International Journal of Advance and Innovative Research
Volume 12, Issue 2 (XVII): April - June 2025
ISSN 2394 - 7780

FRAUDULENT TRANSACTION IN THE FIELD OF FINANCE

[1]Neel Naik, [2]Fatima Shaikh and [3]Balkrishna Parab
[1]Student, Department of Big data Analytics
[2, 3]Assistant Professor, Department of Big data Analytics, Jai Hind College, Mumbai 400020, India

ABSTRACT
Financial scam is a crucial problem in the banking industry and the detection of fraudulent transactions is an important task for banks to protect their customers and maintain confidence in the financial system. This has led to an exponential rise in daily transactions. In this article, three to four methods are used to form normal transactions and fraudulent transactions, namely (Support Vector Machine), Logistic Regression, Decision Tree, Random. Forest, Naïve Bayes, ANN, Bagging, Boosting and K-Nearest Neighbour. The tools and technologies used are sklearn, imblearn and benchmark metrics to evaluate model performance. In addition, in this paper, we want to combine the model that gives the most accurate accuracy result, using the classification ratio to check the accuracy and recall value of the model and help detect the fraudulent transaction faster. This evaluation provides comprehensive guidance for choosing an optimal algorithm based on the type of fraud, and this article shows the result with an appropriate performance metric.

Keywords: *Financial fraud, ML technique*

I. INTRODUCTION
The level of fraud has significantly risen as there is development in the advanced technology and communication globally. Fraud can be detected in two main ways: prevention and detection. Prevention makes it completely impossible for any attack to be instituted by fraudsters since it is an encasing layer. Detection comes when the entire process of prevention has been tried without success. The primary goal of the system fraud detection is the identification of the fraud in early stages so that the required measures can be taken against it.

Machine learning is the solution of this generation that replaces these Techniques and approaches can be applied to extensive data sets, which is not feasible for humans. Machine learning techniques can be broadly classified into two categories: supervised learning and unsupervised learning. Fraud detection analysis can be performed using two different methods, and the choice of method can be determined on the database only. Supervised learning is contingent on categorization of anomalies beforehand. Over the past few years, several supervised algorithms have been applied to fraudulent credit card detection.[3]

One strategy that is conspicuous in the credit card fraud detection issue is data mining. Credit card fraud detection means the ability to distinguish between fraudulent and legitimate transactions into two categories of genuine and fraudulent.[1].Credit card fraud detection is also centered on the spending pattern of a card. Several ways have been used in credit card fraud detection, one of them being artificial neural networks. Imbalanced datasets are a usual problem in data mining and classification particularly when the datasets are un-proportioned. Current studies have shown significant interests in the class imbalance problem. [2]. This study focuses on improving data sampling techniques by combining oversampling techniques with random under sampling [2].

Therefore, this article studies several combinations of oversampling techniques (derived from the family of Synthetic Minority Oversampling Techniques (SMOTE)) with random subsampling techniques designed to address some of the related problems. The data in this article is taken from Kaggle which contains 1,296,675 rows and 23 columns and the data is unbalanced in 0.5822% of the transactions are fraudulent in the whole data set. The main contributions of this article are briefed as follows. To tackle the problem of fraud detection, several machine learning algorithms and Deep learning algorithm are used. The combination of models is done based on the highest accuracy score. From the results of the experiments, some conclusions were drawn that may be useful for future work.

II. REVIEW OF LITERATURE
Real-time credit card fraud detection using machine learning [3] The article explores a method to detect fraud transactions in real-time using machine learning models. This system categorizes fraud into four main types and addresses challenges such as the unbalanced distribution of data, which is common in fraud detection. Using machine learning models (SVM, Naive Bayes, logistic regression and K-Nearest Neighbour) and resampling techniques (such as SMOTE), the approach It significantly improves detection accuracy and response speed. This real-time capability allows financial institutions to act quickly on reported transactions, potentially

ISSN 2394 - 7780

reducing financial losses and improving fraud management. Credit Card Fraud Detection Using a Naive Bayesian Model and KNN Classifier [4]This article explores the application of machine learning techniques, specifically Naive Bayesian and K-Neighbour Neighbour (KNN) algorithms, to detect fraudulent transactions of credit cards. The authors, Kiran Sai et al. , highlights the challenges of fraud detection due to the large number of daily transactions and the unbalanced nature of the data set, where only a small fraction of transactions are fraudulent. The authors conclude that relying on a single fraud detection algorithm may not be effective, and suggest that combining multiple algorithms can leverage their respective strengths to improve accuracy. Overall, the article contributes to ongoing efforts to improve fraud detection methodologies in the financial sector. Integration of a machine learning-based fraud detection system based on a risk management framework [5]

