A Comprehensive Study of SMOTE-Enhanced Machine Learning Models on Learning Models on Credit Card Fraud Dataset

MSc Research Project

Fintech

Jeevan Kumar A

X21225940

School of Computing

National College of Ireland

Supervisor: Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	……Jeevan Kumar A……
Student ID:	……X21225940……
Programme:	…MSc Fintech…… **Year:**2022/2023
Module:	……Msc Research Project……
Supervisor:	……Brain Byrne……
Submission Due Date:	……14/08/2023……
Project Title:	**A Comprehensive Study of SMOTE-Enhanced Machine Learning Models on Learning Models on Credit Card Fraud Dataset**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary.

Signature: ……Jeevan Kumar A……

Date: ……14/08/2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comprehensive Study of SMOTE-Enhanced Machine Learning Models on Learning Models on Credit Card Fraud Dataset

X21225940

Abstract

Credit card fraud is a pervasive issue that puts both people and financial institutions at significant financial risk. Due to the increase in the number of online transactions, Effective and trustworthy fraud detection technologies are urgently needed. This study uses utilizing the Synthetic Minority Over-sampling Technique (SMOTE) to examine the productivity before and after balancing.

The results of the study show the differing degrees of efficiency reported among the various approaches. Notably, after class balancing, certain models demonstrated improved performance. This work serves as a compelling reminder of the importance of selecting proper machine learning techniques and preprocessing processes with care. These measures are critical in building robust fraud detection systems capable of withstanding the ever-changing landscape of fraudulent operations.

Keywords: Credit card fraud detection, SMOTE (Synthetic Minority Over-sampling Technique), Machine learning algorithms, Class imbalance

1 Introduction

The increase in social media utilization and the reliance on credit cards for online transactions lead to fraudulent activities within the domain of contemporary technology, culminating in substantial financial losses. Most businesses accept credit cards as a form of payment. Many approaches like big data technology and machine learning technologies have been used till now to determine fraud transactions. Building a suitable model to capture the fraudulent transaction is a challenge due to the imbalance class problem, where the quantity of recognized transactions is upward to the number of fraudulent transactions. The following discussion shows some common types of fraud cases associated with cardholders (Al-amri *et. al.*, 2021). Stolen cards: Private information, name of the person, SSN number, and credit card number, can be obtained by scammers and used to make purchases. Card skimming, hacking, and phishing schemes are just a few examples of how this might occur. This means fraudsters win customers' trust and then take their emails and personal information or small gadgets are placed on card readers by scammers to steal card data. Card testing: This occurs when fraudsters utilize automated software to check the accuracy of stolen card numbers. They do generate a lot of transactions in this situation and check which credit card numbers are legitimate and which are not. Alternative refunds method: In this case, fraudsters pay more than they should for a product or service, and later they do demand saying it happened accidentally entering the wrong number.

There are many research and experiments going on to deal with the issue of fraudulent card payment transactions. Some of the methods include AI-based techniques including deep learning, data mining, and big data cloud technologies. Fraud detection not only saves millions of rupees in payments but also helps to build better customer relationships (Alerfai *et. al.*, 2022). In this study,

we assess the outcomes based on the model accuracy and use machine learning prediction algorithms to analyze fraud and legitimate situations.

1.1 Motivation and Project Background

The incredible ease afforded by online transactions has generated unprecedented growth in credit card usage in this quickly expanding digital era. While this extraordinary expansion has undeniably aided frictionless trade, it has also created the path for a slew of illicit activities, the most prominent of which is credit card theft. This pernicious type of fraud not only causes enormous losses for clients and financial institutions, but it also corrodes the entire basis of trust on which our financial systems rely. As a result, the capacity to recognize and prevent such fraudulent acts quickly has become critical. The ever-changing nature of fraudsters' strategies adds to the urgency of the situation, leaving standard detection measures more useless.

Therefore, the creation of creative, flexible, and highly effective fraud detection technologies is urgently needed.

There are several forms that credit card theft can happen, including card-present fraud, card-not-present fraud, account takeover, and fake applications. Financial organizations and banks have traditionally depended on rule-based systems as their principal line of defense against these fraudulent operations. These systems would alert transactions based on predefined parameters, such as unusually large transaction amounts or transactions originating in high-risk geographical areas. However, these traditional approaches frequently produced an alarming number of false positives, causing significant annoyance to actual clients.

Fortunately, The development of machine learning and artificial intelligence led to a paradigm shift in fraud recognition techniques. Machine learning demonstrates are exceptional in their capacity to learn from prior transaction data, identify underlying trends, and forecast the validity of future transactions. These models can drastically minimize false positives and improve the Total efficacy of identifying fraud procedures of responding to emerging fraud strategies.

Nonetheless, due to the inherent class imbalance in the data, training these models is a considerable issue. Because real transactions substantially outweigh fraudulent ones, algorithms that forecast genuine transactions have a bias. As a result, the essence of this research is resolving this imbalance properly and critically assessing the efficacy of machine learning models functioning inside such environments.

1.2 Research Question

"How well do several machine learning models, such as Logistic Regression, GaussianNB, SVM, Random Forest, XGBoost Classifier, and AdaBoost Classifier, perform in class balancing both before and after credit card fraud detection using the Synthetic Minority Over-sampling Technique (SMOTE)?"

1.3 Research Objectives

Objective.1: To evaluate the performance of six distinct machine learning algorithms for detecting credit card fraud, we considered the following models: Logistic Regression, GaussianNB, Support Vector Machine (SVM), Random Forest, XGBoost Classifier, and AdaBoost Classifier.

Objective.2: To assess the impact of class imbalances on the predictive accuracy of the selected models and determine the necessity of fraud detection using class balance approaches.

Objective.3: Putting the Synthetic Minority Over-sampling Technique into action for class balancing and analyzing its impact on the machine learning models' performance metrics.

Objective.4: To evaluate each model's precision, recall, F1-score, and accuracy before and after class balancing with SMOTE.

Objective.5: To determine, after class balancing, which machine learning model has the highest F1 score and accuracy for detecting credit card fraud.

2 Related Work

As more individuals move into the digital world, cybersecurity is becoming more and more important in daily life. The major problem when discussing digital life security is spotting unusual behavior in regular life transactions. Many consumers choose credit cards while transacting or making online purchases of any kind. The “credit limit” on “credit cards” can occasionally enable purchases even when the necessary funds are not available. On the other side, cybercriminals abuse these features (Mathew, 2023). There is a mechanism that can stop a transaction if it notices abnormalities to address this problem. Here, a system that can monitor the patterns of all transactions is required, and if any patterns are abnormal, the transaction should be stopped. Financial fraud incidences, particularly credit card fraud, have risen since the new "e-payment" and "e-commerce" advances. Therefore, developing tools that can detect (credit card fraud) is essential. The characteristics of the features must be carefully chosen when using "machine learning" to identify "credit card fraud."

2.1 Concept

Credit card fraud is an ongoing and significant challenge in the finance sector, affecting both customers and financial institutions alike. Scammers have created new, highly sophisticated approaches to exploit vulnerabilities in "credit card systems" because of the growing reliance on digital transactions and electronic payment systems. Hence, the demand for proficient "fraud detection systems" is on the rise. to protect consumers' financial resources and uphold confidence in electronic payment systems (Taha and Malebary, 2020). The goal of this literature study is to look into modern techniques to make use of "machine learning techniques" to identify "credit card fraud”. The mainstays of “traditional fraud detection techniques” were rule-based algorithms and manual inquiry, both of which have limitations in terms of their ability to identify new and difficult fraud forms. Over the past ten years, the adoption of "machine learning techniques" has considerably improved the accuracy of fraud detection. (Al-amri *et al*., 2021). Decision trees, logistic regression, and naive Bayes classifiers were early examples of "machine learning models.". These techniques had great potential, but they were constrained by their incapacity to handle huge datasets and non-linear connections. "Supervised learning algorithms" have grown in popularity in the identification of "credit card fraud" due to their capability to analyze labeled data Random Forests, Neural Networks, and Support Vector Machines (SVMs) have all been extensively studied in this area. These algorithms may successfully detect fraudulent transactions by finding patterns and relationships in earlier data. They may struggle to manage datasets with imbalances, where the fraction of fraud incidents is significantly smaller than that of real

transactions, and they commonly encounter overfitting. An alternate strategy is to use unsupervised learning techniques, which do not need labeled training data. Similar transactions are grouped together using clustering techniques like K-means and DBSCAN, which aid in finding abnormalities and possible fraud situations (Ghosal *et al*., 2020). These techniques, however, could produce false positives and lack the accuracy provided by supervised techniques. Researchers have investigated hybrid models to overcome the drawbacks of both supervised and unsupervised procedures. Combining the advantages of the two techniques enhances the accuracy of fraud discovery. By leveraging the concept of anomaly finding, the most popular hybrid technique known as the Isolation Forest algorithm successfully distinguishes between fraud cases.

To capture temporal correlations and identify intricate fraud patterns. Sequential data, including transaction sequences, have been subject to "Long short-term memory (LSTM) networks" and "recurrent neural networks (RNNs)". Regardless of the strategy chosen, effective feature engineering and data preparation are required for fraud detection models. To boost the model's efficiency and speed up computation, feature selection, dimensionality reduction strategies, and class imbalance must be addressed (Zebari *et al*., 2020,). In recent years, significant progress has been made in the detection of "credit card fraud" using "machine learning-based" methods. The accuracy and efficiency of fraud detection systems have significantly improved thanks to research. Moving from conventional supervised models to cutting-edge “deep learning algorithms”. The management of enormous volumes of transactional data, coping with changing fraud tendencies, and lowering false positives are still problems. To successfully fight the constantly changing dangers posed by credit card theft, future research should concentrate on creating more reliable and flexible models.

2.2 Literature Review

According to (Alarfaj *et al*., 2022,) consumers can utilize credit cards to make purchases online since they offer a timely and practical method. The ability for credit card fraud has advanced besides using the assist by credit cards. The theft of a "credit card" causes significant financial failure for the two credit card holders as well as most economic institutions. This study's immediate goal is to identify such "frauds" that include the convenience of general data, large class inequalities in the data, differences in the type of “fraud” an increased false rate warning. Multiple "Machine learning-based methods" for credit card recognition are presented in the relevant literature. Some of these methods include the "Extreme Learning Method," "Decision Tree," "Random Forest," "Support Vector Machine," "Logistic Regression," and "XG Boost." However, due to the low accuracy, modern "deep learning algorithms" must be utilized to minimize fraud losses. The most recent development of "deep learning" algorithms has received the most attention. To provide sufficient results, a comparison of the "deep learning" and "machine learning" algorithms was carried out. A thorough investigation has been completed using the "European card" standard data for "fraud detection". First, the dataset was subjected to a "machine learning technique" that slightly increased the level of identifying fraudulent activity.

Credit card usage has substantially expanded while the globe progressively transitions to cashless commerce. And digitization. Additionally, there have been more fraud-related operations, which cost financial institutions a great deal of money. As a result, must study and differentiate between planning and legitimate transactions. In this paper, (Tiwari *et al*., 2021,) provide a full overview of the major fraud detection techniques. “Random Forests”, “Logistic Regression”, “Genetic Algorithms, Neural Networks”, and “Bayesian Belief Networks” are a few of these techniques.

There is a detailed analysis of multiple methods suggested. The benefits and drawbacks of the topic described in the relevant journals are contained in the student's conclusion to the paper.

According to (Chen and Lai, 2021,) Multiple corporations, the exponential growth of Internet use has led to the establishment of online advantages, particularly by those in the financial and commercial sectors. The growth of financial fraud on a global scale has resulted in enormous financial losses. It is now possible for "Advanced bank fraud detection systems" to actively identify threats like unauthorized trades and sophisticated attacks. Current ages have seen greater using "machine learning techniques" and "data mining" to handle these issues. These tactics still ought to be enhanced in a variety of areas, such as huge data analytics, computation speed, and the detection of previously unknown assault patterns. A “Deep neural network” convolution (DCNN) solution for “money laundering detection” is made in this paper utilizing an "algorithm for deep learning." When there is a large amount of data to be identified, doing so can improve detection accuracy. Using "real-time" credit card fraud data, the suggested model's implementation is evaluated in comparison to other deep learning models, auto-encoder models, and pre-existing "machine learning models". The testing outcomes show which suggested model detects a rate of "99%" in the period of 45 seconds.

In a research project by (Zhang *et al*., 2021,) Owed to fraud influencing credit card transactions, card issuers suffer annual losses of billions of dollars. Instances of financial fraud, specifically credit card fraud, have seen a rise due to recent advancements in "e-payment" and "e-commerce" technologies. The "fraud detection method" that is built using a "deep learning architecture" and a sophisticated quality engineering practice based on "homogeneity-oriented behavior analysis (HOBA)" is the primary contribution to the researcher's work. The researchers launched a comparison study established on a perfect dataset from a few of the largest saleable financial institutions in “China” to evaluate the significance of the suggested framework. The practical results indicate that the researchers presented methodology as a good and beneficial device that determines credit card fraud. In this research, the authors suggested a technique, with a modest false positive rate, that can detect much more fraudulent dealings than the standard methods. The study's conclusions have organizational ramifications for credit card investors, who may employ the suggested approach to promptly identify "fraudulent transactions", safeguard their client's interests, and reduce fraud losses and regulatory expenses.

A significant issue in the world of electronic payments is credit card fraud. To determine forged trades that have been made illegally in the name of legitimate cardholders, (Lucas, and Jurgovsky., 2020,) In this study, the researchers used the dataset and its components, the measured selection, and certain trading methods with such erratic datasets to carry out a normal credit card recognition task. Every issue in detecting credit card fraud starts with these inquiries. Then the researchers concentrate on dataset shift, also known as concept drift, which is the gradual evolution of the underlying dispersal that produced the dataset: For example, cardholders' shopping habits may vary during the year, and fraudsters may change their methods.

Any action with the intent to harm another party financially is categorized as fraud. The prevalence of digital money fraud has increased along with its expansion across the globe. These fraudulent operations cost banking and credit card companies millions of pounds in lost revenue each year and harm the careers of countless workers. In this paper, (Azhan and Meraj., 2020,) have discussed There are numerous credit card fraud methods active nowadays. Researchers have mentioned how “Machine learning” & It is possible to employ "Neural Networks" While all of those can't be dealt

with at once, potential fraudsters can be identified using the prior mistakes and characteristics of previous crooks.

In the study that was done by (Krishna Rao *et al*., 2021,) a business's intolerance for fraudulent payments may rely on a mixture of factors, including its earnings margin (sales value minus the value of the products marketed). Patience for fraud in payments reduces as the margin falls. Overall, fraud poses a significant financial risk to the client and the providing bank. Technologies such as “chip” and “pin”, 3D security, and fraud detection methods are utilized to reduce fraud. But why is fraud detection required if “3D-safe chip” and pin technologies already exist? There are principally two causes. Compared to the value of fraud detection, the whole value of “3D security”, “chip” and “pin” technology is quite large. For instance, while online retailers are concerned about conversion, “3D secure” significantly (>5%) lowers it. Therefore, when given the option, many online retailers decide to disable “3D Secure” and take control of the risk of payment fraud themselves. The goals of MasterCard fraud detection are to provide opportunities for businesses to earn more money while decreasing penalties resulting from fraudulent payments for both providers banks and retailers.

According to (Lebichot *et al*., 2020,) Although “credit card fraud” only appears a small number of transactions, the resultant financial losses could be enormous. Due to the various qualities of fraudster conduct, automated systems for fraud detection must be developed to catch fraud dealings with high accuracy. In fact, the kind of fraud behavior might change greatly depending on the payment process (such as a business or stockpile terminal), the country, and the population group. It is becoming more and more crucial for transactional organizations to make use of current platforms and modify them to various genres and settings because creating "data-driven FDSs" is expensive. The researchers specifically discuss and present two domain-adaptive techniques in the context of deep neural networks: the first is a novel domain adaptation strategy that relies on the result of new features for transferring effects from the original part to the target domain. Three cutting-edge benchmarks and 5 months' worth of data from "a lot more over 80 million online shopping and in-person purchases" provided by a sizable card issuer allowed for the examination of the two types.

Driven by the rapid progress of electronic commerce technology, "credit card" use has significantly expanded. Credit cards are the most used method of payment, thus as a result, there are an increasing number of fraud instances involving them. Therefore, to counteract these schemes, the researchers desire a robust fraud detection system that correctly identifies fraud. (Bhanusri *et al*., 2020,) This essay has examined the idea of "credit card" scams. Make use of several "machine learning techniques" in this case, such as "random forest," "Naive Bayes," and "logistic regression" to analyze an unbalanced dataset. employing a group of classifiers. The suggested and deployed systems intended for identifying credit card scams have been thoroughly examined, and several methods have been compared. To classify the data, multiple classification models are employed, and statistical metrics involving "accuracy," "precision," "recall," "f1 score," "support," & "confusion matrix" was used to assess the models' efficiency. The finding of this study illustrates how to train and evaluate a classifier using supervised approaches to get a superior result.

In this research work, (Trivedi *et al*., 2020,) present a "machine learning-based" system with a response Credit card fraud identification mechanism. Its response method helps to increase the classifier's cost-effectiveness and detection rate. The effectiveness of various methods was then assessed using slightly skewed "credit card fraud" data sets using the "random forest," "tree

classifiers," "artificial neural networks," "support vector machines," "Naive Bayes," "logistic regression," and "gradient boosting classifier" algorithms. These data sets comprise credit card transaction info obtained from “284,807” trades made by “European account holders”. Both raw content, including pre-processed stuff, and these procedures are comparable.

In this article, financial fraud detection with unbalanced data is discussed. (Izotova and Valiullin, 2021,) Compare different techniques for catching "credit card fraud". On the one hand, employed the different intensity parametric functions to calculate the likelihood of fraud prediction using homogeneous and heterogeneous Poisson processes. On the other hand, to "handle classification" problems, utilize "machine learning algorithms" and several families of "ensemble approaches," "including boosting". The results of the two methods are contrasted. The article also touches on the "false positive" issue.

Considering the status of the economy right now, using credit cards has grown in popularity. The user can use these cards to make significant purchases without lugging around a lot of cash. They have revolutionized cashless transactions and made this thing easy for clients to submit any form of payment. The risks associated with this electronic payment method are impressive despite how useful it is. Like the increase in consumers, credit card fraud is also on the rise. Credit card information can be fraudulently obtained and used to make transactions. Machine learning methods may be used to gather data. This study which is made by (Khatri *et al*., 2020,) examines a variety of popular supervised learning methods for spotting fraudulence in legitimate transactions.

The flow of credit card fraud is a major problem in the financial sector. These frauds prevent cardholders from making purchases, which causes considerable losses for both businesses and financial institutions. Availability of publicly available information is one of the key issues with credit card theft, the significant degree of discrepancy in data, and the growing type of fraud. The most recent advancements in "deep learning" have been employed to combat difficult problems in the variation of domains. In this study, deep learning practices for the flow in "credit card fraud detection" are thoroughly investigated, which is made by (Nguyen *et al*., 2020). Additionally, using three different financial datasets, comparisons of their performance to that of several "machine learning techniques" were made. The results of the experiments show that while comparing the traditional (machine learning models) and the (deep learning methods) the deep leaning methods perfume better.

In this research, (Dileep *et al*., 2021,) new business-making processes appeared in the economic sector as technology advanced. The credit card mechanism is just one of them. Nevertheless, various challenges with the “credit card scamming” approach have arisen because of systemic inadequacies. As a result, both the sector and customers who use "credit cards" are suffering severely. Two methods are used in this context: "Decision Tree-based" "fraud detection" for "credit cards" and "Random Forest-based" "fraud detection." A sample of openly accessible data is employed to rate the model's efficacy. Then, a financial institution's actual global credit card information group is looked at. The data samples also receive additional noise to test the systems' robustness. The study's first method is essential since it builds a user behavior tree that may be used to spot fraud. The creation of an applicant based on activities forest, which will be utilized to try to identify the suspect, is the second technique. The analysis's findings demonstrate unequivocally that the typical approach selected detects credit card theft cases in an accurate and reliable manner.

3 Methodology

3.1 Introduction

Now, let's look at credit card fraud detection research approaches, which often focus on frameworks like Knowledge Discovery in Databases (KDD) or CRISP-DM. The KDD technique is judged more appropriate for this study since the major goal is not the direct deployment of models in a financial setting. However, the objective is to enhance financial institutions' capacities for detecting credit card fraud. While this study does not directly relate to a specific financial application, it does serve as a basic step in that direction. As a result, Azevedo and Santos' (2008) modified KDD approach is used.

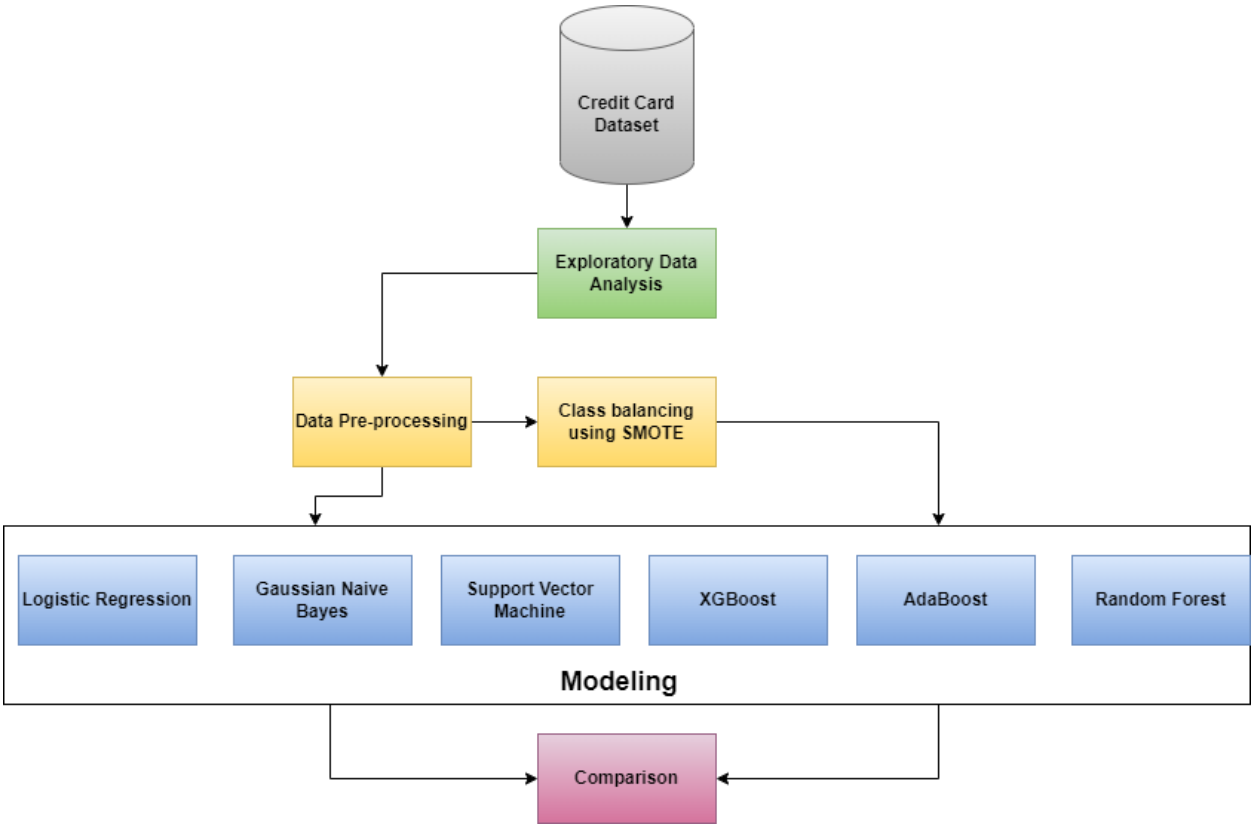

Figure 3.1: Methodology to identifying Credit Card Fraud

3.2 Data Collection

Several phases are included in the updated Knowledge Discovery and Data Mining. Figure 3.1 illustrates an approach to finding credit card fraud. Initially, transactional data is taken from a dataset containing transactions made by European cardholders in September 2013. 284,807 transactions made. The information includes transactions from European cardholders during two days in September 2013. Since just 492 of these transactions were identified as fraudulent, the dataset is substantially distorted, of fraud representing only 0.172% of entire transactions.

The dataset is mostly made up of numerical input variables obtained by a PCA revolution. However, for privacy concerns, specifics about primary characteristics and conditions of material are being withheld. The main elements produced by PCA are the features with the designations

V1 through V28; however, the 'Time' and 'Amount' features are not altered. The (Time) feature shows how many seconds have occurred connecting individual transactions and the dataset's initial transaction. Cost-sensitive learning can use the 'Amount' function, which displays the transaction's value. The 'Class' feature, the final answer variable, has a value of 1 for fraud and 0 for non-fraud.

3.3 Modeling

Different machine learning models are applied in the study to detect credit card fraud. The Random Forest Classifier, Logistic Regression, Gaussian Naive Bayes, XGBoost, AdaBoost, and Support Vector Machine (SVM) classifiers are implemented. Each of these models offers unique strengths, with SVM being known for handling high-dimensional data and XGBoost and AdaBoost being popular for their robustness. These models are applied to the data with imbalanced classes first and balanced classes afterward. Due to the dataset's inherent class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance it. efficiency of each type was then reviewed on the bases of recall, accuracy, precision, and F1 score, giving researchers a thorough picture of how well each model could identify fraudulent activity.

3.4 Conclusion

The KDD (Knowledge Discovery in Databases) approach was precisely customized to address the objectives of the credit card fraud detection inquiry. This performance is perfectly aligned with the project's procedure, which includes the gathering of a large transaction dataset. We will now look at the system design, practical implementation, evaluation, and consequences of the trained models in the following section. The major goal of this study is to distinguish authentic transactions from possibly fraudulent ones, providing crucial insights for improving financial security.

4 Design Specifications

4.1 Introduction

In this section we will examine the challenges engaged in the development of systems for detecting credit card fraud. The versatile Python programming language is used to carry out this study, with the well-known Jupyter IDE serving as our preferred platform. Python, as a powerful language, offers us many libraries that are critical to the effective execution of our system. Sci-kit Learn, Matplotlib, and imblearn are notable libraries that all play critical roles implementation of the fraud finding system.

4.2 Project Design Process Flow

Let us now look more closely at the system design workflow. The design workflow is depicted in Figure 4.2, which is divided into two distinct layers: the presentation layer and the business logic layer. To provide model interpretations and undertake preliminary data analysis, the client presentation makes use of Python visualization tools such as Matplotlib. These visualizations help you comprehend the outcomes and support informed decision-making.

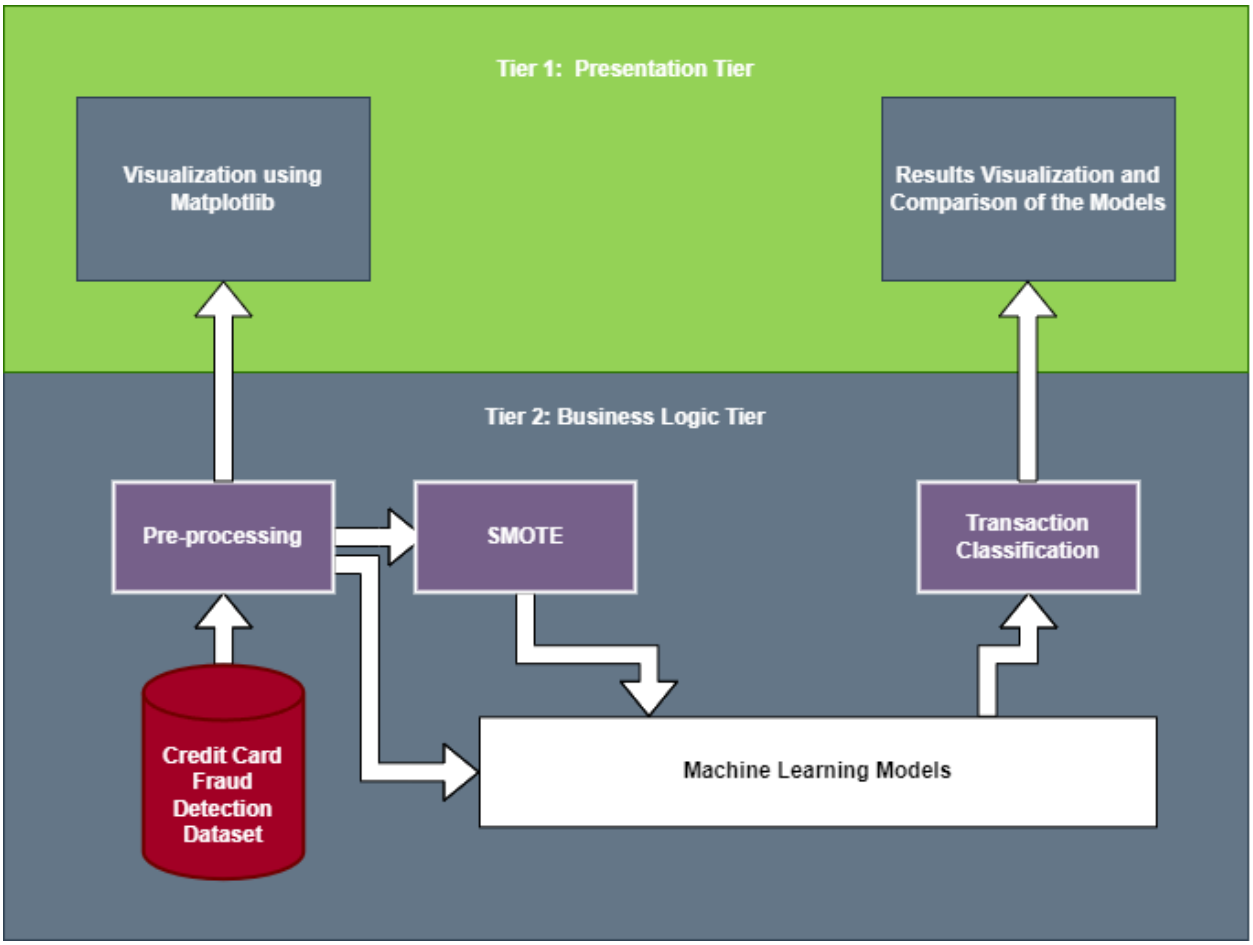

Figure 4.1: Project Design Flow for Detecting Credit Card Fraud

The business logic layer, contrasted with it, includes numerous critical steps which will be critical for the success of our fraudulent detection system. These stages include data acquisition, feature extraction, transformation, model training, and evaluation. By combining these phases, we provide a strong and complete framework for transaction analysis. This method creates a system capable of identifying credit card fraud.

4.3 Synthetic Minority Oversampling Technique (SMOTE)

The method has grown in prominence for dealing with the problem of imbalanced datasets, notably in the context of classification difficulties. In many real-world circumstances, the data shows an imbalance in which one class, frequently the class of interest, is considerably underrepresented in comparison to other classes. In fraud detection, for example, how many fraudulent transactions there were, compared to all other transactions, where a minority class is frequently dwarfed by the count of genuine transactions, the primary class.

The occurrence of such an imbalance can lead to serious problems. Machine learning models may, if improperly trained, tend to predict the majority class for all inputs, achieving good accuracy while failing to find instances of the minority class. This shortcoming is especially troubling because minority class instances are frequently the most significant to identify.

SMOTE comes to the rescue in this situation. SMOTE is a method for rebalancing datasets that generate synthetic minority class samples. SMOTE successfully producing these simulated samples, increases the resemblance for the minority class, allowing the machine learning model to gain insight from an additional equal and extensive dataset.

This method has grown in prominence for dealing with the problem of imbalanced datasets, notably in the context of classification difficulties. In many real-world circumstances, the data shows an imbalance in which one class, frequently the class of interest, is considerably underrepresented in comparison to other classes.

4.4 Gaussian Naïve Bayes Model

A statistical classifier based on the Bayes theorem, specifically the Gaussian Naive Bayes approach, relies on a notable assumption concerning feature independence. It presupposes that a single feature's occurrence or nonexistence within a class is independent of any other feature's existence or non-existence. The model is best suited to high-dimensional datasets. GaussianNB, despite its simplicity, can be surprisingly effective. Because of its probabilistic architecture, which gives a level of interpretability by outputting the chance of a specific event occurring, it is frequently employed as a baseline in text categorization tasks.

4.5 Logistic Regression

In contrast to what its name suggests, logistic regression is a categorization model. It determines the probability that a given instance falls under a specified category. The possibility of a transaction being fraudulent is calculated by fraud detection. The chance which is a specified record viewpoint belonging to a particular class output is then converted into a binary outcome using a threshold. For instance, if the output probability exceeds 0.5, it is classed as class 1; otherwise, it is classified as class 0. One of logistic regression's merits is its ability to provide probabilities, and its model training and prediction timeframes are extremely rapid.

4.6 Random Forest

Random Forest is a versatile collaborative structure that can perform classification and regression. It creates a lot of decision trees by preparing and producing the mode for categorization. One of Random Forest's unique capabilities is the ability to rank the importance of variables in a dataset, giving a clearer picture of the data's structure and the variables that need more attention. It excels at handling huge datasets with increased dimensionality. Furthermore, it can handle missing values, and its ensemble nature aids in the prevention of overfitting.

4.7 Extreme Gradient Boosting (XGBoost) model

Extreme Gradient Boosting is a high-performance gradient-boosted tree implementation. XGBoost, which was created for speed and efficiency, plays a significant role in machine learning competitions due to its performance and adaptability. It is more than simply a classifier; it may also be used for regression, ranking, and custom prediction tasks. The capability of XGBoost to tolerate sparse input and missing values is one of its distinguishing characteristics. Due to the use of L1 (Lasso Regression) and L2 (Ridge Regression) regularization, it avoids overfitting and enhances performance.

4.8 AdaBoost Model

AdaBoost (Adaptive Boosting) is a boosting method for enhancing the performance of underperforming learning systems. It works by focusing on cases that are difficult to categorize and giving them more weight in each succeeding iteration. The goal is to assign weights to both classifiers and data points (or samples) in such a way that classifiers are forced to focus on difficult-to-classify observations. The model's performance is enhanced by this iterative process that minimizes bias and variation.

4.9 Support Vector Machine Model

Support Vector Machines are resilient and flexible supervised machine learning methods. They are utilized for both regression and classification. SVMs work by translating and entering information into a higher-dimensional space, everywhere a hyperplane can be split using or categorize the data points. The brilliance of SVM is that it uses "kernels" to create non-linear decision boundaries. For binary classification problems, SVM tries to determine the "maximum margin" hyperplane that best separates the dataset into classes, making sure that the difference between data points of different classes is as wide as possible.

5 Implementation

5.1 Introduction

This section of the thesis deals with the implementation part wherein the steps mentioned in the methodology (section 3) are discussed. In the presented study, we have implemented various machine learning modalities viz. Support vector machines, XGBoost, AdaBoost, and Random Forest, along with Logistic Regression and Gaussian Naive Bayes.

Four significant metrics are applied to evaluate the output of each of the models that have been applied. F1-score, Recall, Precision, and Accuracy are these measurements.

5.2 Data Collection

During the first stage of our research, we began by importing the dataset. This dataset, which was handily saved as a CSV file, was easily downloaded from a predefined place on Google Drive. Using the panda library's strong features, we easily transferred the dataset into a data frame for pandas, ready for additional research.

After this critical phase, conducted a thorough exploratory data analysis, leaving no stone untouched. We gained a thorough consideration of the dataset's intricate shape and subtle subtleties because of our comprehensive examination. We determined the dataset's flawless integrity, free of missing values or anomalies, through this extensive investigation.

The absence of missing records allowed us to focus our attention on other important parts of the data. This meant that we could devote our full attention to the analysis without having to worry about imputation or data cleaning due to missing values. We were able to extract important insights and draw appropriate conclusions from the dataset thanks to our simple and basic approach.

5.3 Data Preparation

The dataset for the study was sourced from a CSV file on Google Drive, comprising 30 numerical columns with the "Class" column as the target, indicating genuine or fraudulent transactions. No

missing values were discovered after initial tests, however, there was a notable class imbalance because only 0.172% of the 284,807 transactions were fraudulent. To address potential model sensitivities to feature scales, the "Amount" column was standardized using the Standard Scaler, making sure the mean is 0 and that the average deviation is 1. A heatmap was employed to assess feature correlations, leading to the removal of certain features, including 'Time' and several 'V' columns, that exhibited weak correlations with the target "Class" column. After that, the cleaned-up dataset was divided in half, 80/20 split between training and testing. Ensuring a solid analysis of the machine learning models used to detect credit card fraud.

5.4 Experiment 1: Modeling imbalanced dataset

It is significant to emphasize the usage of a variety of algorithms, including ensemble techniques like XGBoost, AdaBoost, and Random Forest, as well as traditional machine learning models like Gaussian Nave Bayes, Logistic Regression, and Support Vector Machines. The inherent nature of the dataset and the complexities of the situation at hand heavily influenced the selection of these models.

Table 5.1 enlists the metrics obtained for the models on the class imbalanced dataset.

Method		Precision	Recall	F1-score	Accuracy (%)
Logistic regression	0	1	1	1	99.8
	1	0.67	0.67	0.67	
GaussianNB	0	1	0.99	1	99.4
	1	0.18	0.67	0.28	
SVM	0	1	1	1	99.8
	1	.52	.13	.21	
Random Forest	0	1	1	1	99.9
	1	.93	.71	.81	
XGBoostClassifier	0	1.0	1.0	1	100
	1	.96	.77	.85	
AdaBoostClassifier	0	1.0	1.0	1.0	99
	1	.84	.67	.75	

Table 5.1: Performance of the models for class imbalanced dataset

The LR model fared well in conditions of exactness, as shown in the table above. However, the model had trouble correctly categorizing fraudulent transactions; its precision, recall, and f1-score were all only 67%, demonstrating its bias towards the class that was in the majority.

The Gaussian Naive Bayes model showcased a commendable level of accuracy, achieving an impressive 99.4%. However, when it came to the task of identifying fraudulent transactions, the model's precision dropped significantly to a mere 18%. This signifies that although the model successfully identified 67% of the actual instances of fraud (as indicated by the recall), a considerable portion of its fraud predictions proved to be incorrect. The F1 score stands at 28%, shedding light on the challenges that the model faces in effectively classifying fraudulent transactions. This low score emphasizes the need for further improvement and refinement to enhance its performance in this specific area.

