

A Survey on Machine Learning Handling Imbalanced Dataset in Credit Card Fraud

Pawan Panchole, Rajesh Dhakad

Department of Computer Engineering SGSITS Indore

Abstract- In the era of digital transaction people prefer to make online payments and purchases due to the convenience of time, transportation, etc. Credit card fraud has also increased significantly due to the growing trend of e-commerce. Fraudsters try to take advantage of card and internet payment information. Credit card and online payment information is often used by fraudsters for fraudulent purpose. Imbalanced dataset and high dimensionality of data are the key issues observed in credit card fraud detection. The use of various machine learning algorithm has been utilized for identifying anomalies in credit card transaction, focusing on the problem of imbalanced dataset and reduction of dimension which were carefully reviewed and studied. The study investigates the impact of imbalanced datasets on PCA-based fraud detection and provided detailed techniques such as Random Oversampling, SMOTE & Random Undersampling to handle imbalanced datasets and various classification as well as anomaly detection methods. Additionally, given the labelled nature of the dataset, various methods are reviewed like Logistic Regression, Random Forests, and Decision Trees. This study analyses and compares the performance of these methods before and after applying PCA and addressing data imbalance to assess their effectiveness in detecting credit card fraud.

Index Terms- Credit Card Fraud, Machine Learning, Principal Components Analysis, SMOTE.

I. INTRODUCTION

Credit card was first introduced in 1914 in the United States for bank, gas and retail holder. In 1950's, credit card took a new advancement and made travel and entertainment card. The growth of credit card increased up to 12.2% across worldwide. The use of credit card has gained popularity in between the years 1973 to 1977. With the introduction of bank cards, average use of credit card increased by 28% and then increased up to 38.4% in 1977 [1]. In terms of security, the introduction of magnetic stripes on cards in 1969 and European Europay MasterCard Visa chips (EMV) has been a milestone event which took place at the end of 21th century. In recent times, electronic payments have resulted in a sudden rise in the use of online payments [1]. Credit cards are one among these strategies that is very widely used. It is possible to make purchases using this credit line. Credit cards are a very important part of today's life as any financial loss of credit card or loss of sensitive credit card information is called CCF [2]. Financial fraud is a very significant issue that affects business and the economy, affecting the cost of living for customers [3]. Presently, almost 51% of all transactions include card payments. A credit card is issued to customers which allow merchants to make payments for products and services, from which the user can make use of money for payment [4], [5]. This is also true of credit card theft [6],

where reports increased by more than half (53%) in 2023 while the increase was even greater during that period last year.

CCF is a major problem in the digital economy, posing a threat to financial security and user trust [4], [7]. This study is to improve fraud detection techniques and handle imbalanced dataset which will strengthen financial security in the digital world [3]. Since the data is highly imbalanced and represents only a small portion of the total dataset, various machine learning algorithms have emerged as effective tools for detecting fraudulent activity in credit card transactions in response to fraud-related issues. Principal Component Analysis (PCA) is one of these techniques that is useful for feature extraction and dimensionality reduction, which helps to identify anomalies in large datasets [2]. PCA can enhance fraud detection but faces difficulties due to imbalanced datasets [3]. Data balancing techniques include undersampling, oversampling & hybrid [2] that improve the performance of fraud detection models and also investigates the relationship with PCA and data balancing approaches in enhancing fraud detection performance [3]. This study aims to investigate a relationship between imbalanced dataset handling and PCA performance in CCF detection, focusing on data balancing techniques as well as anomaly detection algorithms.

This study includes different methods that may be utilized for categorizing transactions as fraud or legitimate transactions [2]. Multiple machine learning techniques like Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN) [3] & PCA can be used for fraud detection [5]. Analysing anomaly detection methods using PCA-transformed features will shed light on how well various strategies work to reliably identify fraudulent transactions [7]. In order to ensure that there is sufficient fraud data for PCA to learn from, this study examines various approaches to data balancing. Our objective is to improve fraud detection so that everyone can transact safely online.

Background Study

Credit card fraud is considered as illegal use of credit card data by the fraudster. Fraud can be categorized into two types: Behavioural fraud & Application fraud [7]. Behavioural fraud is considered as multiple types of fraud which occurred due to inappropriate misuse of existing cards. Stolen or lost card fraud occurs when third person takes or finds a lost credit card and uses it [3]. Application fraud occurs when third person applies for a new card with fraudulent information or someone else's information. This can include submitting many applications with the same set of fake information, known as duplication fraud, or submitting multiple applications with identical fraudulent information, known as identity fraud [5], [8]. Fake card fraud is the creation of using stolen card information, and non-existent cardholder fraud is the use of card details without presenting the actual card, which is typically done using mail, message, phone or by internet transactions [3], [5]. Traditionally, fraudulent operations were carried out remotely using card data sent by phone, email, messages or the internet [9]. Fraudsters may now use these approaches to create fake cards using stolen card info [5].