This article explores the application of machine learning methods, particularly ensemble approaches such as random forests, to identify fraudulent activity in digital financial transactions. It demonstrates how traditional statistical techniques have given way to present machine learning models, emphasizing how well random forests tackle the difficulties presented by unbalanced datasets, which are prevalent in fraud detection situations. In conclusion, there is opportunity in the use of machine learning in financial risk management; nevertheless, to keep up with the constantly developing financial landscape and technological obstacles, continuous innovation and development are required. Preventing fraudulent banking transactions with a deep learning algorithm [6] The paper deals basically with the fraud of financial transaction and the demand for complex detection methods that find and stop such frauds. The authors indicate the possibility of automatically extracting attribute values from data on transactions with the use of deep learning algorithms, in the form of a multilayer perceptron or MLP. Deep learning algorithms have been emphasized as a potent instrument for handling big data sets and shifting to new fraud trends, overcoming the drawbacks of conventional rule-based methods.

The article discussed the basic building blocks of the bank fraud predictor, including risk assessment, data gathering, data analysis, monitoring, and continued improvement. Both the level and volume of the training data would affect the precision of the MLP model. Ensuring that there is a diverse dataset which contains a broad spectrum of fraudulent behavior is necessary for the model to be able to identify the characteristics and patterns that define fraudulent behavior. Identifying Credit Card Fraud Using Random Forest [7] Two kinds of different random forest models that can be used to detect credit card fraud are presented in this research. For the first model, based on random forest, the basic classifier is the direct implementation of decision trees. For the second model, it employs CART as its initial classifier since it uses CART random forest. Both models are trained on historical transaction data that includes both legitimate and fraudulent transactions, so to learn the characteristics of regular and anomalous transactions. The performance of two different kinds of random forest models was investigated in this research. Our experiment uses a genuine B2C dataset of credit card transactions. Random forest still has several issues, such as inaccurate data, even if it produces good results on little data. Detecting fraud with credit cards using a machine learning algorithm from an inconsistent data set [8]. The purpose of this article is to evaluate various classifiers by analyzing various machine learning approaches using various metrics. Instead of incorrectly labeling a legitimate transaction as fraudulent, this strategy seeks to enhance fraud detection. It also addresses the issue of imbalanced data. They therefore employed strategies like undersampling and oversampling to address this issue.

In addition, clustering techniques, such as the use of k-means clustering and genetic algorithms, can be effective in dealing with unbalanced datasets by generating new examples of minority classes. Credit Card Fraud Detection Using Machine Learning Techniques: A Study in Comparison [1] Three machine learning approaches—Naive Bayes, k-nearest-neighbors model, and logistic regression—are evaluated in this article for the purpose of detecting credit card fraud. Only 0.172% of the credit card transactions in the seriously skewed data set used by the researchers were fraudulent. To balance the dataset, they adopted a hybrid strategy that undersampled transactions which were valid and oversampled transactions which were fraudulent. The article highlights the difficulties in detecting credit card fraud, including the ever-changing character of fraudulent activity and The datasets' high degree of skewness shows how effectively hybrid sampling techniques work to enhance machine learning model performance. Detecting Fraudulent Financial Transactions Using Machine Learning [9]

In this research, the use of machine learning approaches to predict the legitimacy of financial transactions with accuracy and efficiency is studied. Among the machine learning algorithms that the researchers compared were an MLP Regressor, Random Forest Classifier, Complement NB, Gaussian NB, Bernoulli NB, LGBM Classifier, Ada Boost Classifier, K Neighbors Classifier, Logistic Regression, Bagging Classifier, Decision Tree Classifier, and Deep Learning. The data was collected from the Kaggle database, with 10 columns and 6,362,620 rows. The Random Forest Classifier was the best classifier in the unbalanced informational collection with 99.97%

ISSN 2394 - 7780

exactness, 99.96% F1scores, 99.97% remember, and 99.96% precision. With 99.96%, 99.97%, and 99.97% accuracy, the Bagging Classifier was the top classifier for the balanced dataset. 95%, 99.98% F1 efficiency, and 99.98% recall. Sampling and subsampling combined methods for imbalanced classification: A analyzing credit card fraud data using machine learning [2]

In order to detect credit card fraud, this article analyzes how well classification models perform when oversampling and undersampling strategies are used. Credit card fraud represents an escalating problem because fraudulent transactions are difficult to identify accurately because of data set imbalances. The authors tackle class imbalance by combining multiple oversampling techniques from the SMOTE family and random undersampling methods. Models were evaluated using random forest classifiers and performance metrics like precision, recall, and F1 score modified for unbalanced datasets. The findings demonstrate that approaches to increase average precision, recall, and F1 score are used in both oversampling and undersampling by about 0. 80%

III. EXPREMENTAL METHODOLOGY

A. Data Description

The dataset titled "Transaction Train Fraud" was obtained from Kaggle Discovery.With a total of 1,296,675 records and 23 features, the database is heavily weighted towards the positive class, with fraudulent transactions accounting for only 0.562% of total transactions. And the dataset is in CSV format, that is, in a format where data values are separated by commas.

Table.1: Data Description

Features	Description
trans_date_time	Time when the transaction takes place
cc_num	Credit Card Number
merchant	Name of merchant who sold the product
Category	What type of product is sold
Amt	Amount of the transaction in local currency.
first, last	First-name and last name of Buyer
gender	Gender of Buyer
City, street, zip, lat, long city_pop	Address
job	Jobe of buyer
dob	Date of Birth
Trans_num	Transaction number of payment happened
Merch_lat and long	Merchant locatin
Is _fraud	This is the transactions made by the fraudulent agents inside the simulation.

B. Data Preprocessing:

Finding the most relevant factors in a dataset through feature selection is an essential method that helps to decrease overfitting, increase accuracy, and shorten training times. [3] Feature extraction is done by checking the influence of the input variable on the output variable. For accurate results and to train a model with accurate data, this article transforms the input features using PCA. The principal components acquired from the PCA transformations are only the numerical values under the attributes [4], and the only feature that has not been transformed with the transformation of the principal component analysis is the "amt" attribute. "amt" contains data that represents nothing in addition to the amount of transactions and this function can also find its use for automatic learning of the cost-sensitive and of the instance. After PCA to transform the data to the same scale, this article also used Standard Scaler to scale the data to the normal temporal format. And Last but not least, the response variable "is_fraud" has the value "1" in the event of a fraudulent transaction, or a positive result, and "0" in the event of a genuine transaction.

ISSN 2394 - 7780

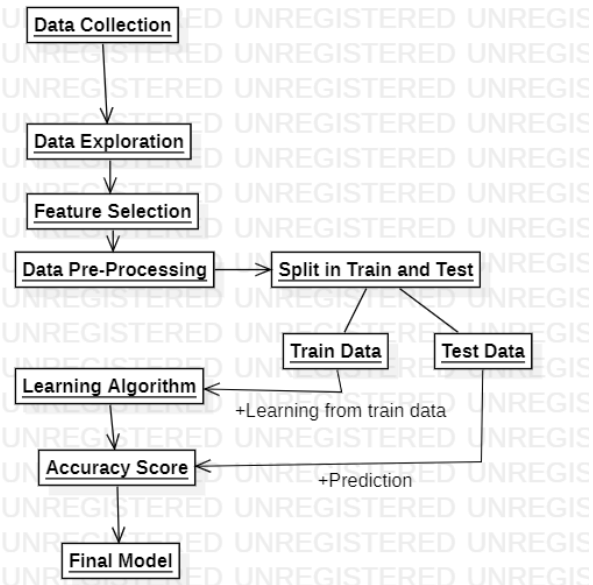

Fig.1 Data Flow

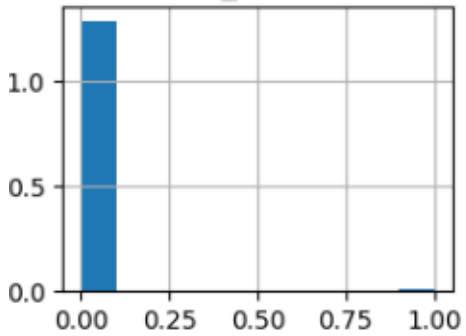

Fig.2 Imbalanced dataset

C. Sampling Method:

One method for changing the size of training groups is sampling. Oversampling alters the training samples by repeating the samples from the minority training set while the under sampling from a smaller majority training set. It is expected that both approaches will improve the situation by decreasing the degree of imbalance and class imbalance. In data mining, classification with unbalanced datasets has emerged as one of the most difficult issues. Three primary methods can be taken into consideration when sampling data: undersampling, oversampling, and a combination of the two. [2]

1. Over Sampling Technique:

In machine learning,over sampling is the process of replicating the records of lesser representation to alleviate the problem of uneven data sets, particularly where one side is substantially large as compared to the other side. This disparity may lead to models which are biased and accomplish nothing for the smaller class. In order to achieve equal distribution, more instances of the minority class need to be added.

2. Under Sampling Technique:

In machine learning, under sampling is a specific method meant to solve the problem of class imbalance within a dataset. It consists of decreasing the count of samples in the overwhelming majority class to have a more balanced data set. So, each model can concentrate on learning the features of the smaller class, which typically tends to be the focused class.

3. Combine Sampling Technique:

The mixer of oversampling and undersampling techniques is a powerful approach in the domain of machine learning, particularly in addressing class imbalance. This method attempts to reduce any overfitting issues while dealing with the class imbalance problem. In this work, SMOTE-Tomek Links was implemented where SMOTE generates artificial samples for the underrepresented class. Moreover, Tomek Links removes the nearest neighbors of the minority class samples from the majority class, thereby increasing the separation of the classes.

ISSN 2394 - 7780

D. Machine Learning Models:

1. Logistic Regression:
One technique for categorization tasks is logistic regression. It simulates the likelihood that a particular class or event will occur, depending on one or more independent variables. Logistic regression differs from linear regression, which forecasts a continuous numerical value. Logistic Regression predicts a categorical outcome, such as "yes" or "no", "spam" or "no spam", or "positive" or "negative". The sigmoid function, maps the linear combination of the independent variables to a probability between 0 and 1 [1].

Mathematically, the sigmoid function is defined as: sigmoid(z) = 1 / (1 + e^(-z))

2. Support Vector Machine:
Support Vector Machines (SVM) are a classification method that seeks to identify the ideal "hyperplane" or boundary to separate different groups. The SVM will then identify the line that best separates these two groups, choosing line that maximizes the difference or distance from the nearest host on each side. These nearest guests are known as *support vectors* and help determine the position of the boundary [9].

3. Naïve Bayes:
Based on the training data's probabilities and conditional probabilities of occurrence, the Naïve Bayes machine learning classifier attempts to predict a class known as the result class. This type of learning, which is also known as supervised learning, is incredibly efficient, quick, and practical. Conditional probabilities using Bayes' theorem constitute the initial stage of the Naïve Bayes classifier. The class is "C," and the known data sample is "x." [4]:

P (C / x) = P(x/C)/P(x)

4. Decision Tree:
A decision tree is a model that separates data into discrete to sort people into groups by particular features to provide judgments and predictions. The structure of the tree is like a flow chart where the final branches at the end hold. for the ultimate classification or suggestion, as well as each path represents one option based on a trait. A Decision Tree is an intuitive and efficient tool for classification and decision-making because of its systematic division, which enables it to make conclusions quickly by reducing the number of options.

5. Random forest:
Random Forest for classification and regression tasks is the ensemble learning method which create multiple decision tress and composite them to give predictions that is more accurate and reliable. Random Forest enhances, accuracy, handles missing data, Prevents Overfitting combining the knowledge of multiple trees. Because each tree Random Forest is particularly suitable because it makes decisions independently, The following process, if done iteratively and not sequentially, can help the developer(s) build a robust and generalizable model that is more stable with respect to variations in the data than a single decision tree would be [7].

6. ANN (Artificial Neural Network):
An artificial neural network (ANN), also known as a deep neural network, is a computer system that follows after the structure and functions of the human brain. It is mostly used for pattern recognition, classification, and prediction. An input layer, one or more hidden layers, and an output layer are the layers made up of interconnected nodes, or "neurons."Each neuron receives input, processes it by applying weights (which represent the importance of each input), then passes the result through an activation function in determines whether it should be "enabled" or enabled. This signal is then transmitted to the next layer of neurons in the network. Through a process called training, the ANN adjusts these weights by comparing its predictions with actual results, using methods such as back propagation to reduce errors. [6].

7. Bagging:
Bagging, also known as Bootstrap Aggregating, is an ensemble learning approach which utilizes several iterations of a model trained on various subsets of the data to boost the accuracy and stability of machine learning models. From the initial training dataset, several random samples, or *bootstraps*, initially develop. Each sample is made with replacement, permitting certain data points to appear in a number of sample. After that, a model—typically a decision tree—is trained independently on each of these bootstrapped data. The findings of multiple models are combined to produce the final prediction, generally by averaging predictions in regression tasks or using the majority vote in classification tasks.

