card Fraud Detection Model Using Machine Learning. Electronics 2022, 11, 662. https://doi.org/10.3390/electronics11040662. https://www.mdpi.com/journal/electronics
[10] Naresh Kumar Trivedi, Sarita Simaiya, Umesh Kumar Lilhore and Sanjeev Kumar Sharma (2020). An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods. International Journal of Advanced Science and Technology Vol. 29, No. 5, (2020), pp. 3414 - 3424 3414 ISSN: 2005-4238
[11] Hassan Najadat, Ola Altiti, Ayah Abu Aqouleh, and Mutaz Younes (2020). Credit Card Fraud Detection Based on Machine and Deep Learning. 2020 11th International Conference on Information and Communication Systems (ICICS)
[12] Andhavarapu Bhanusri, K. Ratna Sree Valli, P. Jyothi, G. Varun Sai, R. Rohith Sai Subash (2020). Credit card fraud detection using Machine learning algorithms. Quest Journals. Journal of Research in Humanities and Social Science, Volume 8 Issue 2 (2020) pp. 04-11. ISSN (Online): 2321-9467. www.questjournals.org
[13] Varun Kumar K S, Vijaya Kumar V G, Vijay Shankar A and Pratibha K (2020). Credit Card Fraud Detection using Machine Learning Algorithms. International Journal of Engineering Research & Technology (IJERT) IJERTV9IS070649. Vol. 9 Issue 07, July-2020. ISSN: 2278-0181. http://www.ijert.org
[14] Mohammed Azhan and Shazli Meraj (2020). Credit Card Fraud Detection using Machine Learning and Deep Learning Techniques. Proceedings of the Third International Conference on Intelligent Sustainable Systems, [ICISS 2020] IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3
[15] Ruttala Sailusha, V. Gnaneswar, R. Ramesh and G. Ramakoteswara Rao (2020). Credit Card Fraud Detection Using Machine Learning. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020). IEEE Xplore Part Number: CFP20K74-ART; ISBN: 978-1-7281-4876-2
[16] Praveen Kumar Sadineni (2020). Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms. Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) IEEE Xplore Part Number: CFP20OSV-ART; ISBN: 978-1-7281-5464-0
[17] Olawale Adepoju, Julius Wosowei, Shiwani lawte and Hemaint Jaiman (2019). Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques. 2019 Global Conference for Advancement in Technology (GCAT) Bangalore, India. Oct 18-20, 2019
[18] Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla (2019). Credit Card Fraud Detection - Machine Learning methods. 18th International Symposium INFOTEH-JAHORINA, 20-22 March 2019
[19] Vaishnavi Nath Dornadula and Geetha S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. International Conference on Recent Trends in Advanced Computing 2019, ICRTAC 2019
[20] S P Maniraj, Aditya Saini, Swarna Deep Sarkar and Shadab Ahmed (2019). Credit Card Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research & Technology (IJERT), IJERTV8IS090031. Vol. 8 Issue 09, September-2019. ISSN: 2278-0181. http://www.ijert.org
[21] Maja Puh and Ljiljana Brkic (2019). Detecting Credit Card Fraud using Selected Machine Learning Algorithms. MIPRO, May 20 – 24, 2019, Opatija, Croatia
[22] John O. Awoyemi, Adebayo O. Adetunmbi and Samuel A. Oluwadare (2017). Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis. 978-1-5090-4642-3/17/$31.00 ©2017 IEEE

[23] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download
[24] Ramya Bhaskar Sundaram (2018). An End-to-End Guide to Understand the Math behind XGBoost https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/
[25] Niklas Donges (2021). Random Forest Algorithm: A Complete Guide. All you need to know about the random forest model in machine learning. https://builtin.com/data-science/random-forest-algorithm
[26] Sai Patwardhan (2021). Simple understanding and implementation of KNN algorithm! https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/
[27] Nagesh Singh Chauhan (2022). Decision Tree Algorithm, Explained. All you need to know about decision trees and how to build and optimize decision tree classifier. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
[28] Jain Y., Namrata T., Shripriya D. and Jain, S. "A comparative analysis of various credit card fraud detection techniques". Int. J. Recent Technol. Eng. 2019, 7, 402–403
[29] Sahin, Y. and Duman E. "Detecting credit card fraud by ANN and logistic regression". In Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15–18 June 2011; pp. 315–319
[30] Sunil Ray (2017). 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[31] Analytics Vidhya (2020). https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/
[32] Dal Pozzolo, A., Caelen, O., Bontempi, G. When is under-sampling effective in unbalanced classification tasks? In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Porto, Portugal, 7–11 September 2015; pp. 200–215
[33] Cieslak, D.A., Chawla, N.V. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Notre Dame, IN, USA, 15–19 December 2008; pp. 143–152
[34] Batista, G. E. A. P. A., Bazzan, A. L. C., and Monard, M. A. (2003). Balancing Training Data for Automated Annotation of Keywords: Case Study. Proceedings of the Second Brazilian Workshop on Bioinformatics, pp. 35–43
[35] Scikit-Learn Developers. 3.1. Cross-validation: Evaluating Estimator Performance. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 22 January 2022)
[36] Manish Pathak (2020). Quick Guide to Evaluation Metrics for Supervised and Unsupervised Machine Learning. https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/
[37] Aniruddha Bhandari (2020). AUC-ROC Curve in Machine Learning Clearly Explained. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY