FRAUD RECOGNITION IN DIGITAL TRANSACTIONS BY USING SMOTE ALGORITHM

V. Priyadarshini[1] and A. Pushpa Latha[2]
[1]Department of Computer Science and Engineering, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India
[2]Department of CSE, SRKR Engineering College, Bhimavram, Andhra Pradesh, India
E-Mail: priyavoosala@gmail.com

ABSTRACT

Digital transactions make our lives easier. At the same time, every human is facing fraud issues by using Digital Transactions like credit cards with the growing number of transactions. Many intruders try to steal credit card details using various internet sources and cheat credit card holders. The intruders play unique tricks to cheat users, like sending trustworthy messages and emails. An enhanced fraud detection technique has become necessary to keep users sustainable to overcome the problem. An Ensemble model is constructed in this study utilizing the SMOTE algorithm to detect fraudulent transactions and alert users. The model performance is evaluated by using Machine Learning Models like KNN, Logistic Regression, and SMOTE. Among these, the SMOTE algorithm has the highest accuracy in detecting fraud.

Keywords: fraud detection (FD), digital transactions, machine learning (ML), data mining (DM), credit card fraud (CCF), synthetic minority oversampling approach (SMOTE).

1. INTRODUCTION

Data mining-based fraud detection tools are still a source of worry for people, even though they have been around for decades. The fraudulent use of a credit card to make a transaction is the main focus of the crime known as "credit card fraud." Credit card transactions can be done in person or digitally [1]. During the transaction, a credit card is required for use in the case of physical transactions. It may occur by phone or on the WWW in the case of digital transactions. The cardholder submits card information via phone or online.

Massive growth in credit card usage may be attributed to the proliferation of online shopping during the last decade [2]. For example, consumers in Malaysia have completed 320 million different types of transactions. After the allotted time has passed, it has climbed to 360 million. CCF has increased with the popularity of using these cards [3]. Despite the prevalence of many authorization methods, credit card theft continues to thrive. Online fraud is expected because the perpetrator's location and identity may be concealed. There has been a significant effect on the banking sector from the increase in credit card theft. In 2015, worldwide credit card theft cost victims a whopping $21.89 billion [4].

A. Fraud Detection

Monitoring the behaviour of a user community to anticipate, evaluate, or steer clear of improper activity constitutes fraud detection. This significant problem has attracted more interest in some fields, such as ML and data science, where the answers may be automated [4]. Because of its multifaceted nature, which includes class imbalance, this subject presents a significant learning challenge. In terms of total volume, legitimate transactions far exceed fraudulent ones. Also, the statistical features of the transaction patterns often shift over time.

Deception perpetrated to gain material or immaterial benefit is considered fraud [4]. Two different methods may be used to prevent financial loss caused by fraud. There are two types of fraud protection: preventative measures and detection methods. The best way to combat fraud is with a preventive strategy that stops it in its tracks before it ever begins. On the other side, fraud detection is essential if a criminal attempts to carry out a fraudulent transaction. But these aren't the only difficulties to overcome when putting a fraud detection system into practice in the real world. In the actual world, automated technologies will quickly scan a massive stream of payments to choose which ones to approve. Figure-1 shows a visual representation of this procedure.

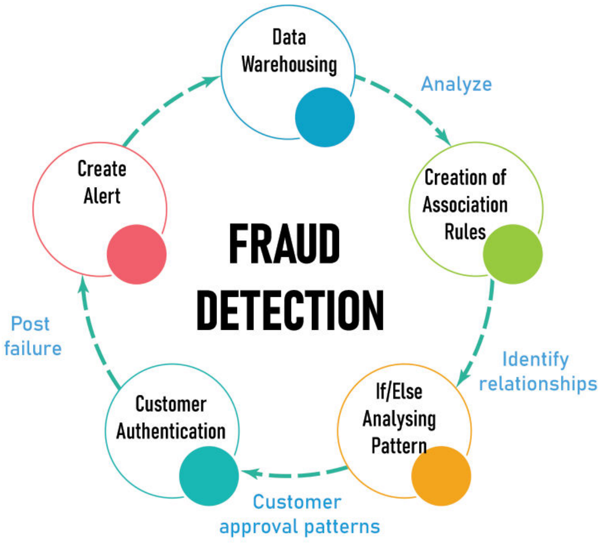

Figure-1. Fraud detection process.

To assess all the lawful transactions and those that seem fraudulent, algorithms are designed to detect fraud. Professionals evaluate these records and call cardholders to verify transactions [2, 5, 6]. When investigators supply data to the automated system, it may be utilized to train and update the algorithm, improving fraud detection over time.

B. Classifications of Credit Card Frauds

a) **Application fraud:** A thief takes control of an application, obtains a customer's credentials, and then sets up a phoney account.
b) **Electronic or manual card imprints:** In this scam, the crook takes the magnetic strip data from the card.
c) **Card not present:** The transaction with this kind of credit card does not include using a physical credit card at any point.
d) **Counterfeit card fraud:** Massive fraud in which the culprit copies a magnetic strip onto a phoney card that looks and acts like the original. This card is utilized in the commission of fraudulent acts.
e) **Lost/stolen card:** This fraudulent behaviour may occur if a cardholder fails their card or has it stolen from them. It can also take place if the cardholder misplaces their card.
f) **Card id theft:** Identity theft uses a cardholder's personal information fraudulently.
g) **Mail non-received card fraud:** There will be a method for sending mail to the receiver while supplying the credit card, and fraud may occur here by either tricking the letter or phishing the information included in the credit card.
h) **Account Takeover:** Now that they have complete control over the account holder, the fraudster will perform fraudulent acts against them.
i) **Fake fraud on the website:** Criminals that commit fraud will plant harmful code on a website so that they can carry out their plans.
j) **Merchant collision:** Without the cardholder's authorization, the merchants in this scam will disclose the cardholder's personal information to a third party or the fraudster.