ISSN 2394 - 7780

8. Boosting:
Boosting is an ensemble learning strategy used in machine learning that combines the skills of numerous weak learners to increase the accuracy of models. A weak learner is a model, including a small decision tree (violation), that performs moderately better than a random guess, typically because due to its simplicity. With each new model emphasizing on the errors caused by prior ones, the improvement aims to instruct these weak learners in order. The original data is used to create the first model, then subsequent models are trained with altered data, emphasizing observations that the first models did not fully grasp. By fixing the mistakes of its predecessors, each model in the series "boosts" performance, and the final improved model aggregates all of the weak learners into a single strong model.

9. KNearest Neighbour:
The KNearest Neighbours (KNN) algorithm is a simple yet powerful method, is used for classification and regression tasks. How it works: It compares a new data point to homogenization, where "k" is the number of the closest object points in its environment. training on n neighbours is a given number of neighbours The choice of "k" affects the workings of the algorithm: a low “k” centers in on the nearest neighbours, so it is more sensitive to noise, whereas a biggish “k” takes a broader scope to the stability [4].

IV. PERFORMANCE EVALUATION AND RESULTS
The experiments are assessed using four fundamental metrics: Four basic metrics evaluate the experiments which include false positive rates (FPR), false negative rates (FNR), true positive rates (TPR), and true negative rates (TNR). Cases that receive positive classifications and are confirmed to be true positives are referred to as true positives. True negative situations receive the correct classification. False positive cases receive positive classification despite being negative. False negatives are cases that receive negative classification when they should be positive. The Naive Bayesian, Logistic Regression, SVM, Decision Tree, Random Forest, Trunk, Boost, and Kneighbor models' ANN, sensitivity, specificity, and accuracy are evaluated. How these assessment criteria are applied depends on how well they assess the unbalanced binary classification challenge.

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Sensitivity = TP / (TP + FN)

Specification = TN / (FP + TN)

Precision = TP / (TP + FP)

TP (True Positive): Number of positive cases correctly identified as positive. TN (True Negative): Number of negative cases correctly identified as negative. FP (False Positive): Number of negative cases incorrectly identified as positive. FN (False Negative): Number of positive cases incorrectly identified as negative. Sensitivity (Recall) gives precision in the classification of positive cases (fraud). The specification provides precision in classifying negative (legitimate) cases. Accuracy gives accuracy in cases classified as fraud (positive). [1]

A. RESULTS
Nine classifier models are created in this study, based on bagging, boosting, kneighbour, decision trees, random forests, naïve bayes, logistic regression, svm, and ann. To evaluate these models, 0.7% of the dataset is used for training, and 0.3% is set aside for testing and validation. Accuracy, sensitivity, specificity, and precision are used to evaluate the performance of the three classifiers. Classifier accuracy for the original dataset distribution, 0.9724:99.0262 dataset distribution, The sampled 70:30 distributions are with Over, under and combine sampling technique are presented in Tables 2,3 and 4

Table 2. Accuracy results for Over sampled data distribution

Metrices	Classifiers								
	Naïve Bayes	Logistic Regression	Decision Tree	Random Forest	SMV	ANN	Bagging	Boosting	Kneighbour
Accuracy	0.82	0.86	0.99	0.99	0.46	0.95	0.99	0.95	0.99
Precision	0.97	0.94	0.99	0.99	0.15	0.96	0.99	0.97	0.97
Sensitivity	0.67	0.76	0.99	0.99	0.02	0.94	0.99	0.93	0.99
Specificity	0.98	0.95	0.99	0.99	0.91	0.96	0.99	0.98	0.97

ISSN 2394 - 7780

Table 3. Accuracy results for Under sampled data distribution

Metrices	Classifiers								
	Naïve Bayes	Logistic Regression	Decision Tree	Random Forest	SMV	ANN	Bagging	Boosting	Kneighbour
Accuracy	0.81	0.86	0.94	0.95	0.3	0.85	0.95	0.94	0.85
Precision	0.97	0.94	0.94	0.95	0.27	0.94	0.95	0.94	0.88
Sensitivity	0.64	0.77	0.95	0.94	0.23	0.76	0.94	0.94	0.84
Specificity	0.98	0.95	0.94	0.95	0.37	0.95	0.95	0.94	0.85

