

**DEVELOPMENT OF A BANKING RISK-ENGINE TO PREDICT
CREDIT, MARKET AND OPERATIONAL RISKS USING
MACHINE LEARNING**

by

MD. NOOR ALAM

POST GRADUATE DIPLOMA IN INFORMATION AND COMMUNICATION
TECHNOLOGY




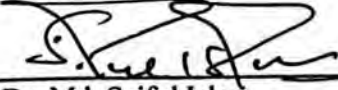
Institute of Information and Communication Technology (IICT)
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY (BUET)


November 2022

This project titled “DEVELOPMENT OF A BANKING RISK-ENGINE TO PREDICT CREDIT, MARKET AND OPERATIONAL RISKS USING MACHINE LEARNING” submitted by MD. NOOR ALAM, Roll No: 1017311002, Session: October/2017, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Post Graduate Diploma in Information and Communication Technology on November 5, 2022.

Board of Examiners

- 
1.

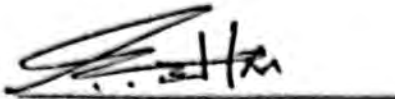
Dr. Md. Rubaiyat Hossain Mondal
Professor
IICT, BUET, Dhaka. Chairman
(Supervisor)
- 
2.

Dr. Md. Saiful Islam
Professor
IICT, BUET, Dhaka. Member
- 
3.

Dr. Md. Liakot Ali
Professor
IICT, BUET, Dhaka. Member

Candidate's Declaration

It is hereby declared that this report or any part of it has not been submitted elsewhere for the award of any degree or diploma.



MD. NOOR ALAM

ID: 1017311002

Dedicated
To
My Parents and Family Members

Table of Contents

Board of Examiners.....	ii
Candidate’s Declaration.....	iii
Dedication.....	iv
Table of Contents.....	v-vi
List of Tables.....	viii
Abbreviation & Key Terms.....	ix
Acknowledgment.....	x
Abstract.....	xi
1 Introduction.....	01-03
1.1 Overview.....	01-02
1.2 Motivation and Objectives of the research.....	03
1.3 Organization of the report.....	03
2 Discussion on Credit Card Operations	04-11
2.1 Guidelines on Credit Card	04-05
2.2 Issuing Authority, Shareholder and Card Types.....	05-06
2.3 Marketing Strategies.....	06-07
2.4 Consideration for Card issuance.....	07
2.5 Wrongful Billing and Recovery.....	08
2.6 Protection of Customer Rights and Dispute Resolution.....	08
2.7 Credit, Market and Operational Risk.....	09-10
2.8 Internal control, monitoring, and Fraud Control.....	10-11
3 ML Algorithms for CCFD.....	12-15
3.1 ML.....	12-13
3.2 ML effectiveness.....	13
3.3 Different ML Method.....	14-15
3.4 Data Acquisition and Processing in ML.....	15-16
4 The Proposed CCFD.....	17-25
4.1 Overview.....	17
4.2 Experimental Setup.....	17-18
4.3 Explanation of the Dataset.....	18-19
4.4 Methodology	20
4.5 Performance of RF Classifier	20-22

4.6 Performance of AdaBoost	22-23
4.7 Performance of CatBoost	23-25
5 Conclusion.....	26-27
5.1 Conclusion.....	26
5.2 Future Work.....	27
References.....	28-31

LIST OF FIGURES

Figure No.	Figure Caption	Page No.
Figure 3.1	Block diagram of decision flow architecture for Machine learning systems	13
Figure 4.1	Time density plot	19
Figure 4.2	Work flow diagram of the methodology	20
Figure 4.3	Important feature selection using RF	21
Figure 4.4	Basics of confusion matrix	22
Figure 4.5	Confusion matrix using RF	22
Figure 4.6	Important feature selection using AdaBoost	23
Figure 4.7	Confusion matrix using AdaBoost	23
Figure 4.8	Important feature selection using CatBoost	24
Figure 4.9	Confusion matrix using CatBoost	24

LIST OF TABLES

Table No.	Table Caption	Page No.
Table 4.1	Statistical Analysis of Classes	19
Table 4.2(a)	Performance evaluation of classifiers	24
Table 4.2(b)	Performance evaluation of classifiers	24

ABBREVIATION & KEY TERMS

BB	Bangladesh Bank
CCFD	Credit Card Fraud Detection
ML	Machine Learning
RF	Random Forest
TPR	True Positive Rate
FPR	False Positive Rate
TP	True Positive
TN	True Negative
CM	Confusion matrix
FDR	False Detection Rate
CR	Credit Risk
MR	Market Risk
OR	Operation Risk

Acknowledgment

First of all, I would like to convey my gratitude to Almighty Allah for giving me the opportunity to accomplish this project. I want to thank my supervisor Dr. Md. Rubaiyat Hossain Mondal, Professor, IICT, BUET for giving me the chance to explore such an interesting field of research and providing help and advice whenever I needed it. Without his proper guidance, advice, continual encouragement, and active involvement in this process of this work, it would not have been feasible.

A big thank also goes to all the teachers, officers, and staff of Information and Communication Technology (IICT) for giving me their kind support and information during the study.