The scam's perpetrators want to benefit themselves, either personally or financially, by using dishonesty [7]. Based on this, the two most important ways to avoid loss caused by fraud are identification and prevention of fraud. Contrast fraud detection, which is the reactive method of identifying fraudulent transactions by fraudsters, with fraud prevention, which is the proactive method of preventing fraudulent activities from occurring in the first place. These days, you can get just about anything you want on a card that can be used as a form of payment, from a credit or charge card to a debit or prepaid card. In certain regions, they have surpassed all other forms of currency as the preferred method of exchange. Indeed, the advent of digital technology has prepared the way for shifts in how we deal with money, particularly in payment methods, which have shifted from requiring physical action to being conducted digitally and through electronic means.

Dramatically altered the field of monetary policy, with implications for how businesses of all sizes conduct their operations. The unauthorized use of a credit card to make a purchase is known as credit card fraud. Anybody with any technology may complete these deals [8]. Credit cards are often used in in-person purchases. Digital transactions, however, are conducted remotely, often over the web or the phone. The cardholder's number, verification number, and expiry date are frequently requested over the phone or through a website. Credit card use has skyrocketed in recent years, paralleling the meteoric development of online shopping. Figure 2 shows online Fraud Detection.

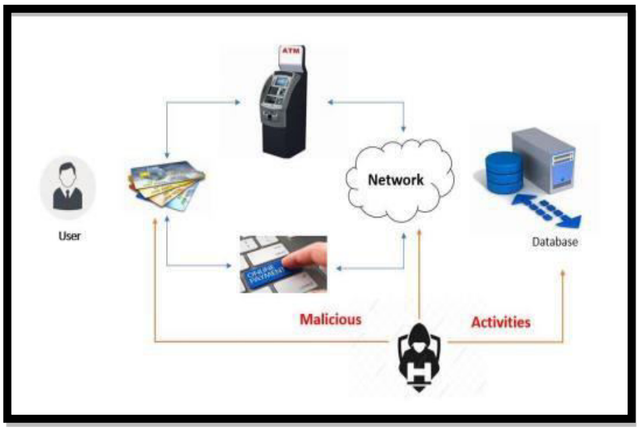

Figure-2. General scenario of online fraud.

2. RELATED WORK

CCFD was initiated by Y. Sahin *et al.* [4], who relied heavily on seven different categorization strategies. They have used DT and SVM to help reduce potential losses for financial institutions in their study. They suggested that LR classification models and ANN would be more beneficial for enhancing performance in identifying fraud. Using examples from the literature, they contrast the performance of ANN classifiers with that of LR classifiers and show why the former is superior to the latter. As a result, all detection algorithms were less effective at identifying fraudulent transactions, and the distribution of training data sets became more skewed.

Classification methods for CCFD were shown effective by A. Shen *et al.* [5], who also suggested three models based on a DT, a NN, and LR. There are other methods. However, the DT is outperformed by NN and LR. Decision-making in the face of uncertainty is addressed by the probability theory framework given by M. J. Islam *et al.* [6]. After going through the basics of the Bayesian approach, they used the k-nn and NB classifiers on the credit card system dataset.

Along with SVM, ANN, BN, HMM, KNN, FL, and DT, Y. Jain *et al.* [7] have studied many other ways to find credit card fraud. Their article found that KNN, DTnaiv, and SVM had medium. Accuracy. The accuracy of fuzzy logic and LR is the worst compared to other methods. A high detention rate may be achieved using NN, NB, fuzzy systems, and KNN. Using Logistic Regression, SVM, and DT, K. Randhawa *et al.* [8] presents a model for detecting fraud with a high detection rate in the intermediate stage. Some algorithms, such as ANNs and NB Networks, outperform others. Training one of them will set you back a significant amount of money. Each algorithm has the same major flaw: it does not provide the same result in different settings. They offer better results with one dataset but produce poor results with every other dataset. Algorithms like KNN and SVM provide excellent results with limited data sets, while algorithms like LR and fuzzy logic systems have acceptable accuracy with unprocessed and unsampled data.

N. Nisar *et al.* [9] have developed and deployed a voting-based ensemble model to identify fraudulent emails, including information about credit card transactions and details. To accommodate this, they have constructed the model in two distinct stages. Classifiers such as SVM, NB, k-NN, DT, RF, and AdaBoost were separately applied in the initial stage to identify fraudulent emails. Second, an Ensemble Voting Classifier was employed, which took the aggregated results of all the algorithms and ranked them according to the number of votes each received. H. Naik I [10] have studied various algorithms, including NB, LR, J48, and AdaBoost. Like most classification algorithms, the NB algorithm is based on the Bayes theorem, which determines whether or not an event is likely to occur. A technique known as LR may be used in predicting or forecasting the values of numerical variables. Specifically, they use the J48 method for the classification issue, which is employed to create a DT. The time factor is used to choose the optimal algorithm since they are all equally accurate. After considering timing, they determined that the Adaboost algorithm effectively spotted credit card fraud.

Twain's fundamental algorithmic procedures, which are known as the Whale Optimization Techniques (WOA) and SMOTE, have been broken down by V. Sahayasakila *et al.* [11] in their article (Synthetic Minority Oversampling Techniques). They were primarily interested in increasing the convergence speed and resolving the issue of data imbalance. Using the SMOTE and the WOA method, they solved the problem of unequal educational opportunities. All synthesized transactions are differentiated using the SMOTE method, resampled to double-check the correctness of the data, and optimized with the WOA method. The technique enhances the system's convergence speed, dependability, and efficiency.