One way to control fraud is through both Detection and Prevention. The goal of fraud prevention is to come to fraudulent activity by spotting and preventing fraudulent transactions before they are approved [9]. Whereas the primary aim of fraud detection is to recognize and distinguish between transactions that are fraudulent and those that are valid. As show in below figure-1 there are two main types of Credit Card Fraud:

- Card Not Present Fraud
- Card Present Fraud

Types of Credit Card Fraud

Card Not Present Fraud: Card Not Present (CNP) [3] fraud happens when scam performers utilize credit card details that have been stolen to carry out internet or phone purchases without actually having the card present [5], [10]. An actual card is needed in order to swipe it. Whenever using the virtual card [11], some information is required, such as the cardholder's name, password, security question, and CVV number and all [3], [5]. Or with internet banking. In most of

online or internet buying purchases do not require a physical card [6], [10]. Rather, provide the card number, date of expiry, and CVV number [3], [12]. This information can be stolen by scammers, who can then use it to conduct fraudulent online transactions [11]. Card Not Present (CNP) Fraud can further be classified into following types.

Phishing: Phishing is type of cyber crime where scammers collect personal information of the user such as credit card details, account details, and their passwords. Fraudsters make that by sending out fake mails, SMS, and web pages via the internet or without using the internet appear to be legitimate [9]. These kinds of crimes take advantage of trust and frequently generate urgency in order to get people to provide their information as soon as possible [3].

Identity Theft: Identity theft in CCF happens when someone obtains personal data such as DOB & Social Security Number, of any person to use the data for getting money [13]. This might involve making purchases, opening new accounts, or ruining the customer's credit history by using personal information. Detecting and resolving identity theft may be difficult and irritating for customers, underlining the need of protecting personal information from unauthorized use [6]. Another method which attempts to get credit cards under false pretences, which results in the financial losses of the card holder and ruining the target's credit history known as application fraud [4].

Card Present Fraud: The use of a physical credit card for a manual or hand transaction is known as credit card present fraud. Physical card fraud happens when a retail transaction is completed and the cardholder has to show the card to the retailer [5], [11]. There is a possibility that fraudsters will steal and use the customer's card without them knowing them [9], [12]. This type of fraud occurs when a physical credit card has been stolen, duplicated and faked at a physical place, mostly at store. At a point-of-sale terminal, the scammer usually inserts for swipes, or taps the card to inform the merchant about the scam [11]. Card Present (CP) Fraud can further be classified into following types.

Skimming: Skimmers are devices that collect credit card information through the magnetic strip present at the back of the card [9]. Scammers connect them to credit card reader devices at ATMs, retail shops, petrol pumps [5], [9]. The information is then either sold to other fraudsters or used to charge your card [3].

Card Cloning: Card cloning is copying information from a real credit card and transferring it to a fake card. In this method, fraudsters make illegal purchases using people's credit card information without their awareness [9]. Card cloning is often accomplished using skimmers, which get card

information during valid transactions, or by hacking methods that take card data from unsecure networks.

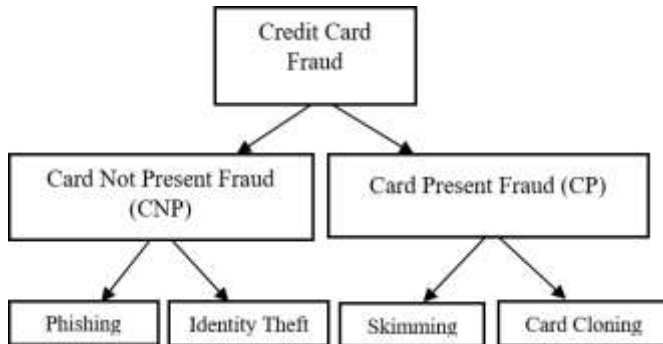


Fig. 1: Types of Credit Card Fraud

Challenges in CCF

There were following challenges and problems that were faced while dealing with credit card fraud detection that are mentioned below:

- Training fraud detection algorithm has become more difficult due to the imbalanced dataset, as there are more fraudulent transactions as compared to fraud transactions [3].
- Most fraudulent cases are difficult to find anomalies that might result in significant financial losses if misclassified [10].
- Credit card dataset contain more legitimate credit card transactions than fraudulent credit card transactions, which makes detecting CCF more difficult, leading to false results [8].
- In digital era, where fraud detection transaction takes place, balancing and applying PCA on the data is a very challenging issue which need to be solved [3].
- The imbalance poses an important challenge to machine learning models, they are biased toward the majority class, which causes a reduce in fraud detection rates. [11], [12]. Frequently leading to a high percentage of false negatives and incorrect fraud detection. Anomalies might not always be fraudulent they could just be unusual but ligament transaction leading to high false positive rate.

II. LITERATURE REVIEW

Dal Pozzolo et al. [14], identified that by use of Random Forests for improving fraud detection problem. It functions similarly to the technology that can analyse through large amounts of transaction data, even when there is not much fraud present.

Yeh and Lien [15], proposed a comparison of several ML algorithms for the identification of CCF. To test how well they could identify fraudulent transactions, they used methods that logistic regression, random forests & decision trees. To

address the problem of class imbalance, this study used dataset and presented a hybrid technique that included random oversampling & undersampling. The outcomes were demonstrated in terms of accuracy and precision.

Ibomoiye, Domor Mienye et al. [2], provided a detailed review of different machine learning techniques for CCFD they emphasized more on the use of supervised machine learning algorithms such as LR, RF, DT, SVM, Adaboost & ANN, as labelled data are more better in classifying fraudulent data.

Popat and Chaudhary [3], collectively proposed various supervised machine learning algorithms for CCF detection that include DT, Fuzzy Logic Systems, Neural Networks, DL, LR, Naive Bayes (NB), & SVM. The goal of these algorithms was to identify fraudulent transactions. To analyse how well these machine-learning algorithms performed in identifying CCF the authors focused more on the performance factors such as prediction, clustering, and outlier identification.

Vimala Devi. K. S. Kavitha et al. [10], identified various ML algorithms like RF, DT, & SVM for fraud detection along with fraudulent transactions. To improve the performance, variety of indicators were utilized and finally outcomes and performances of various algorithms were compared to each other.

Deepa and Akila [11], proposed various techniques such as the Anomaly Detection Algorithm, KNN, RF, K-Means, with DT. In a specific case, they proposed several methods and predicted the most effective algorithm for detecting fraudulent transactions. The system generated data using a range of rules and algorithms to forecast the likelihood of fraud.

Xiaohan Yu et al. [16], proposed a deep neural based network algorithm for CCF detection. The neural network technique provided various advantages that helped in CCF detection. Additionally, it also focused on the preparation techniques and how to utilize the loss to fix dataset bias problems.

Mengran Zhu, Ye Zhang, et al. [17], noted that in CCFD, imbalanced class distribution is a typical problem. SMOTE is a technique which generates data points of fraud cases in order to counter this imbalance. This is partially due to the rise of neural networks, which are particularly effective at identifying complex patterns in transaction data that include conventional methods. The utilization of ensemble approaches like bagging and boosting, as well as evolutionary algorithms for feature selection have also played a crucial role in further improving model accuracy. Additionally, more and more importance is being given to interpretable machine learning and Explainable AI (XAI).

Mahmoudi, N, Duman, E. et al. [18], analysed that the rise of credit card transactions nowadays has significantly increased the chances of fraud, leading to the development of advanced fraud detection techniques. An overview of this paper provides details of current methods & highlight gaps in the existing literature. A key issue identified in the literature was the equal treatment of false positives and false negatives, despite false negatives being more costly. This paper introduces the Fisher Discriminant Analysis as a novel method for credit card fraud detection, modifying the traditional approach and pointed out the profit maximization by weighting transactions based on their importance. The authors address the under-explored area of profit-based classification, proposing a model that improves classification accuracy while maximizing profits.

Siddhant Bagga et al. [19], analyzed the role of machine learning techniques such as NB, LR, KNN, and RF on credit card fraud detection. The dataset was balanced using ADASYN method and performance was evaluated using several parameters such as accuracy, recall, and F1 score.

Thudumu et al. [20], worked on anomaly detection for high-dimensionality data, but faced some challenges such as curse of dimensionality. To solve the problem, the authors came up with a new framework that helped in improving the accuracy and performance for detecting anomalies in large and complex datasets. Their work provides a foundation for future research in optimizing anomaly detection techniques for high-dimensional big data.

Prajal Save Pranali Tiwarekar et al. [21], forecasted that there is a higher chance of credit card theft as a result of the growth in e-commerce. To solve the problem there are various fraud detection techniques. Authors proposed several traditional supervised learning methods like ANN and DT but faced challenges of complexity and overfitting that restrict their usage across various applications and datasets.

The table 1 shows the summary of CCF detection and different study techniques, limitation, result and accuracy.

III. METHODS OF HANDLING IMBALANCED DATASET

An imbalanced dataset is a problem where one class (majority) contains more number of instance as compared to other class (minority). In CCF dataset, there are two classes i.e., fraudulent and non-fraudulent that are imbalanced by the nature [6], [17]. This is a typical problem of CCF detection, in which the number of genuine transactions are more than fraudulent transactions, they are very small as compared to all the transactions [3]. Due to imbalance dataset problem, machine learning models are difficult to train, models tend to

become biased in favour of the majority class [2], making them less effective at detecting the minority class [13]. Figure 2 shows the fraudulent transactions and genuine transactions in the total dataset.

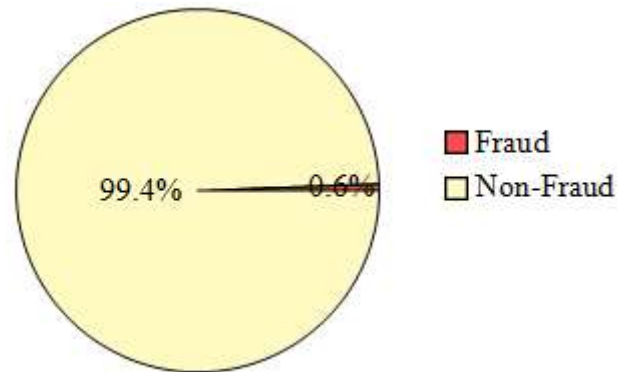


Fig. 2. Fraud and Non-Fraud Transactions

Fraud = 0.6% Non-Fraud = 99.4%

Ways of handling imbalanced dataset

- Oversampling
- Undersampling
- Hybrid sampling (Combination of oversampling and undersampling)

Oversampling: Oversampling techniques, including copying existing samples or producing new virtual samples that are implemented to enhance the quantity of samples in the minority class [6], [26]. Oversampling method helps in balancing the dataset and thus improve machine learning model's ability to identify patterns in the minority class [16], [27]. It takes more time to train the scientific model and it also increases overfitting problem. Due to overfitting issue, it becomes difficult to fit their data into the same curve, so the accuracy of the model decreases [6], [27]. The most often used oversampling methods are:

Random Oversampling
Synthetic Minority Oversampling Technique (SMOTE)
Adaptive Synthetic Sampling (ADASYN)

Random Oversampling: One of the basic oversampling methods, known as Random Oversampling, involves randomly selecting duplicate samples from the minority class in the training dataset, referred to as Random Oversampling [28], [29].

Synthetic Minority Oversampling Technique (SMOTE): The most popular successful oversampling method is called SMOTE [2], [27]. SMOTE methods work by identifying nearby instances in the feature space, then drawing a line between them and a new sample as a point along that line. It

generate synthetic samples for the minority class to balance the dataset [27], [29]. SMOTE creates synthetic data points by interpolating between neighbouring instances. This can help addressing class imbalance problem without overfitting [26]–[28].

Adaptive Synthetic Sampling (ADASYN): ADASYN is similar method as SMOTE [2] the density of instances in the minority class is inversely proportional to the number of fake samples produced by the procedure. In feature spaces with low minority example densities, it is designed to produce synthetic instances; in feature spaces with high minority example densities, it is meant to produce fewer or no examples. ADASYN is a data resampling technique that is used to balance imbalanced datasets by generating synthetic samples for the minority class [27]. ADASYN works by first identifying the minority class samples that are most difficult to learn, and then generating synthetic samples for these hard samples, therefore it is much effective than SMOTE [29]. This helps to improve the performance of machine learning models on imbalanced

Table 1: Literature Review

Title	Year	Techniques	Dataset	Limitation	Result	Accuracy
“A Survey on Credit Card Fraud Detection Using Machine Learning” [3]	2018	Neural Network and SMOTE	Kaggle: European card transactions	Logistic regression, Neural network, Support vector machine and Decision tree.	The paper discusses challenges in credit card fraud detection, including data imbalance, slow classification speeds. As compared to traditional methods, machine learning techniques for credit card fraud detection high accuracy and detection rates.	99%

“Credit Card Fraud Detection Using State-of-the-Art Machine Learning Algorithms” [13]	2022	Extreme Learning Method, Decision Tree, Random Forest, Support Vector	Neural networks, Bayesian models, decision trees	“Credit Card Fraud Detection in Card-Not-Present Transactions: Where to Invest?” [2]	2021	99%
Kaggle, credit card fraud detection Predictive Models		impact algorithm performance, they don't address the difficulties or biases associated with using these algorithms for fraud detection, and they don't provide specifics about the processing capacity required for deep learning.	Kaggle: “Paysim” Synthetic Financial Datasets For Fraud Detection	scalability, concentration on CNP transactions, dependence on undersampling, issues, and limited computing resource discussion.		
credit card fraud detection using deep learning and sophisticated machine learning, especially 20-layer CNNs, using a dataset of European transactions from October 2018				Accuracy and effectiveness of fraud detection systems in card-not-present transactions is enhanced by feature engineering and lower feature set cardinality.		93.4%

<p>“Credit Card Fraud Detection using Machine Learning and Data Science” [24]</p> <p>2019</p>	<p>Local Outlier Factor & Isolation Forest on PCA-transformed kaggle : credit card fraud detection.</p>	<p>Real-world data and more precise performance indicators are needed for a genuine implementation, the proposed credit card fraud detection system</p>	<p>False positives were identified & compared with actual values to calculate accuracy and precision</p>	<p>99.67%</p>
<p>“Credit Card Fraud Detection Using Machine Learning Algorithms” [22]</p> <p>2023</p>	<p>Local Outlier Factor and Isolation Forest, Principal components analysis ,</p>	<p>Kaggle: Credit card fraud detection</p>	<p>Real-world use is limited by the small data unit, and a more accurate system requires more bank data.</p>	<p>97%</p>
<p>“A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection” [23]</p> <p>2022</p>	<p>LSTM, Neural Network, AdaBoost, SMOTE(ENN)</p>	<p>European credit card benchmark dataset : 492 out of 2,84,807</p>	<p>Generalizability, Robustness, Scalability, and Interpretability</p>	<p>99.6%</p>
<p>“Ensemble of Deep Sequential models for Credit Card Fraud Detection ” [25]</p> <p>2020</p>	<p>Ensemble of LSTM, GRU, and FFNN with a learning-based voting mechanism</p>	<p>European cards dataset, Brazilian dataset</p>	<p>Challenges like data skewness and need for efficient computational resources are discussed; specific limitations not detailed</p>	<p>95%</p>

Datasets.

Undersampling: In undersampling, it reduces the number of samples or instances of the majority class to balance with the minority class [16], [27]. Randomly removing instances from the majority class is one approach, but more sophisticated methods can be used to preserve important information [6], [28]. Some of the more widely used and implemented undersampling methods include:

- Random Undersampling
- Near Miss Undersampling
- Tomek Links Undersampling

Random Undersampling: It is a most simple undersampling method. This technique selects samples at random from the majority class and then removes the instances from the training dataset [28], [29].

Near Miss Undersampling: Near Miss helps in balancing an imbalance dataset by randomly selecting samples from the majority class. This method consists of three types: The majority class samples with the least average distance to the three closest minority class samples are chosen using NearMiss-1 [30]. The samples from the majority class that are closest to the three farthest samples from the minority class on average are chosen by NearMiss-2 [27]. NearMiss-3 selects a specified number of majority class samples that are closest to each example in the minority class. [30].

Tomek Links Undersampling: The samples in the training dataset that are closest neighbors but fall into distinct classes are referred to as Tomek Links [27]. In addition to being

removed from the majority class, these samples are incorrectly categorized at the class border.

Hybrid Sampling (Combination of oversampling and undersampling): This technique works by combining both oversampling and undersampling method, hence increases the model's ability to identify and classify the occurrences of minority classes [16], [27]. In combination of these techniques, if oversampling is done on its own, the model may over fit and learn from the noise in the data [28]. These techniques helps in the betterment of the data and thus reduce the threat of overfitting by reducing the feature space according to the defined parameters [26].

IV. CREDIT CARD FRAUD DETECTION TECHNIQUES

The credit card dataset utilized in the study is described in this part, along with a thorough discussion of the several algorithms and techniques that went into creating the recommended credit card fraud detection approach [2]. Credit Card Transactions Fraud Detection Dataset, is freely available on kaggle, comprises of 10,48,576 records and total 23 attributes (both categorical and numerical) with some missing values. It is intended for thorough analysis, especially for identifying fraudulent transactions, and includes a variety of attributes, including transaction details and customer information. The table 4 shows dataset attributes and its entries. Few machine learning algorithms that may be applied to the identification of CCF [13].

- Logistic Regression
- Decision Tree
- Random Forest
- Principal Components Analysis

Logistic Regression: Logistic Regression can be used for classification & regression problems but it is mostly used for binary classification problems [3], [33]. LR uses a linear model which is used to predict categorical variables using dependent variables [27]. The algorithm computes the probability value ranging between 0 and 1 [8], [34].

The variable used for binary classification is described below:

$$y_i = \begin{cases} 0 \\ 1 \end{cases}$$

LR creates this cost function, which is referred as logistic or sigmoid function, which is mentioned below [33]:

$$h(x) = \frac{1}{1 + e^{-x}}$$

Prediction Formula: Where:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

- Y=1 means a transaction is fraudulent.
- X1, X2, . . . are the features of the transaction.
- $\beta_0, \beta_1, . . .$ are the model coefficients learned during train- ing.

The regression standardization parameters are W0 and W1, where, W1 stands for the coefficient vector, while W0 is for the intercept [28], [33].

Working of LR on CCF is discussed below:

Class Weighting: LR allows assigning different weights to the classes, increasing the penalty for misclassifying minority class fraudulent transactions within the majority class [33]. LR makes the model pay more attention to fraudulent transactions even though they are sparse.

Threshold Tuning: Instead of using the default threshold value as 0.5, value of the threshold is reduced to increase the sensitivity of the fraudulent cases [3], [29]. A lower threshold value of the model increases the chances to classify borderline transactions as fraud.

Resampling Techniques

- **Oversampling (like SMOTE, ADASYN):** Oversampling techniques are classified into two types i.e., SMOTE and ADASYN. These techniques create synthetic fraudulent transactions that means new identity is created by combining real and fake information to balance the dataset [27], [28].
- **Undersampling:** This technique helps to reduce the total amount of non-fraudulent transactions to balance the dataset [27].