With a 99.8% accuracy rate, the Support Vector Machine (SVM) model performed brilliantly. However, the results for detecting fraudulent transactions were less than ideal. The SVM model's precision was just 52%, implying that a large percentage of legitimate transactions were incorrectly

labeled as fraudulent. Furthermore, at 13%, the recall rate was shockingly low, indicating that the model missed a significant fraction of actual fraudulent cases.

The Random Forest algorithm delivered outstanding outcomes. It also demonstrated a significant precision of 93% when identifying fraudulent transactions, with a nearly flawless accuracy rate of 99.9%. The model effectively identified a considerable fraction of genuine frauds, with a recall rate of 71%. Furthermore, the F1-score of 81% shows balanced performance regarding precision and recall.

The XGBoostClassifier model also stood out for its 100% accuracy. A recall rate of 77% and a high precision rate of 96% for spotting fraudulent transactions. The model demonstrates its ability to differentiate between valid and fraudulent transactions with an F1-score of 85%.

The AdaBoostClassifier model also performed superbly, with a 99% accuracy rate. While detecting fraudulent transactions, it had a recall rate of 67% and a precision rate of 84%. On the other hand, the F1-score of 75% suggests that there is still room for improvement, notably in memory.

Experiment 2: Modeling a balanced dataset

Table 5.2 enlists the metrics obtained for the models on the class imbalanced dataset.

Method		Precision	Recall	F1-score	Accuracy
Logistic regression	0	1	0.97	0.99	0.973
	1	0.06	0.96	0.11	
GaussianNB	0	1	0.98	0.99	.976
	1	0.06	0.90	0.11	
SVM	0	1	.96	0.98	.96
	1	0.04	.95	0.07	
Random Forest	0	1	.99	1	
	1	.20	.91	.32	.993
XGBoost Classifier	0	1.0	.99	1	.993
	1	.17	.92	.29	
AdaBoost Classifier	0	1	.98	.99	.975
	1	.06	.93	.11	

Table 5.2: Performance of the models for a class-balanced dataset

Regression analysis with Logistic Regression (LR) is one such method. This model earned an excellent 97.3% accuracy. It performed exceptionally well in categorizing authentic transactions, with a precision of one, indicating that it seldom misclassified actual transactions. Its ability to detect fraudulent transactions, however, was restricted, since it only reached a 6% precision. Despite this constraint, the model's recall of 96% for fraudulent transactions shows that it detected a significant fraction of actual frauds. The F1-score of 11% for fraud demonstrates the discrepancy between precision and recall, demonstrating that the LR model incorrectly classifies many legitimate transactions as fraudulent, resulting in a large proportion of false positives.

The accuracy of the GaussianNB model was 97.6%. It obtained a high precision of 1 for real transactions, like the LR model, but struggled with a precision of 6% for fraudulent transactions. However, with a fraud recall of 90%, it appears that the model successfully identified most of the genuine false transactions. The F1-score of 11% of frauds indicates the difficulty in balancing precision and memory for the minority class.

The Support Vector Machine (SVM) model will now be discussed. For genuine transactions, it achieved an accuracy of 96% and a precision of 1. However, it encountered difficulties due to a poor precision of 4% for fraudulent transactions. On the plus side, its fraud recall of 95% shows that it might detect most true fraudulent occurrences. Regrettably, the F1-score for frauds is the lowest of the models, demonstrating a considerable imbalance among recall and precision for the minority class.

Checking out the Random Forest model, which has a 99.3% accuracy rate. It obtained flawless precision and recall for authentic transactions, indicating that it rarely misclassified actual transactions. It achieved a precision of 20% and a commendable recall of 91% when it came to fraudulent transactions. The F1-score for frauds is 32%, which is significantly higher than the prior models, demonstrating improved precision and recall performance.

Moving on, the XGBoost Classifier attained a 99.3% accuracy rate. It perfectly classified authentic transactions with precision and recall of one, just like the Random Forest model. It achieved a high recall of 92% for fraudulent transactions and a precision of 17% overall. When compared to the LR, GaussianNB, and SVM models, the F1-score of 29% for frauds shows a better balance of precision and recall, although there is still potential for improvement.

Finally, the AdaBoost Classifier achieved 97.5% accuracy. It was extremely precise, 1 aimed at authentic dealings although suffered of precision by 6% for fraudulent transactions, as did the LR and GaussianNB models. However, with a recall of 93% for fraud, the algorithm appears to have identified a considerable fraction of actual scams. The LR and GaussianNB models have trouble finding a compromise between precision and recall for the minority class, which is shown in the F1-score of 11% for frauds.

6 Discussion

The domain of fraud detection heavily depends on the application of machine learning models. which may become quite complex due to the dataset's innate imbalance. In most cases, The number of legitimate transactions is vastly outweighed by genuine ones, posing a challenge for accurate detection. Utilizing various methods like SMOTE were introduced, but they come with their own set of complexities.

Before implementing SMOTE, the models used in fraud detection generally achieved high accuracy rates. For example, the Logistic Regression (LR) model boasted an accuracy of nearly 99.8%. However, a closer examination of the metrics revealed that while these models excelled at identifying genuine transactions, they struggled with correctly classifying fraudulent ones. The precision for fraudulent transactions in the LR model was only 67%, indicating that a substantial portion of flagged fraudulent transactions were genuine. This bias towards the majority class is a common problem with models trained on imbalanced datasets.

After implementing SMOTE, there was a noticeable shift in model performance. Although the accuracy of models like LR slightly decreased to 97.3%, the recall for fraudulent transactions increased significantly to 96%. This means that a greater proportion of actual frauds were now detected by the models. However, this improvement came at the cost of precision, which dropped to a mere 6%. This decline indicates that while the models were identifying more frauds, they were also misclassifying many genuine transactions as fraudulent.

The challenges associated with SMOTE arise from its synthetic sample generation process. By creating synthetic samples, SMOTE can introduce noise into the dataset, leading the models to overfit these synthetic samples The loss in precision for models like LR, GaussianNB, and SVM after incorporating SMOTE is indicative that the models may struggle to generalize successfully on the real test data. These models began flagging a higher number of genuine transactions as fraudulent, resulting in an increased number of false positives.

Another challenge with SMOTE is that it operates in the feature space rather than the data distribution. This means that the synthetic samples may not accurately represent the characteristics of actual fraudulent transactions, potentially misleading the models during training.

