Credit Card Fraud Detection through Class Balancing Framework and Machine Learning Algorithms

[1] Aparna Suresh, [2] S.V Annlin Jeba
[1]PG Student, [2] Professor
[1]CSE, Sree Buddha College of Engineering,Pattor,
[2]CSE, Sree Buddha College of Engineering,Pattor,
[1] aparnasureshsm@gmail.com, [2] sureshannlin@gmail.com

ABSTRACT

Nowadays, people use credit cards for online transactions as it provides an efficient and easy-to-use facility. With the increase in usage of credit cards, the capacity of credit card misuse has also enhanced. Credit card frauds cause significant financial losses for both credit card holders and financial companies. In this paper, the main aim is to detect such frauds, including the accessibility of public data, high-class imbalance data, the changes in fraud nature, and high rates of false alarm. The relevant literature presents many machines learning based approaches for credit card detection, such as Extreme Learning Method, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and XG Boost. However, due to low accuracy, there is still a need to apply state of the art deep learning algorithms to reduce fraud losses. The main focus has been to apply the recent development of deep learning algorithms for this purpose, also PCA and SMOTE are used for Feature Selection and Data Balancing. PCA is the widely used tool in data analysis and in machine learning for predictive models. It will be more called as a dimensionality reduction method, then as a feature selection method. SMOTE is commonly used oversampling methods to solve the imbalance problem. SMOTE is used to generate artificial or synthetic samples for the minority class. Comparative analysis of both machine learning and deep learning algorithms was performed to find efficient outcomes. Machine learning algorithms are applied to the dataset, which improved the accuracy of detection of the frauds.

Keywords: Fraud detection, deep learning, machine learning, online fraud, credit card frauds, transaction data analysis

INTRODUCTION

Credit card fraud (CCF) is a type of identity theft in which someone other than the owner makes an fraud transaction using a credit card or account details. A credit card that has been stolen, lost, or counterfeited might result in fraud. Card-not-present fraud, or the use of your credit card number in e-commerce transactions has also become increasingly common as a result of the increase in online shopping. Increased fraud, such as CCF, has resulted from the expansion of e-banking and several online payment environments, resulting in annual losses of billions of dollars. In this era of digital payments, CCF detection has become one of the most important goals. As a business owner, it cannot be disputed, that the future is heading towards a cashless culture. As a result, typical payment methods will no longer be used in the future, and therefore they will not be helpful for expanding a business. Customers will not always visit the business with cash in their pockets. They are now placing a premium on debit and credit card payments.

As a result, companies will need to update their environment to ensure that they can take all types of payments. The goal of supervised CCF detection is to create a machine learning (ML) model based on existing transactional credit card payment data. The model should distinguish between fraudulent and nonfraudulent transactions, and use this information to decide whether an incoming transaction is fraudulent or not. The issue involves a variety of fundamental problems, including the system's quick reaction time, cost sensitivity, and feature pre-processing. ML is a field of artificial intelligence that uses a computer to make predictions based on prior data trends. ML models have been used in many studies to solve numerous challenges. Deep learning (DL) algorithms applied applications in computer network, intrusion detection, banking, insurance, mobile cellular networks, health care fraud detection, medical and malware detection, detection for video surveillance, location tracking, Android malware detection, home automation, and heart disease prediction. We explore the practical application of ML, particularly DL algorithms, to identify credit card thefts in the banking industry in this paper.

RELATED WORKS

In the field of CCF detection, several research studies have been carried out. This section presents different research studies revolving around CCF detection. Moreover, we strongly emphasise the research that reported fraud detection in the problem of class imbalance. Many techniques are used to detect credit cards. Therefore, to study the most related work in this domain, the main approaches can be categories, such as DL, ML, CCF detection, ensemble and feature ranking, and user authentication approaches.

The paper ''An Efficient real time model for credit card fraud detection based on Deep Learning'' [1] briefs thatIn the last decades Machine Learning achieved notable results in various areas of data processing and classification, which made the creation of real-time interactive and intelligent systems possible. The accuracy and precision of those systems depends not only on the correctness of the data, logically and chronologically, but also on the time the feed-backs are produced. It focuses on one of these systems which is a fraud detection system. In order to have a more accurate and precise fraud detection system, banks and financial institutions are investing more and more today in perfecting the algorithms and data analysis technologies used to identify and combat fraud. Therefore, many solutions and algorithms using machine learning have been proposed in literature to deal with this issue.

''Facilitating User Authorization from Imbalanced Data Logs of Credit Cards Using Artificial Intelligence,''[2] briefs that an effective machine learning implementation means that artificial intelligence has tremendous potential to help and automate financial threat assessment for commercial firms and credit agencies. The scope of this study is to build a predictive framework to help the credit bureau by modelling/assessing the credit card delinquency risk.

"Auto loan fraud detection using dominance-based rough set approach versus machine learning methods,''[3], briefs thatfinancial fraud is escalating as financial services and operations grow. Despite preventive actions and security measures deployed to mitigate financial fraud, fraudsters are learning and finding new ways to get around fraud prevention systems, thereby, challenging quantitative techniques and predictive models..

The "Interleaved Sequence RNNs for Fraud Detection,''[4], Payment card fraud causes multibillion dollar losses for banks and merchants worldwide, often fueling complex criminal activities. To address this, many real-time fraud detection systems use tree-based models, demanding complex feature engineering systems to efficiently enrich transactions with historical data while complying with millisecond-level latencies.

‘‘Adversarial Attacks for Tabular Data: Application to Fraud Detection and Imbalanced Data,’’ [5], Guaranteeing the security of transactional systems is a crucial priority of all institutions that process transactions, in order to protect their businesses against cyberattacks and fraudulent attempts. Adversarial attacks are novel techniques that, other than being proven to be effective to fool image classification models, can also be applied to tabular data.

‘‘Credit Card Fraud Detection Using Machine Learning,’’[6], The usage of credit cards for online and regular purchases is exponentially increasing and so is the fraud related with it. A large number of fraud transactions are made every day. Various modern techniques like artificial neural network Different machine learning algorithms are compared, including Logistic Regression, Decision Trees, Random Forest, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, and K-means clustering etc. are used in detecting fraudulent transactions.

‘‘Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks,’’ [7], With the increasing use of credit cards in electronic payments, financial institutions and service providers are vulnerable to fraud, costing huge losses every year. The design and the implementation of efficient fraud detection system is essential to reduce such losses. However, machine learning techniques used to detect automatically card fraud do not consider fraud sequences or behavior changes which may lead to false alarms. In this paper, we develop a credit card fraud detection system that employs Long Short-Term Memory (LSTM) networks as a sequence learner to include transaction sequences.

‘‘Ensemble of deep sequential models for credit card fraud detection,’’ [8] In the recent years, the fast development of e-commerce technologies made it possible for people to select the most desirable items in terms of suggested price, quality and quantity among various services, facilities, shops and stores from all around the world. However, it also made it easier for fraudsters to abuse this huge opportunity. As credit card has become the most popular mode of payment, the fraudulent activities using credit card payment technologies are rapidly increasing as a result.

‘‘Deep Residual Learning for Image Recognition,’’ [9] Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth.

.“Fraud detection for job placement using hierarchical clusters-based deep neural networks,’’ [10] Fraud detection is becoming an integral part of business intelligence, as detecting fraud in the work processes of a company is of great value. Fraud is an inhibitory factor to accurate appraisal in the evaluation of an enterprise, and it is economically a loss factor to business. Previous studies for fraud detection have limited the performance enhancement because they have learned the fraud pattern of the whole data. This paper proposes a novel method using hierarchical clusters based on deep neural networks in order to detect more detailed frauds, as well as frauds of whole data in the work processes of job placement.

PROPOSED SYSTEM

The goal of supervised CCF detection is to create a machine learning (ML) model based on existing transactional credit card payment data. The model should distinguish between fraudulent and nonfraudulent transactions, and use this information to decide whether an incoming transaction is fraudulent or not. The issue involves a variety of fundamental problems, including the system’s quick reaction time, cost sensitivity, and feature pre-processing. Principal Component Analysis is an unsupervised feature reduction method for projecting high dimensional data into new lower dimensional representation of data which describes the variance in the data. SMOTE is one of the methods used in oversampling to solve the imbalance problem.It will balance the class distribution by randomly increasing the minority class by replicating them.

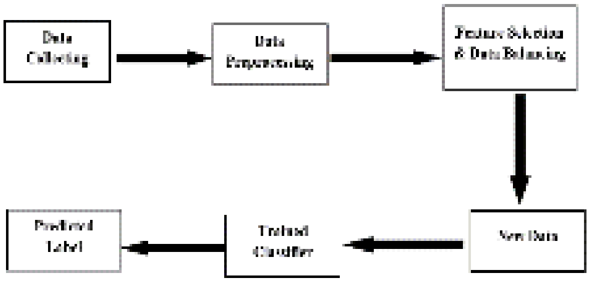


Fig 1: Architecture of Credit Card Fraud Detection

The main aim is to detect fraudulent transactions using credit cards with the help of ML algorithms and deep learning algorithms. Feature selection algorithms are used to rank the top features from the CCF transaction dataset, which help in class label predictions.The deep learning model is proposed by adding a number of additional layers that are then used to extract the features and classification from the credit card farad detection dataset.To analyse the performance CNN model, apply different architecture of CNN layers.To perform a comparative analysis between ML with DL algorithms and proposed CNN with baseline model,the results prove that the proposed approach outperforms existing approaches. To assess the accuracy of the classifiers, performance evaluation measures, accuracy, precision, and recall are used. Experiments are performed on the latest credit cards dataset.

In credit card fraud detection,firstly we preprocess the data to handle the missing value .Data selection as well as data balancing is performed and then machine learning and deep learning algorithms are used to train the model.At last ,predict which model has provided more accuracy.

```
Algorithm: Credit Card Fraud Detection
Input: Credit Card details
Output: Predecting accurate result to detect credit card fraud using machine learning and deep learning
START
Step 1: Data Preprocessing
    1.1 The dataset is preprocessed to check the missing values and unwanted columns before executing the algorithms.
Step 2: Feature Selection and Data Balancing
    2.1 Perform Feature selection using PCA Technique
    2.2 Data balancing using SMOTE Technique
Step 3: Machine learning approaches
    3.1.Machine learning and Deep learning approaches like Random Forest,Decision Tree, KNN and Naive Bayes are used to find out the accuracy of each model.
Step 4: Comparison of model accuracy
    4.1 Prediction
    4.1.1 Input Data
    4.1.2  Loading the model which have high accuracy
    4.1.3 Prediction using the model
    4.1.4 View Result
STOP
```

A. DATA PREPROCESSING

Data preprocessing, is a component of data preparation, which describes any type of processing performed on raw data to prepare it for another data processing .In data preprocessing, we will handle the missing value by ignoring the tuple or by filling the missing values,remove the unwanted columns as well as split the dataset into training and testing set.

B. FEATURE SELECTION

A feature is an attribute which has an impact on a problem and it will choose the important features for the model is known as feature selection. Each machine learning process depends on feature, which mainly contains two processes; which are Feature Selection and Feature Extraction. Although feature selection and extraction processes may have the same objective,but both are completely different from each other.Feature selection is about selecting the subset of the original feature set, whereas feature extraction creates new features. Feature selection is a way of reducing the input variable for the model by using only relevant data in order to reduce overfitting in the model.

PCA

Feature selection is done by using PCA technique.PCA stands for Principal Component Analysis is a dimensional reduction method which is used in field of data science. It is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables. PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. It will be more called as a dimensionality reduction method, then as a feature selection method.

Firstly, take the input dataset and divide it into two subparts X and Y, where X is the training set, and Y is the validation set. Represent dataset into a structure. Such as represent the two-dimensional matrix of independent variable X. Here each row corresponds to the data items, and the column corresponds to the Features. The number of columns is the dimensions of the dataset. Standardizing the data and the calculate covariance,eigen values.Then sort the eigen vectors and remove unimportant features.

C. DATA BALANCING

Imbalanced data is a common problem in machine learning, which challenges to feature correlation, class separation and evaluation, which results in poor model performance. Balancing a dataset makes training a model easier because it helps as to prevent the model from becoming biassed towards one class. The model will no longer favour the majority class just because it contains more data.A balanced data set for a model would generate higher accuracy models, higher balanced accuracy and balanced detection rate.

SMOTE

Data balancing is done by using SMOTE. SMOTE stands for Synthetic Minority Oversampling Technique is one of the most commonly used oversampling methods to solve the imbalance problem. SMOTE is used to generate artificial/synthetic samples for the minority class. This technique works by randomly choosing a sample from a minority class and determining K-Nearest Neighbors for this sample, then the artificial sample is added between the picked sample and its neighbors. It aims to balance class distribution by randomly increasing minority class examples by replicating them.

```
SMOTE ALGORITHM STEPS
Input:Preprocessed data containing minority class
data points
Output: Balanced dataset with equally distributed
Step 1: Analyze the minority class vector.
Step 2: Identify the k nearest neighbour.
Step 3: Take one of those neighbours and identify
        the vector between the current data point
        and the selected neighbour and place a
        synthetic point.
Step 4: Repeat the above step for all minority data
        points and their k neighbours,till data is
        balanced.
```

D.MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

1. RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

2. K- NEAREST NEIGHBOURS (KNN)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

3. DECISION TREE

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

4. NAIVE BAYES

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is mainly used in text classification that includes a high-dimensional training dataset.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

RESULTS & DISCUSSION

This section explains about the experimental results obtained by applying the proposed mechanism.The implementation is carried out using Python with SMOTE module imblearn.Classification algorithms like Decision Tree,Random Forest,KNN and Naïve Bayes are tested using balanced as well as imbanced dataset.The performance measures of proposed credit card fraud detection is evaluated using the matrix such as Accuracy,Precision,Recall,F1-Score and AUC.

Accuracy: The degree of the closeness to the true value.It is used to measure the performance in the domain recovery and processing of the data,which can be represented as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision: It is a performance assessment which measures the ratio of correctly identified positives and the total number of identified positives,which can be represented as follows:

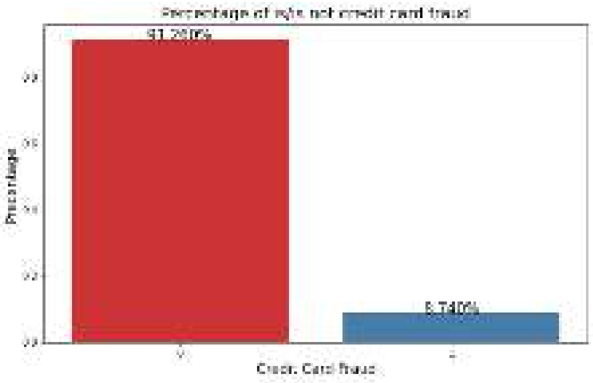
$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall : It is the ratio of connected instances retrieved over the total number of retrieved instances which can be seen as follows:

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score:It considers both the precision as well as the recall.The ratio of connected instances retrieved over the total number of retrieved instances which can be seen as follows:

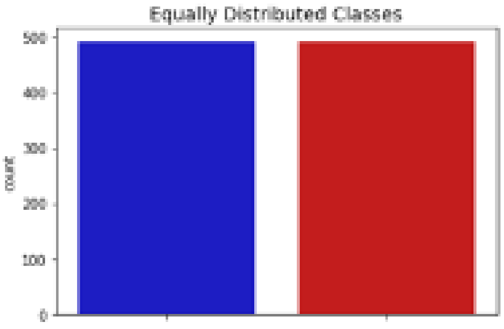
$$F = \frac{2XPrecision\ xRecall}{Precision+Recall}$$



0 - Not fraud 1- Fraud

Fig 2: Percentage of is or is not fraud

Fig 2,illustrates the original dataset containing 8.740% fraud and 91.260% non fraud.Here,the data is unbalanced.Applying classification algorithms to this unbalanced dataset will not provide an accurate classification ,so to improve the accuracy of the model tp preprocess the dataset.The main aim of data balancing is to convert imbalanced to balanced data.In the proposed mechanism we have use SMOTE algorithm for data balancing.



1-Fraud 0- Not Fraud

Fig 3: Equally distributed classes after data balancing

Fig 3, illustrates the balanced dataset obtained using Smote,ie, as per the figure balanced dataset contains 50:50 equally distributed classes.

Also, we have measure the performance of the algorithms such as Random Forest,KNN,Decision Tree,Naïve Bayes through the matrix Precision,Recall,F1-Score and AUC.