DT, RF, SVM, and LR are all topics that R. Sailusha *et al.* [12] have covered. They've used a severely biased dataset in their research. Precision, sensitivity, specificity, and accuracy are the metrics used to assess performance. The results show that the accuracy of LR is 98.6%, that of DT is 95.5%, that of Random Forest is 98.6%, and that of the SVM classifier is 97.5%. Their comparisons found that the RF algorithm is the most accurate of the bunch, making it the top pick for detecting fraudulent activity. Consequently, they concluded that the SVM algorithm is flawed due to imbalanced data and does not improve upon previous methods of identifying credit card fraud.

R. R. Popat *et al.* [13] studied identifying fraudulent credit card use, focusing on the three major types of fraud: bank, business, and insurance. They have concentrated on the two ways credit card transactions may be made, virtually and ii) physically. LR, SVM, NN, KNN, NB, Genetic Algorithm, DT, etc., were all the centre of their attention. DM techniques such as classification, cluster analysis, prediction, outlier discovery, regression, and visualization have also been outlined. Therefore, they reasoned, all the ML approaches could guarantee a high Accuracy for the detection rate. The business world eagerly awaits the discovery of innovative strategies for cutting expenses, boosting profits, and introducing new standards.

D. Varmedja *et al.* [14] have suggested and studied several ML algorithms (CCDFM) to identify credit card fraud. MLP, LR, NB, and RF are just a few of the ML variants available. ANN with four hidden layers is employed here for MLP. The Relu activation function avoids negative values and uses Adam as the optimizer. LR yielded an accuracy of 97.46%, using a dataset of 56962 samples, of which 98% were fraudulent transactions. The accuracy scores for NB and RF on the same dataset are 99.23% and 99.96%, respectively. Overall, ANN achieved an accuracy of 99.93%, with the best result being achieved by RF in identifying CCF.

C. Jiang *et al.* [15] present a four-stage fraud-detection approach. First, they used past transaction data to group transactions with similar behaviour. After that, they devised a method to aggregate the transactions using a sliding window. This method may characterize a cardholder's typical pattern of behaviour. Feature extraction is done after aggregation in a new window. Patterns of behaviour and responsibilities are finally sorted out via categorization. Their LR with raw data (RawLR), RF with aggregation data (AggRF), and RF with feedback technique with aggregation data (AggRF +FB) approaches have 80% accuracy compared to other methods.

K. Fawagreh *et al.* [16] created ML techniques by testing them on a data set derived from the actual world. They used those algorithms to construct a super classifier via ensemble learning. After that, they examined how well their implementation of a super classifier fared compared to the performance of supervised algorithms. They implemented 10 ML algorithms: RF, LR, Stacking Classifier, XGB Classifier, Gradient Boosting, Logistic Regression, MLP Classifier, SVM, DT, KNN, and NB. And then, they compared the super classifier's output to the results in terms of accuracy, recall, precision, and confusion matrix. Since the LR provides more accurate predictions of fraudulent transactions, it was chosen.

K. Randhawa *et al.* [17] employed twelve ML algorithms to identify CCF. To do this, they track how to benchmark and real-world datasets fare. Related research describes the differences between single and hybrid models and adds that AdaBoost and majority voting approaches are used to create hybrid models. They reported the outcomes of applying their chosen twelve algorithms to both parameters (Benchmark and real-world datasets). Under benchmark data, the RF algorithm achieves the highest accuracy and sensitivity compared to other conventional algorithms employing AdaBoost and majority voting techniques. Experiments conducted using real-world data show an accuracy rate of over 90%, despite 30% noise in the dataset. Modelling competency criteria, or MCC, is the industry norm for evaluating model effectiveness.

S. Kiran *et al*. [18] have proposed a CCDFM titled "Naïve Bayes improved K-nearest Neighbor method (NBKNN)". The results of the experiments show that the two classifiers operate differently on the same dataset. They have used the European cardholders' transaction dataset, which contains two-day data. For this dataset, KNN and Naïve Bayes have detected credit card fraud transactions with 90% and 95% accuracy, respectively. Sara *et al*. [19] have implemented a CCDFM to prevent and detect fraud. They have considered class imbalance the most well-known and critical issue in their research. They have conducted an experimental study with various solutions for tackling the class imbalance problem. In that process, they identified that imbalanced classification methods are ineffective, specifically for large datasets.

F. Carcillo *et al*. [20] combined supervised and unsupervised CCF detection approaches. In this work, they have produced unsupervised outlier scores at various levels of granularity by following the best-of-both-worlds principle. These scores have been tested with a real, annotated CCDF dataset. The results have proved that the combination is effective in accurate fraud prediction.

A model for the identification of fraud that is based on DL has been suggested by Chen *et al*. [21]. They have structured the model into three stages: based before, during, and after applying the model. Within 24 hours, 5 million transaction data records were utilized for implementation. The statistics are skewed since there are only 6,223 fraudulent transactions, despite their 5,000,000 total. They have used DCNN to build a model that can anticipate fraudulent activity with 99% certainty. The authors also noted that the model holds up well with more extensive datasets.

E. Esenogho *et al*. [22] have proposed a fraud identification approach for credit cards with different classifiers such as LSTM and SMOTE-Edited Nearest Neighbours (SMOTE-ENN) and also tested the model's performance against the classifiers SVM, MLP, AdaBoost, and LSTM. According to the experimental findings, classifiers worked better when trained on resampled data. The suggested LSTM ensemble surpassed the other techniques by achieving sensitivity and specificity values of 0.996 and 0.998, respectively.

A. Alharbhi *et al*. [23] have implemented a deep learning-based fraud detection model with text-to-image conversion. After converting text into images, the images are fed into CNN architecture for class imbalance problem resolution. The proposed model utilized image processing and ML and DL techniques for fraud detection. For their research, they used 284,315 records datasets with 28 variables. Initially, they converted these variables into image features with time and amount as the first and last features. After that, they applied the CNN approach to finding the fraud and obtained good results in detecting fraudulent transactions.