The failure of some models after class balancing with SMOTE can be attributed to several factors. Firstly, overfitting synthetic samples can reduce the models' performance on actual test data. Secondly, while SMOTE expands the sample size for the minority class, it may not capture the intricate patterns of genuine fraudulent transactions, leading to models being trained on data that does not truly represent the underlying fraudulent behaviors. Lastly, the decision boundary between fraudulent and genuine transactions becomes more complex after implementing SMOTE, which can pose challenges for models like SVM that rely on finding an optimal boundary. This complexity often results in reduced precision.

When comparing different models, Random Forest and XGBoost stood out for their balanced performance after implementing SMOTE. These ensemble methods are designed to handle the noise and complexity introduced by SMOTE better than simpler models like LR or SVM. They can effectively capture intricate patterns and are less prone to overfitting, making them more robust in such scenarios. On the other hand, models like SVM displayed a significant drop in performance after implementing SMOTE, with precision for fraudulent transactions as low as 4%. This indicates their struggle to find an optimal decision boundary in the transformed feature space.

In conclusion, SMOTE is a valuable tool for data scientists dealing with imbalanced datasets in fraud detection. However, it is crucial to understand its implications and consider various factors such as the choice of models, dataset intricacies, and the nature of the problem. To effectively evaluate model performance, especially in vital applications like fraud detection, it is crucial to delve deeply into metrics beyond accuracy.

7 Conclusion and Future Work

The complexity and difficulties of detecting fraudulent transactions in a sea of legitimate ones are revealed through research on credit card fraud detection using machine learning. The dataset's intrinsic imbalance, in which the genuine transactions, are more than the fraud, provides a unique problem. While standard accuracy measurements may show remarkable model performance, a closer look at precision, recall, and F1-score reveals the true picture. Techniques such as SMOTE, which are expressly developed to address this imbalance, present their own set of challenges, emphasizing the value of a comprehensive strategy for model training & evaluation.

Machine learning models, particularly ensemble approaches such as Random Forest and XGBoost, have shown the potential in discovering fraud trends that can be identified. However, the voyage is far from over. The dynamic nature of fraudulent actions, paired with the ever-changing financial landscape, needs ongoing updates and analyses of models to ensure their effectiveness.

Using neural network developments such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) to create deep learning models, future studies could look at credit card fraudulent recognition. These models have demonstrated potential in tasks requiring sequence prediction, and they could be able to identify the temporal patterns present in credit card transactions.

Furthermore, the current research can be expanded by going deeper into feature engineering. The prediction power of the model can be improved by generating new features from current ones and incorporating domain-specific knowledge.

Future studies can investigate anomaly detection techniques rather than just classification-based approaches. These strategies are intended to discover patterns that depart from the norm and may be especially useful in the identification of fraud.

Future research can also concentrate on employing real-time machine learning models to assess and forewarn against fraud suspicions while transactions take place. This real-time fraud detection technique supports overall fraud prevention efforts while enhancing the security of credit card transactions.

References

Al-amri, R., Murugesan, R.K., Man, M., Abdulateef, A.F., Al-Sharafi, M.A. and Alkahtani, A.A., 2021. A review of machine learning and deep learning techniques for anomaly detection in IoT data. Applied Sciences, 11(12), p.5320.

Alarfaj, F.K., Malik, I., Khan, H.U., Almusallam, N., Ramzan, M. and Ahmed, M., 2022. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access*, *10*, pp.39700-39715.

Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H. and Saif, A., 2022. Financial fraud detection based on machine learning: a systematic literature review. Applied Sciences, 12(19), p.9637.

Azhan, M. and Meraj, S., 2020, December. Credit card fraud detection using machine learning and deep learning techniques. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 514-518). IEEE.

Bhanusri, A., Valli, K.R.S., Jyothi, P., Sai, G.V. and Rohith, R., 2020. Credit card fraud detection using Machine learning algorithms. Journal of Research in Humanities and Social Science, 8(2), pp.4-11.

Chen, J.I.Z. and Lai, K.L., 2021. Deep convolution neural network model for credit-card fraud detection and alert. Journal of Artificial Intelligence and Capsule Networks, 3(2), pp.101-112.

Dileep, M.R., Navaneeth, A.V. and Abhishek, M., 2021, February. A novel approach for credit card fraud detection using decision tree and random forest algorithms. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 1025-1028). IEEE.

Ghosal, A., Nandy, A., Das, A.K., Goswami, S. and Panday, M., 2020. A short review on different clustering techniques and their applications. Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, pp.69-83.

Izotova, A. and Valiullin, A., 2021. Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection. Procedia Computer Science, 186, pp.721-726.

Khatri, S., Arora, A. and Agrawal, A.P., 2020, January. Supervised machine learning algorithms for credit card fraud detection: a comparison. In 2020 10th international conference on cloud computing, data science & engineering (confluence) (pp. 680-683). IEEE.

Krishna Rao, N.V., Harika Devi, Y., Shalini, N., Harika, A., Divyavani, V. and Mangathayaru, N., 2021. Credit card fraud detection using spark and machine learning techniques. In Machine Learning Technologies and Applications: Proceedings of ICACECS 2020 (pp. 163-172). Singapore: Springer Singapore.

Lebichot, B., Le Borgne, Y.A., He-Guelton, L., Oblé, F. and Bontempi, G., 2020. Deep-learning domain adaptation techniques for credit cards fraud detection. In Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, held at Sestri Levante, Genova, Italy 16-18 April 2019 (pp. 78-88). Springer International Publishing.

Lucas, Y. and Jurgovsky, J., 2020. Credit card fraud detection using machine learning: A survey. arXiv preprint arXiv:2010.06479.

MATHEW, D.T.E., 2023. AN ENSEMBLE MACHINE LEARNING MODEL FOR CLASSIFICATION OF CREDIT CARD FRADULENT TRANSACTIONS. Journal of Theoretical and Applied Information Technology, 101(9).

Nguyen, T.T., Tahir, H., Abdelrazek, M. and Babar, A., 2020. Deep learning methods for credit card fraud detection. arXiv preprint arXiv:2012.03754.

Taha, A.A. and Malebary, S.J., 2020. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, pp.25579-25587.

Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J. and Singh, A.K., 2021. Credit card fraud detection using machine learning: a study. arXiv preprint arXiv:2108.10005.

Trivedi, N.K., Simaiya, S., Lilhore, U.K. and Sharma, S.K., 2020. An efficient credit card fraud detection model based on machine learning methods. International Journal of Advanced Science and Technology, 29(5), pp.3414-3424.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D. and Saeed, J., 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, 1(2), pp.56-70.

Zhang, X., Han, Y., Xu, W. and Wang, Q., 2021. HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, pp.302-316.