Table 4. Accuracy results for Combined sampled data distribution

Metrices	Classifiers								
	Naïve Bayes	Logistic Regression	Decision Tree	Random Forest	SMV	ANN	Bagging	Boosting	Kneighbour
Accuracy	0.82	0.86	0.99	0.99	0.47	0.95	0.99	0.95	0.99
Precision	0.97	0.94	0.99	0.99	0.25	0.95	0.99	0.97	0.97
Sensitivity	0.66	0.76	0.99	0.99	0.06	0.95	0.99	0.92	0.99
Specificity	0.98	0.95	0.99	0.99	0.63	0.95	0.99	0.97	0.97

B. Comparative Analysis

The comparative analysis evaluated three sampling techniques—oversampling, under sampling, and combined sampling—using multiple classifiers. Oversampling and combined sampling consistently show the highest performance across most classifiers like decision tree, random forest, bagging, boosting, and k-nearest neighbour (KNN), with high accuracy, precision, sensitivity, and specificity. Both techniques lead to nearly identical results for these classifiers. Under sampling shows slightly lower sensitivity and accuracy for some classifiers and struggles more with class imbalance, especially with KNN and SVM. Decision tree, random forest, bagging, and boosting emerge as the top-performing classifiers in all sampling techniques. Conversely, SVM consistently struggles, particularly in sensitivity and precision, regardless of the sampling method.

Decision tree, random forest, bagging, and boosting were used for making hybrid model

V. CONCLUSION

The current study demonstrates that addressing class imbalance is crucial to successful financial transaction fraud detection. The analysis of several classifiers, including Decision Tree, Random Forest, Bagging, and Boosting, finds that they perform better than any other sampling technique, particularly when dealing with imbalanced data sets. The top performance of these models in fraud detection is due to their superior sensitivity coupled with high accuracy and precision levels. When utilizing SVM and K-Nearest Neighbors classifiers they face challenges with class imbalance especially during undersampling which results in reduced accuracy for fraud detection tasks.

The research demonstrates the importance of using combined sample techniques like undersampling and oversampling to address class imbalance issues. Hybrid models which integrate Decision Tree, Random Forest, Bagging, and Boosting offer dependable solutions that improve both accuracy and reliability for fraud detection tasks. This research offers important findings that enhance financial security frameworks by enabling the creation of fraud detection systems designed to manage dynamic and high-stakes financial transactions effectively.

ACKNOWLEDGEMENT

I would like to appreciate my college faculty and my parents for the support without this research project wouldn't have been possible.

REFERENCE

[1] A. John O, A. Adebayo O. and O. Sumuel A., Credit card fraud detection using Machine Learning, IEEE, 2017.

[2] Haziqah Shamsudin, Umi Kalsom Yusof, Andal Jayalakshmi and Mohd Nor Akmal Khalid, Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent, Sapporo, Hokkaido, Japan: 0 IEEE 16th International Conference on Control & Automation (ICCA), 2020.

ISSN 2394 - 7780

[3] Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga and Nuwan Kuruwitaarachchi, Real-time Credit Card Fraud Detection Using, IEEE, 2019.

[4] Sai Kiran , Jyoti Guru , Rishabh Kumar , Naveen Kumar, Deepak Katariya and Maheshwar Sharma, Credit card fraud detection using Naïve Bayes model based and KNN classifier, International Journal of Advance Research, Ideas and Innovations in Technology , 2018.

[5] Lingfeng Guo, Runze Song , Jiang Wu , Zeqiu Xu and Fanyi Zhao, Integrating a Machine Learning-Driven Fraud Detection System Based on a Risk Management Framework, Preprints.org, 2024.

[6] P. Manikandaprabhu, S. Prasanna, K. Sivaranjan and R. Senthilkumar, Fraudulent Banking Transaction Classification Using Deep Learning Algorithm, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) , 2023.

[7] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang and Changjun Jiang, Random Forest for Credit Card Fraud Detection, IEEE, 2018.

[8] S. Warghade, V. Patil and S. Desai, Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm, International Journal of Computer Trends and Technology (IJCTT), 2020.

[9] Mosa M. M. Megdad, Bassem S. Abu-Nasser and Samy S. Abu-Naser , Fraudulent Financial Transactions Detection Using Machine Learning, International Journal of Academic Information Systems Research (IJAISR) , 2022.