	Accuracy	Precision	Recall	F1	AUC
Decision Tree	0.97	0.97	0.97	0.97	0.97
Random Forest	0.97	0.97	0.97	0.97	0.97
KNN	0.96	0.95	0.97	0.95	0.97
Naïve Bayes	0.54	0.12	0.6725	0.2356	0.58

Table 1 Comparison of performance of classifier without balanced dataset

Table 1,represents the evaluation matrix and values obtained through the evaluation matrix for the different algorithms using balanced dataset.

	Accuracy	Precision	Recall	F1	AUC
Decision Tree	0.99	0.98	0.99	0.99	0.99
Random Forest	0.99	0.99	0.99	0.99	0.99
KNN	0.99	0.98	0.99	0.98	0.99
Naïve Bayes	0.63	0.15	0.73835	0.25999	0.68

Tabe 2,Comparison of performance of classifier with balanced dataset.

Table 2,represents the evaluation matrix and values obtained through the evaluation matrix for the different algorithms using balanced dataset.

On comparing the experimental results,it is proved that accuracy of performance and efficiency of algorithms using balanced dataset is better compared to imbalanced dataset.

CONCLUSION

The machine learning models such as decision tree, random forest, K-nearest neighbors, plain Bayes, support vector machine, and logistic regression after handling imbalanced samples, we found that by oversampling credit card fraud data, the model achieves the best performance We found that by oversampling the credit card fraud data, the model achieves accurate classification of fraudulent transactionsin credit card fraud data and improves the accuracy of classification of fraudulent transactions, and the random forest algorithm outperforms other machine learning algorithms among all the machine learning models compared. Using these methods for the detection of credit cards yields better performance than traditional algorithms. Accuracy of machine learning algorithms depends on quatity and quality of dataset used during traiing. Therefore,accuracy issues may occur. In future, instead of SMOTE we can implement SMOTE-NC or SMOTE-N.It contains numerical as well as categorical features .Thus it improves the performance.

REFERENCES

[1] Y. Abakarim, M. Lahby, and A. Attioui, ''An efficient real time model for credit card fraud detection based on deep learning,'' in Proc. 12th Int. Conf. Intell. Systems: Theories Appl., Oct. 2018, pp. 1–7, doi:10.1145/3289402.3289530.

[2] V. Arora, R. S. Leekha, K. Lee, and A. Kataria, ''Facilitating user authorization from imbalanced data logs of credit cards using artificial intelligence,'' Mobile Inf. Syst., vol. 2020, pp. 1–13, Oct. 2020, doi:10.1155/2020/8885269.

[3] J. Błaszczyński, A. T. de Almeida Filho, A. Matuszyk, M. Szelg ̧, andR. Słowiński, ''Auto loan fraud detection using dominance-based rough set approach versus machine learning methods,'' Expert Syst. Appl., vol. 163,Jan. 2021, Art. no. 113740, doi: 10.1016/j.eswa.2020.113740.

[4] B. Branco, P. Abreu, A. S. Gomes, M. S. C. Almeida, J. T. Ascensão, and P. Bizarro, ''Interleaved sequence RNNs for fraud detection,'' in Proc.26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2020,pp. 3101–3109, doi: 10.1145/3394486.3403361.

[5] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, ''Adversarial attacks for tabular data Application to fraud detection and imbalanced data,'' 2021, arXiv:2101.08030.

[6] V. N. Dornadula and S. Geetha, ''Credit card fraud detection using machine learning algorithms,'' Proc. Comput. Sci., vol. 165, pp. 631–641, Jan. 2019,doi: 10.1016/j.procs.2020.01.057.

[7] I. Benchaji, S. Douzi, and B. E. Ouahidi, ''Credit card fraud detection model based on LSTM recurrent neural networks,'' J. Adv. Inf. Technol.,vol. 12, no. 2, pp. 113–118, 2021, doi: 10.12720/jait.12.2.113-118.

[8] J. Forough and S. Momtazi, ''Ensemble of deep sequential models for credit card fraud detection,'' Appl. Soft Comput., vol. 99, Feb. 2021,Art. no. 106883, doi: 10.1016/j.asoc.2020.106883.

[9] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' 2015, arXiv:1512.03385.

[10] J. Kim, H.-J. Kim, and H. Kim, ''Fraud detection for job placement using hierarchical clusters-based deep neural networks,'' Int.J. Speech Technol., vol. 49, no. 8, pp. 2842–2861, Aug. 2019, doi: 10.1007/s10489-019-01419-2.