John O. Awoyemi *et al*. [24] proposed a hybrid ML approach that detects fraud on bank cards. This approach has been presented with the main focus on two main issues to resolve three ML methods. On the skewed data, a hybrid process consisting of under-sampling and over-sampling is carried out, and both of these techniques are applied to raw and preprocessing data. A hybrid method is used to sample the extremely unbalanced dataset to get distinct data distributions, with the positive class being oversampled and the harmful category being undersampled. The effectiveness, measured at 97.92%, was attained by the KNN method. The extreme data imbalance hurts the sampling strategies used by meta-classifiers and learning while obtaining data.

Changjun Jiang *et al*. [25] introduced an AggRF method to train the data with affected and unaffected cardholders. The proposed approach is a combination of RF and feedback techniques. Initially, the data is clustered according to the holder card behaviour. Then it extracted the data according to the cluster with patterns. It undergoes behavioural patterns and assignments to form a priority queue and has genuine or fraudulent data. The proposed approach has achieved 90%. Compared with two different methods of ML. In this approach, the periodic time is high because of single window execution.

Pumsirirat *et al*. [26] have used a DL Neural Networks of high-levelled technique to identify fraud detection. By using AE and RBM, regular transactions are detected automatically. But it does not contain high efficiency when the dataset has many records, so the proposed approach has a slight change based on Keras, RBM, and H2O. These unsupervised learning techniques extract the meaningful features of the data automatically. AE is used for backpropagation where parameter gradients are realized. RBM is reconstructed where visible and hidden layers are input and output to see the data weights to train them. To find the efficiency of the approaches, the authors have compared two methods in DL, but the proposed system has achieved 96% accuracy in AE and 95% in RBM.

E. Duman *et al*. [27] have chosen two DL approaches to detect fraud in bank cards when online transactions happen. SVM and DT are the two methods used to detect the process. Even DT contains c&rt, c5.0, and CHAID for methods extraction. This c&rt is used to measure c5.0 for gain ratio in impurity and CHAID for splitting and merging. SVM uses kernel functions which are used for training the data and overfitting. It increases the value to improve the classification accuracy. 70% and 30% of data are divided into training and testing. The records are the same for both samples. Therefore, a comparison of the model for each example becomes possible. SVM has achieved 99% accuracy compared to the DT has gained 92%. A Comparison can be made on whether the prediction is sufficient.

Gajendra Singh *et al*. [28] chose the SVM approach to detect fraud attempts performed in transactions using or without cards. To solve this issue, many authors have introduced many methods but using SVM TP and TN have achieved a high-efficiency rate, and FP and FN have low efficiency. Here SVM operates kernel is used for feature space for the data in input space. According to the model, the kernel uses three types: linear, polynomial, and Gaussian. The process starts with reading the data, categorizing, vectorization of fields, and separate the data according to the true and false transaction groups, selecting one of the kernels as discussed, training the SVM data, saving the classifier to read and repeating the initial three steps for present transactions. Then place the classifier and current vector in the classifier. Finally, generate the decision of the classifier. The complete results are saved in RBF. According to it, the efficiency of the approach gains 97%. The main difficulty in the process is using the proper kernel type for the data.

R.S. Shankar *et al* [29] has done their research on various ML and DL models Domain. They have used Epilepsey based on EEG signals [30] and done classification of gender by voice [31] and HMM [32]. They also used NLP to [33] predict stock exchange, Election results [34], feature selection using DRG [35], predict staff attrition [36] using DM techniques, Extract tweets from social media [37], noise retroaction analysis [38] and evaluating easy using system tools [39]. They have designed a parallel computing framework in memory [40] and a comprehensive simulation of the entire performance through recommended scheme examined with the Hadoop scheme [41] and they have used brain-inspired KGS to evaluate the performance [42], optimal data delivery in the cloud [43] and among various AES models to evaluate performance [44].

S Deepthi Kavila *et al.* [45] chose three different ML approaches for credit card detection in the e-commerce industry. Here the author has chosen a large dataset to show that the proposed approach achieves a high accuracy rate. The author has chosen positive classes as fraud transactions and negative as genuine transactions. This process starts with reading the data, oversampling, and then dividing the data to train and test practice and feature selection. Now five ML approaches are chosen each carries the performances. Then travels to test samples to verify the ability to predict the outcomes of the pushed data, and finally, performances are evaluated according to the prediction values. The performance is assessed on the bases of four categories. Therefore, in comparing the three approaches, the random forest has the best accuracy rate of 95% and the remaining two hold 94% and 90%.

V N Dornadula *et al.* [46] have used a sliding window, and SMOTE method is applied to ML to the imbalanced dataset. Initially, data is divided according to the groups based on the transaction amount. By using a sliding window, the features are extracted from behavioural patterns. If a new transaction enters will processing, it undergoes the previous step and mains the data in the process. For training the data, the author has to choose a different classifier for each cluster because of patterns in that group. Now SMOTE is operated on the dataset. This work is completed on the imbalanced dataset for those two ways that have been chosen. Firstly, Matthew coefficient correlation on original data; secondly, only one class is classified. Finally, the training group applied to the cardholder group in this high rating is considered the recent behavioural pattern. Now, the feedback system is used for the updation of rating scores. The proposed approach has been compared with four methods in that random forest and DT have achieved a 99.94% accuracy rate.

3. METHODOLOGY

a. Data collection

The dataset consists of various CC transactions that the European CC holders do. The transactions belonging to September 2013 for two days and overall transactions are 284,807 transactions. Among those, 492 are fraud transactions, and the remaining are nonfraud transactions; overall, 0.173% of transactions are frauds, i.e., positive class. The attributes present in the dataset are illustrated in Table-1.

Table-1. Dataset attributes description.