Evaluation Metrics: LR is usually evaluated using following metrics like precision, recall, and the F1-score, rather than accuracy, when dealing with imbalanced dataset [11]. Other metrics like ROC-AUC are also useful, as they helps to evaluate how well the model separates the fraudulent and non-fraudulent transactions.

Decision Tree: It is most famous algorithm applied to machine learning classification problems [4]. It has a tree like structure and begins with a root node, each leaf node representing a feature from the dataset, each leaf node representing the result, and branches indicating the decision

rules [5], [34]. A decision tree then splits the dataset into smaller trees depending on the solution [8]. Fig 3 shows the structure of the DT [5]. It searches the tree from top and then goes down which is a top=down approach.

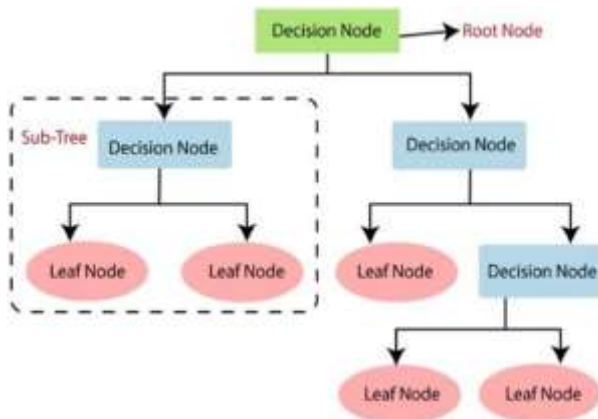


Fig. 3. Structure of Decision Tree

Working of Decision Tree on CCF is discussed below:

Class Weighting: In decision trees, we assign different weights to classes so that the models performs better with fraudulent transactions, even though they belong to the minority class.

Cost-sensitive Learning: This technique assigns more weightage for misclassifying the fraudulent transactions i.e., minority class. The decision tree minimize this data for misclassification during the spitting of the data. This allows the tree to study more about the minority class and make better decisions for fraudulent transactions.

Resampling Techniques

- **Oversampling (e.g., SMOTE, ADASYN):** These techniques create synthetic fraudulent transactions that means new identity is created by combining real and fake information to balance the dataset so that the decision tree has more fraudulent samples to learn from the dataset [27].
- **Undersampling:** This technique reduces the number of non-fraudulent transactions to balance the dataset but might lead to loss of important information from the majority class [6], [27].

Pruning: Decision tree leads to overfitting problem in imbalanced dataset. To solve this problem, Pruning technique has been introduced that reduces the overfitting [35]. Pruning helps in reducing the depth of the tree and removes branches from the tree that doesn't lead to clas- sification which can improve performance in detecting fraud cases [35].

Random Forest: A subset of the data is found to have decision trees using the random forest classifier. This classifier enhances precision and gets rid of the overfitting problem by taking the prediction from each tree and forecasting using a huge amount of data [5]. The majority votes from the forest trees forecasts determine the final outcome [8]. Based on the majority votes of forecasts, The RF uses the predictions from each tree to predict the final result. Excessive numbers of forest trees are used to reduce overfitting and increase accuracy [5], [36]. RF is capable of handling huge amount of data sets with high dimensionality and handling both regression and classification and avoids overfitting problems [5], [6], [34]. Fig 4 shows the structure of the random forest [5].

Table 2: Machine Learning Algorithms and Their Accuracy

Sr. No.	Dataset	Title of the Paper	Algorithm	Accuracy	References
1.	Credit Card Fraud Detection	“Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms”	Extreme Learning Method, DT, RF, SVM, LR, XG Boost & Convolutional Neural Networks (CNNs)	96%	[13]
2.	The bankcard enrolment records	“Credit Card Fraud Detection Based on Machine Learning”	LR-based, RF-based & GBDT-based	75%, 73% & 74%	[31]
3.	Commercial bank in China	“Performance analysis of feature selection methods in software defect prediction: A search method approach”	SVM, RF	97.10%, 96.90%	[32]

Balanced Random Forest: A variation of Random Forest that builds each tree on a balanced subset of data (either by oversampling fraud cases or undersampling non-fraud cases). This helps ensure each tree has enough fraud examples to learn from [36].

Threshold Tuning: Since Random Forest can output class probabilities, the decision threshold can be tuned (lowering it below 0.5) to increase the sensitivity for detecting fraudulent transactions [3], [29].

Evaluation Metrics: Random Forest for imbalanced datasets is typically evaluated using metrics like Precision, Recall, F1-score & ROC-AUC, which are more informative than accuracy in such cases.