S. No	Attributes	Description
1	V1 to v28	These are the principal values gained from the PCA transformation. With some confidentiality problems, the original features are not provided.
2	Time	The 'time' feature shows seconds between each transaction and the commencement of the data collection.
3	Amount	The amount is the variable that describes the transaction amount.
4	Class	The 'Class' attribute will have two values ', 0' and '1', representing fraud and nonfraud transactions, respectively.

b. Objectives

This research is to detect frauds of credit cards using advanced techniques like ML classifiers. 'Fraud' in credit card transactions can be described as unauthorized usage of an account by others instead of the account owner. Preventative actions might be taken to halt fraudulent behaviours and avoid repeat incidents. The main objectives of this research are:

a) If a card is stolen and then used fraudulently, it may be used up to the point when its limit has been reached. Consequently, a solution that reduces the overall limitation on cards vulnerable to fraud is more important than the number of adequately categorized transactions.

b) This study uses an Ensemble classifier with a complex voting method and interval-valued parameters to minimize false alarms.

c. System architecture

Developing practical fraud detection algorithms is essential if these losses are curtailed. An increasing number of these algorithms depend on increasingly complex forms of ML to assist investigators. Figure 3 represents how the system was implemented for the project and how well it functions in terms of detection and optimizing the average accuracy score.

Data preprocessing activities occur when a database is loaded to pick out duplicate and inconsistent records. As records are preprocessed, the records attain training to find fraud detection levels "are there any malpractices?". This process may be repeated until the final transaction falls below the loop rotation. As data is trained, it offers an unbalanced dataset problem. We can first balance the training data using a resampling approach (SMOTE), Logistic Regression, and KNN model by optimizing the average precision rating. The way how CCDFM works.

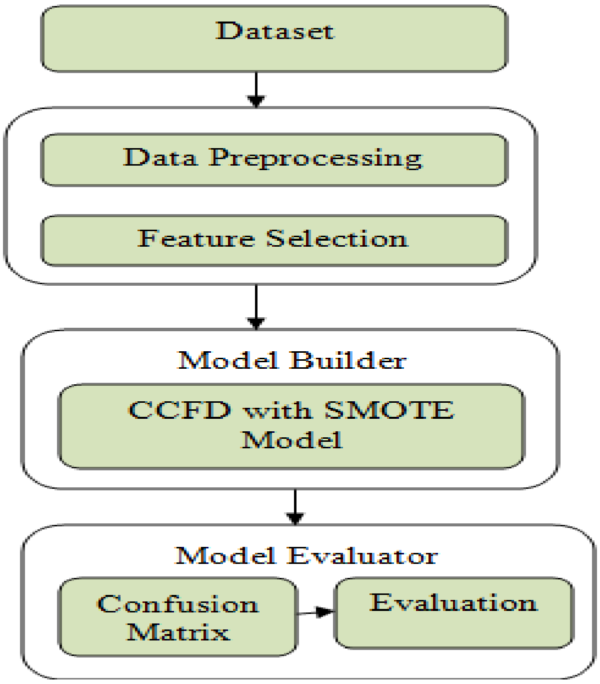

Figure-3. System architecture.

d. Existing models

First, the Internet platform can create user, environment, and behaviour portraits based on the relationship network's users, backgrounds, and behaviors. An ML technique is used to discover aberrant points in the relationship network in real-time to detect fraudulent activity. To detect credit card fraud, the classifiers used in this model are:

- Logistic Regression
- KNN
- SMOTE Algorithm

Logistic Regression

In statistics, LR evaluates a set of plausible variables that may lead to a binary outcome. Explains the influence that the factors under consideration had on the dependent variable that was looked at. On the other hand, multinomial LR is used if the explanatory variables involve a minimum of three unsorted subgroups. Figure 4 depicts the logistic regression process in action.

Our challenge is supervised binary classification, as we have previously discussed; the dataset contains instances. Each example consists of input and output to train the model and predict the output of a new example based on input characteristics. In our work, we have used y ∈ {0, 1} to refer to the output, where "1" represents fraud and "0" means nonfraud. The model has used 31 features: time, amount, class, and 28 other PCA-transformed attributes, as illustrated in Table-1.

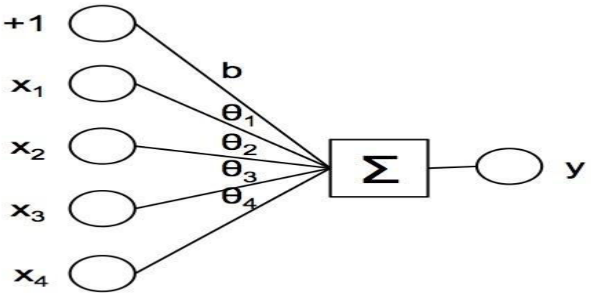

Figure-4. Logistic regression.

KNN

Using the idea of 'nearest neighbour analysis,' several different anomaly detection methods have been developed. The efficiency of the KNN algorithm is primarily affected by three factors:

- The distance measure is used to pinpoint the individuals geographically closest to one another.
- The distance rule determines how k neighbours are grouped.
- The new sample was categorized according to how many neighbours it had.

Everyday use of the K-nearest neighbour technique is in detecting systems. It is further shown that KNN performs very well in systems designed to identify credit card fraud using supervised learning approaches. This approach will use the KNN category to categorize the new instance query. You can see how the KNN operates in Figure-5.

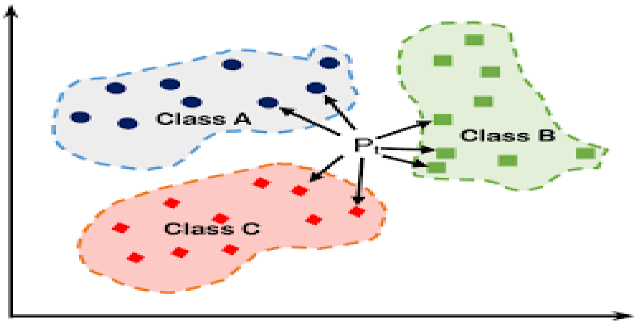

Figure-5. KNN Algorithm.

According to the research problem, the class attribute has to be predicted depending on time, amount, and v1 to v28 PCA transformed characteristics. In this process, the outlier analysis has been used. If the transaction is an outlier, it is classified as fraudulent. Otherwise, it's a normal one. In this model, "class" is the dependent attribute and other attributes are independent.

e. Proposed model

SMOTE Algorithm

As an ML solution, SMOTE (synthetic minority oversampling approach) balances out situations when one group is much larger than another. It is used to find a solution to the issue of the uneven distribution of class. An imbalance in the classes exists when the incidence of one class is much higher than the occurrence of the other classes. With the help of SMOTE, we can tell the fake ones apart from the actual ones and create new ones somewhere between. Accuracy, confusion, precision, recall, and support are metrics used to assess the algorithm's effectiveness. The SMOTE routines are utilized to scale down the minority instances to differentiate the fake from the actual cases. Parameters from the SMOTE function are used to create synthetic examples.

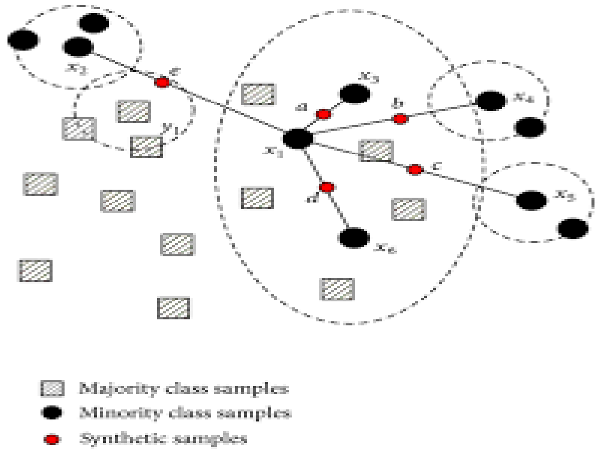

Figure-6. SMOTE model.

SMOTE's primary function is to detect fraudulent purchases made by a cardholder. The system's other primary objective is to hasten the convergence process and even out the data distribution. The method of SMOTE model is shown in Figure-6. The relationship between accurate positive and false positive rates may be seen in the receiver operating characteristics (ROC).

During two days in September 2013, European cardholders made purchases using their credit cards, which are included in the dataset. Out of a total of 284,807 cases, there are 492 instances of fraud. The positive class (frauds) only makes up 0.172% of all trades.

When the data is skewed, the SMOTE method is beneficial. The dataset in this model is unbalanced, making the SMOTE technique the most appropriate when creating a model to forecast fraud transactions.

The algorithm's steps are described below:

Algorithm: 1

Step 1	**Start**
Step 2	Take the collection of credit card transactions T. Here transaction t_iIt can be fraud or nonfraud.
Step 3	From the dataset, preprocess the unnecessary data such as
	A. Remove the missing data transaction t_i from the dataset, B. Remove attributes of any transaction t_i not needed in the model, C. Select remaining attributes as model features after step 3. A and 3. B
Step 4	After Step 3, Split the data into the Training and Testing set
Step 5	Train the model with a classifier using training data
Step 6	Test the model using the test set by forecasting the outcome of the test set transactions T and comparing it with the original work.
Step 7	Evaluate the model with the metrics

We experimented with a variety of oversampling classes in this, ranging from low to balanced oversampling. In the experiments, we looked at the relationships between the features and the type and between them. The fraud class will hereafter be referred to as the minority class, while the not-fraud type will be referred to as the majority class. We use different oversampling proportions in our experiment. Then we classify data and use matrices to run the model.

4. RESULTS

Before training the model, model performance (optimization) must be measured. The proportion of accurate predictions is used to quantify predictive model performance. Let's review two standard metrics for model evaluation.

Precision = TP/(TP+FP)
Recall = TP/(TP+FN)

We utilize 80% of the data for training and just 20% for testing. There are five distinct groups of information used for training. For this two-class problem, we construct a Logistic Regression model. We use a random search strategy to determine the ideal hyperparameters, such as regularising L1 or L2 and regularisation penalty.

To ensure that our training data is evenly distributed between classes, we use SMOTE (Synthetic Minority Over-sampling Technique). To generate more "good" instances, SMOTE is an oversampling method. Our training dataset has an equal amount of fraud and legitimate transactions when SMOTE is applied.

`Over-sampling captures 92% of fraud at the expense of 2% more ordinary transactions being reported as fraud (FP). It's a choice between missing numerous copies and erroneously halting frequent transactions to identify fraudulent transactions. Table-2 compared KNN, LR, and SMOTE with various Evaluation Metrics. Figure-7 shows the graph of evaluation metrics compared with the Models, and Figure-8 shows the Accuracy of the Models.

Table-2. Comparison table with performance evaluation metrics.

Mode/ Metrics	Accuracy	Precision	Recall	Specificity	F1-Score
Logistic Regression	98.72	0.06	0.87	0.87	0.71
KNN	97.83	0.86	0.55	0.91	0.67
SMOTE	99.91	1.00	0.98	0.55	0.12

The dataset used in this research is highly imbalanced. Because of this, we are getting very high Accuracy and AUC scores. The confusion matrix shows us that our model predicts non-fraudulent transactions more efficiently than fraudulent transactions (the F1 score for non-fraudulent transactions is 100%, whereas it is 72%).

We have 99.9% of non-fraudulent transactions and only 0.172% of fraudulent transactions. The objective is to identify fraudulent transactions correctly. We cannot afford to miss predictions of so many fraudulent transactions. The model can predict only 62% of fraudulent transactions (precision score). After changing the threshold to 10%, the F1 score value for identifying fraudulent transactions has improved to 81% from 71% though AUC remains the same.

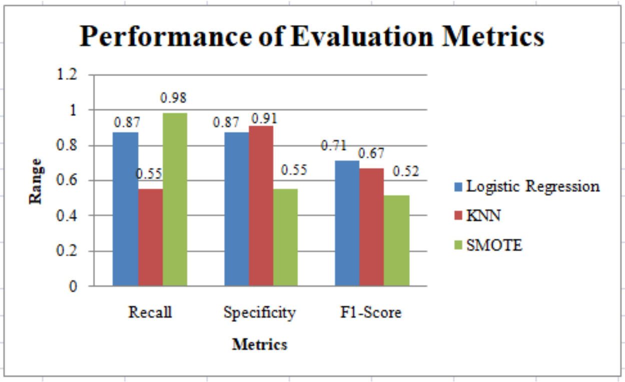

Figure-7. Graph for evaluation metrics.

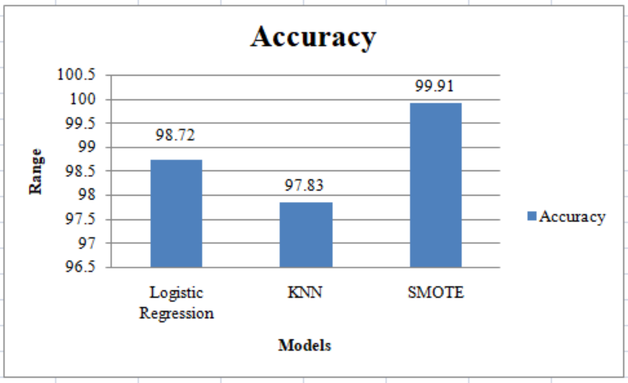

Figure-8. Comparison graph with various models on accuracy.

5. CONCLUSIONS

Fraud is rising due to the development of contemporary technology and the establishment of global communication superhighways, resulting in annual losses of tens of billions of dollars. Even though fraud prevention systems are the best way to stop fraud, con artists are innovative and will find a way to get around them eventually. If we want to catch fraudsters when prevention fails, we need reliable methods of detecting fabrication. Statistics and ML are potent tools for detecting fraud and have successfully identified activities including money laundering, e-commerce, credit card fraud, and more. In this study, we show how the Ensemble method, which is better than other methods, can be used to find credit card fraud. For Fraud transactions, the accuracy is 99.91%. By analyzing these metrics, the proposed algorithms perform better. As observed, there is a significant increase in performance. However, this result was generated over comparatively old data and had a high imbalance.

REFERENCES

[1] O. Adewumi and A. A. Akinyelu. 2017. A survey of machine-learning and nature inspired based credit card fraud detection techniques. Int. J. Syst. Assurance Eng. Manage. 8(2): 937-953.

[2] Srivastava A., Kundu S. Sural and A. Majumdar. 2008. Credit card fraud detection using hidden Markov model. IEEE Trans. Depend. Sec. Comput. 5(1): 37-48.

[3] The Nilson Report. [Online]. Available: https:// www. nilson report.com /upload/ content_promo/ The_Nilson_Report_10-17-2016.pdf. 2016.

[4] Y. Sahin, S. Bulkan and E. Duman. 2013. A cost-sensitive decision tree approach for fraud detection. Expert Syst. Appl., 40(15): 5916-5923.

[5] A. Shen, R. Tong and Y. Deng. 2007. Application of classification models on credit card fraud detection. Service Systems and Service Management 2007 International Conference. pp. 1-4.

[6] M. J. Islam, Q. M. J. Wu, M. Ahmadi and M. A. SidAhmed. 2007. Investigating the Performance of Naive-Bayes Classifiers and K-Nearest Neighbor Classifiers. IEEE International Conference on Convergence Information Technology. pp. 1541-1546.

[7] Y. Jain, N. Tiwari, S. Dubey and S. Jain. 2019. A comparative analysis of various credit card fraud detection techniques. Int. J. Recent Technol Eng. 7(5S2): 402-407.

[8] K. Randhawa, C. K. Loo, M. Seera C. P. Lim and A. K. Nandi. 2018. Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, pp. 14277-14284.

[9] N. Nisar, N. Rakesh and M. Chhabra. 2021. Voting-Ensemble Classification for Email Spam Detection. 2021 International Conference on Communication information and Computing Technology (ICCICT). pp. 1-6.

[10] H. Naik and P. Kanikar. 2019. Credit card fraud detection based on machine learning algorithms. International Journal of Computer Applications. 182(44): 8-12.

[11] Sahayasakila D., Aishwarya Sikhakolli and V. Yasaswi. 2019. Credit card fraud detection system using smote technique and whale optimization algorithm. Int. J. Eng. Adv. Tec. (IJEAT). 8(5): 190-192.

[12] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao. 2020. Credit card fraud detection using machine learning. In 2020 4th international conference on intelligent computing and control systems (ICICCS). pp. 1264-1270.

[13] R. R. Popat and J. Chaudhary. 2018. A survey on credit card fraud detection using machine learning. In 2018 2nd international conference on trends in electronics and informatics (ICOEI). pp. 1120-1125.

[14] Varmedja M., Karanovic S., Sladojevic M., Arsenovic and A. Anderla. 2019. Credit card fraud detection-machine learning methods. In 2019 18th International Symposium, INFOTEH-JAHORINA (INFOTEH). pp. 1-5.

[15] Jiang J. Song, G. Liu, L. Zheng and W. Luan. 2018. Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism. IEEE Internet of Things Journal. 5(5): 3637-3647.

[16] K. Fawagreh, M. M. Gaber, and M. Abdalla. 2022. Pruned Random Forests for Effective and Efficient Financial Data Analytics. In Financial Data Analytics. pp. 225-249.

[17] K. Randhawa C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi. 2018. Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, pp. 14277-14284.