Principal Components Analysis: PCA techniques helps in reducing the number of features by transforming the original dataset into reduced set of principal components [29], that helps in capture the most important information from the dataset [20], [37]. By reducing the number of dimensions or features, PCA make dataset more simpler the dataset, which allows ML techniques like LR, DT & RF to perform faster and more efficiently without losing any important information [30], [34]. This techniques also helps in removing noise or irrelevant information from the data [38]. In fraud detection, some features may not lead to fraud detection but can have some type of complexity [20], [37]. PCA can filter out irrelevant features, improving model performance. Working of PCA on CCF is discussed below:

Data Reduction: PCA converts the data into a new types of dataset that includes principal components that helps in reducing the dimensionality of the original dataset [38]. Variability of data is mostly captured by these principal components, which are derived from the original variables [38]. For example in credit card fraud dataset, PCA can reduce the features from 23 to only 2 principal components [37]. This type of reduction helps distinguish between fraudulent and non-fraudulent transactions while preserving important information.

Identifying Patterns: Using PCA reducing the number of variables in the data, PCA can easily identify underlying patterns in the data. This dimensionality reduction highlights the most significant features, making the detection of fraud data more effective and efficient. This process simplifies complex problems between features, providing a clear view of the data structure.

Finding Principal Components: PCA converts the dataset into new variables that represent the direction of data. The first principal component represents the largest part of the variation, combining features like transaction amount and time. Each principal components represent a smaller amount

of variation, to ensure that only important information is used for analysis [37].

Improving Model Performance: PCA can act as a pre-processing step to improve machine learning models performance [37]. It helps in reducing the number of input variables, remove correlation and simplifies the data. This can prevent over fitting, where a model becomes too complex and can't perform well on new data. In addition, dimensionality reduction helps in decreasing computational complexity, leading to faster and more efficient model training.

V. ANOMALY DETECTION AND CLASSIFICATION TECHNIQUES

Anomaly detection focuses on identifying unusual patterns that are significantly distinct from the majority of the data, referred to as anomalies or outliers [7], [20]. Anomalies, also known as outliers of the dataset, which is apart from majority data classes [12], [39]. It requires identifying unique outliers or anomalies that are different from the data that identify patterns as they could be an indication of fraud, errors, or other unusual activities [22]. Anomalies can be displayed as unique data points or small groups that are different from the real observation cluster [22], [39]. These outliers may point to unique actions or patterns in the dataset. Some anomaly detection techniques are discussed below.

1. Local Outlier Factor

It is an unsupervised outlier detection algorithm that calculates the local density deviation of a given data point with respect to its neighbours [22], [39]. It is a density based detection model that computes similarity and dissimilarity between factors and variants which can be used for performing outlier detection [39], [40]. LOF factors to apply structure to find outliers, in nature [6].

LOF algorithm firstly calculate the difference in density between an object and its neighborhood, then influence of that data point called as local outlier factor (LOF) [22]. Basically, data points with high LOF values are considered outliers, whereas data points with low LOF values are termed as normal points. The LOF scores of almost all instances are probably normal in their attribute.

2. Isolation Forest

Isolation forest is an outlier detection method [40]. In this method, a random feature is selected and further a split value is selected between minimum and maximum values. After a few iterations of this procedure, the average path length is also determined. [6], [22], [39], & the normalized value is taken so that anomalies can be detected instead of normal observations.

V. CONCLUSION

This study has shown that the accuracy with which credit card theft is detected may be improved when using PCA with data balancing strategies. A complete analysis of the two main strategies balancing the dataset first and then using PCA showed that the order in which these methods are used has a significant effect on the model's performance. With the help of Synthetic Minority Over-sampling Technique (SMOTE) dataset is balanced. On the other hand, PCA shown better results than any other supervised anomaly detection models that includes SVM, LR & RF. This study shows how important pre-processing steps are to machine learning processes and offers an easy way to for future research on more advanced data balancing techniques and machine learning algorithms. To provide reliable and flexible fraud detection systems in a dynamic digital environment, future research should focus on improving these approaches and increasing their application to various places within digital transaction security.

REFERENCES

1. G. Garcia, "Credit cards: An interdisciplinary survey," *Journal of Consumer Research*, vol. 6, no. 4, pp. 327–337, 1980.
2. I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," *IEEE Access*, vol. 11, pp. 30 628–30 638, 2023.
3. R. R. Popat and J. Chaudhary, "A survey on credit card fraud detection using machine learning," in 2018 2nd international conference on trends in electronics and informatics (ICOEI). IEEE, 2018, pp. 1120–1125.
4. G. Suresh and R. J. Raj, "A study on credit card fraud detection using data mining techniques," *International Journal of Data Mining Techniques and Applications*, vol. 7, no. 1, pp. 21–24, 2018.
5. A.M. Fayyomi, D. Eleyan, and A. Eleyan, "A survey paper on credit card fraud detection techniques," *Int. J. Sci. Technol. Res.*, vol. 10, no. 09, 2021.
6. S. Warghade, S. Desai, and V. Patil, "Credit card fraud detection from imbalanced dataset using machine learning algorithm," *International Journal of Computer Trends and Technology*, vol. 68, no. 3, pp. 22–28, 2020.
7. W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert systems With applications*, vol. 193, p. 116429, 2022.
8. W. Lovo, "Detecting credit card fraud: An analysis of fraud detection techniques," 2020.
9. K. J. Barker, J. D'amato, and P. Sheridan, "Credit card fraud: awareness and prevention," *Journal of financial crime*, vol. 15, no. 4, pp. 398–410, 2008.
10. J. V. Devi and K. Kavitha, "Fraud detection in credit card trans- actions by using classification algorithms," in 2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC). IEEE, 2017, pp. 125–131.
11. M. Deepa and D. Akila, "Survey paper for credit card fraud detection using data mining techniques," *Int. J. Innov. Res. Eng. Appl. Sci.(IJIRASE)*, vol. 3, no. 6, pp. 483–489, 2019.
12. I. Mekterovic', M. Karan, D. Pintar, and L. Brkic', "Credit card fraud detection in card-not-present transactions: Where to invest?" *Applied Sciences*, vol. 11, no. 15, p. 6766, 2021.
13. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of- the-art machine learning and deep learning algorithms," *IEEE Access*, vol. 10, pp. 39 700–39 715, 2022.
14. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bon- tempi, "Credit card fraud detection: a realistic modeling and a novel learning strategy," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3784–3797, 2017.
15. I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert systems with applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
16. X. Yu, X. Li, Y. Dong, and R. Zheng, "A deep neural network algorithm for detecting credit card fraud," in 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2020, pp. 181–183.
17. M. Zhu, Y. Zhang, Y. Gong, C. Xu, and Y. Xiang, "Enhancing credit card fraud detection a neural network and smote integrated approach," *arXiv preprint arXiv:2405.00026*, 2024.
18. N. Mahmoudi and E. Duman, "Detecting credit card fraud by modified fisher discriminant analysis," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2510–2516, 2015.
19. S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit card fraud detection using pipeling and ensemble learning," *Procedia Computer Science*, vol. 173, pp. 104–112, 2020.
20. S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, pp. 1–30, 2020.
21. P. Save, P. Tiwarekar, K. N. Jain, and N. Mahyavanshi, "A novel idea for credit card fraud detection using decision tree," *International Journal of Computer Applications*, vol. 161, no. 13, 2017.
22. V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Procedia computer science*, vol. 165, pp. 631–641, 2019.

23. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," *IEEE Access*, vol. 10, pp. 16 400–16 407, 2022.
24. S. Maniraj, A. Saini, S. Ahmed, and S. Sarkar, "Credit card fraud detection using machine learning and data science," *International Journal of Engineering Research*, vol. 8, no. 9, pp. 110–115, 2019.
25. J. Forough and S. Momtazi, "Ensemble of deep sequential models for credit card fraud detection," *Applied Soft Computing*, vol. 99, p. 106883, 2021.
26. G. Lemaire, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of machine learning research*, vol. 18, no. 17, pp. 1–5, 2017.
27. Z. Li, M. Huang, G. Liu, and C. Jiang, "A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection," *Expert Systems with Applications*, vol. 175, p. 114750, 2021.
28. H. Shamsudin, U. K. Yusof, A. Jayalakshmi, and M. N. A. Khalid, "Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset," in *2020 IEEE 16th international conference on control & automation (ICCA)*. IEEE, 2020, pp. 803–808.
29. J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *Journal of Big Data*, vol. 10, no. 1, p. 58, 2023.
30. N. M. Mqadi, N. Naicker, and T. Adeliyi, "Solving misclassification of the credit card imbalance problem using near miss," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 7194728, 2021.
31. Y. Fang, Y. Zhang, and C. Huang, "Credit card fraud detection based on machine learning." *Computers, Materials & Continua*, vol. 61, no. 1, 2019.
32. A.O. Balogun, S. Basri, S. J. Abdulkadir, and A. S. Hashim, "Performance analysis of feature selection methods in software defect prediction: a search method approach," *applied sciences*, vol. 9, no. 13, p. 2764, 2019.
33. J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 international conference on computing networking and informatics (ICCNi)*. IEEE, 2017, pp. 1–9.
34. F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, naïve bayes and knn machine learning algorithms for credit card fraud detection," *International Journal of Information Technology*, vol. 13, no. 4, pp. 1503–1511, 2021.
35. Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2011, pp. 1–6.
36. S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*. IEEE, 2018, pp. 1–6.
37. A. Lasisi and N. Attoh-Okine, "Principal components analysis and track quality index: A machine learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 91, pp. 230–248, 2018.
38. N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Computer Science*, vol. 163, pp. 292–299, 2019.
39. A. Toshniwal, K. Mahesh, and R. Jayashree, "Overview of anomaly detection techniques in machine learning," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2020, pp. 808–815.
40. P. Kumar and F. Iqbal, "Credit card fraud identification using machine learning approaches," in *2019 1st International conference on innovations in information and communication technology (ICIICT)*. IEEE, 2019, pp. 1–4.