[18] S. Kiran, J. Guru, R. Kumar, N. Kumar, D. Katariya and M. Sharma. 2018. Credit card fraud detection using Naïve Bayes model-based and KNN classifier. International Journal of Advance Research, Ideas and Innovations in Technology. 4(3): 44.

[19] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine. 2019. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, pp. 93010-93022.

[20] Carcillo Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé and G. Bontempi. 2021. Combining unsupervised and supervised learning in credit card fraud detection. Information sciences. pp. 557, 317-331.

[21] J. I. Z. Chen and K. L. Lai. 2021. Deep convolution neural network model for credit-card fraud detection and alert. Journal of Artificial Intelligence. 3(02): 101-112.

[22] Esenogho I. D. Mienye, T. G. Swart, K. Aruleba and G. Obaido. 2022. A Neural Network Ensemble with Feature Engineering for Improved Credit Card Fraud Detection. In IEEE Access. 10: 16400-16407.

[23] A. Alharbi, M. Alshammari, O. D. Okon, A. Alabrah, H. T. Rauf, H. Alyami and T. Meraj. 2022. A Novel text2IMG Mechanism of Credit Card Fraud Detection: A Deep Learning Approach. Electronics. 11(5): 756.

[24] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare. 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. Proceedings of the IEEE International Conference on Computing, Networking and Informatics. pp. 1-9.

[25] Jiang J. Song, G. Liu, L. Zheng and W. Luan W. 2018. Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism. IEEE Internet of Things Journal. 5(5): 3637-3647.

[26] A. Pumsirirat and L. Yan. 2018. Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine. International Journal of Advanced Computer Science and Applications. 9(1): 18-25.

[27] Y. Sahin and E. Duman. 2011. Detecting credit card fraud by decision trees and support vector machines. International Multi-Conference of Engineers and Computer Scientists. 1: 442-447.

[28] R. Khan, G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel and A. Riyaz. 2012. A Machine Learning Approach for Detection of Fraud based on SVM.

[29] Shankar R. S., Priyadarshini V., Neelima P., Raminaidu C. H. Analyzing Attrition and Performance of an Employee using Machine Learning Techniques. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2021 Dec 2 (pp. 1601-1608). IEEE.

[30] Shankar R. S., Raminaidu C. H., Raju V. S., Rajanikanth J. 2021. Detection of Epilepsy based on EEG Signals using PCA with ANN Model. InJournal of Physics: Conference Series, (2070(1): 012145). IOP Publishing.

[31] Shankar S., Raghaveni J., Rudraraju P., Vineela Sravya Y. 2020. Classification of gender by voice recognition using machine learning algorithms. Journal of Critical Reviews. 7(9): 1217-29.

[32] Vmnssvkr Gupta, R. Shiva Shankar, Harika Devi Kotha, J. Raghaveni. 2020. Voice Identification in Python Using Hidden Markov Model. International Journal of Advanced Science and Technology. 29(06): 8100-8112.

[33] Jyothirmayee S., Dilip Kumar V., Someswara Rao C., Shiva Shankar R. 2019. Predicting stock exchange using supervised learning algorithms. International Journal of Innovative Technology and Exploring Engineering. 9(1): 4081-90.

[34] Kambhampati Kalyana Kameswari, J. Raghaveni, R. Shiva Shankar, Ch. Someswara Rao. 2019. Predicting Election Results using NLTK. 9(1): 4519-4529.

[35] Shiva Shankar R., Ravibabu D. 2019. Digital Report Grading Using NLP Feature Selection. InSoft Computing in Data Analytics, (pp. 615-623). Springer, Singapore.

[36] Shankar R. S., Rajanikanth J., Sivaramaraju V. V., Murthy K. V. 2018. Prediction of employee attrition using datamining. In2018 ieee international conference on system, computation, automation and networking (icscan), (pp. 1-8). IEEE.

[37] Shankar R. S., Murthy K. V., Rao C. S., Gupta V. M. 2018. An approach for extracting tweets from social media factors. In 2018 ieee international conference on system, computation, automation and networking (icscan), (pp. 1-7). IEEE.

[38] Shankar R. S., Srinivas L. V., Ravibabu D., Raminaidu C. 2018. Novice Retroaction Report. ARPN Journal of Engineering and Applied Sciences. 13(24): 9746-9753.

[39] Shankar R. S., Babu D. R., Murthy K. V., Gupta V. 2017. An approach for essay evaluation using system tools. In2017 International Conference on Innovative Research in Electrical Sciences (IICIRES), (pp. 1-9). IEEE.

[40] R. Shiva Shankar, V. Priyadarshini, J. Raghaveni and Ch. Vinod Varma. 2019. Efficient Spatial Data Management by Apache Spark. 14(23): 4097-4103.

[41] Deshai N., Shankar R. S., Sravani K., Ravibabu D. 2019. A Developed Task Allotments Policy for Apache Hadoop Executing in the Public Clouds. In2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), (pp. 1-4). IEEE.

[42] Shankar R. S., Deshai N., Murthy K. S., Gupta V. M. 2019. The Source of Growing Knowledge by Cognitive Artificial Intelligence. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), (pp. 1-6). IEEE.

[43] Mahesh G., Rao V. M., Shankar R. S., Sirisha G. V. 2017. Primal-dual parallel algorithm for optimal content delivery in cloud CDNs. In2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), (pp. 1-6). IEEE.

[44] Gupta V. M., Murthy K. V., Yesubabu A., Shankar R. S. 2012. Recent performance evaluation among various AES algorithm-MARS, RC6, RIJNDAEL, SERPENT, TWOFISH. International Journal of Science and Advanced Technology. 2(2).

[45] Lakshmi S. V., Kavilla S. D. 2018. Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research. 13(24): 16819-24.

[46] N. Dornadula and S. Geetha. 2019. Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science. 165, pp. 631-641.