Finally, I am very grateful to my parents and family members whose continuous support all over my life has brought me this far in my career.

Abstract

Effective management of credit risk is an essential component of any banking system. The identification of credit card fraud in financial transactions is one of the most pressing issues facing financial institutions today. Fraud involving credit cards is significantly on the rise as modern technology continues to advance on a daily basis. Credit card fraud results in annual losses of billions of dollars for a number of different financial institutions. So, the utilization of fraud detection strategies is essential for both banking and non-banking financial institutions in order to reduce the amount of money lost. In order to anticipate credit card fraud detection under credit risks of banks, the use of machine learning (ML) methods can be effective. Appropriate ML algorithms have the potential to identify fraudulent and lawful transactions. However, there is difficulty in exchanging datasets and ideas of fraud detection among different banks due to privacy issues, and the lack of datasets is an obstacle to credit card fraud detection techniques. This work focuses mostly on analyzing the effectiveness of multiple ML classifiers, such as Random Forest (RF), AdaBoost, and CatBoost, with the intention of categorizing fraudulent behaviors involving credit cards. The dataset considered in this research is the transactions performed with credit cards by European cardholders in September 2013 where there are 492 fraudulent transactions, and the total number of transactions is 284,807. Results show that RF and CatBoost achieve an accuracy value of 99.92%, while AdaBoost exhibits an accuracy value of 99.91% for fraud detection. In the future, there will be a need for large real-time datasets to train the model while protecting privacy.

CHAPTER 1

Introduction

1.1 Overview

Most financial institutions now offer internet banking to the public. E-payments are necessary for the competitive financial environment that we live in today. They have made shopping simple. Financial organizations give clients cards so they can shop without cash. Credit cards, like debit cards, protect consumers from damaged, lost, or stolen products. Customers must check credit card transactions with merchants. These stats indicate how card-based transactions become popular with end-users. Due to the volume of global transactions in this area, fraudsters target this group. Despite their perks, credit cards pose security and fraud risks. Banks and financial institutions face credit card fraud. Unapproved people use credit cards to fraudulently get money or property. Unsecured platforms and websites can steal credit card information. Identity theft methods also offer them. Fraudsters can illegally access consumers' credit and debit card numbers. Credit and debit card theft is a major cause of financial losses in the finance business, according to "U.K. finance". Due to technological advancements, worldwide financial losses are significant. Credit card fraud detection (CCFD) is crucial to reducing financial losses.

Managing the risk of credit, market and operation is vital for any banking [1-10] system. In this regard, Basel Accords [10] are introduced globally to set minimum capital requirements for banks. The integrated risk management [5] business intelligence (BI) solution is a system that helps banks to comply with both Basel [10] Accords that are banking supervision regulations [2], and local regulatory reporting [10]. The elements that are considered in credit scoring models are payment history, credit utilization ratio,

length of credit history, different types of credit, soft and hard credit checks/inquiries, diversity of credit, total payment ratio, etc. [3]. Market risk is when an investment's value declines due to market movements. The market risk depends on stock prices, interest rates, credit spreads, foreign exchange rates, commodity prices, and other public-market indicators. Operational risk is the possibility of loss from inadequate or failing procedures, systems, or policies, personnel errors, system failures, fraud, or other criminal behavior, and any occurrence that interrupts business processes. The application of machine learning (ML) algorithms can be useful to predict the above-mentioned three risks of banks. This project focuses on the application of ML algorithms on the management of credit, market and operational risks in the banking sector [4, 6].

Various methods, including k-nearest neighbor (KNN), Bayesian algorithm, support vector machines (SVM), and neural networks, etc., [11, 12] have been used to detect fraud. The statistical methods for identifying fraud have been divided into two broad categories: unsupervised and supervised. In supervised methods of fraud detection, models are implemented according to the sample of genuine and fraudulent transactions. In unsupervised fraud detection algorithms [8,9], however, transactions that are atypical or anomalous are identified as prospective instances of fraudulent transactions. These methods of fraud detection forecast the likelihood of fraud in all specified transactions.

Multiple researchers implemented multiple ML algorithms. Masoumeh Zareapoor made use of the Nave Bayes (NB), KNN, SVM classifier, and Bagging ensemble learning algorithms to predict credit card fraud [13-14]. Vaishnavi Nath Darnadula employed logistic regression (LR), support vector machine (SVM), random forest (RF), and decision tree (DT) to analyze the behavioral patterns [7] in the credit card fraud dataset.

1.2 Motivation and Objectives of the research

ML has the potential to distinguish between fraudulent and legal transactions [15-26]. However, due to concerns over confidentiality and safety, the free flow of real-time data and ideas between various financial institutions might be challenging in many instances [27-36]. The lack of data is an obstacle to developing reliable CCFD techniques. CCFD uses complicated approaches to produce an effective detection system. Numerous and diverse credit card transactions occur. Users utilize credit cards for different purposes based on geography and currency, therefore fraudulent transactions are diverse. CCFD also struggles with fraudulent transactions and uneven datasets. Obtaining real-time credit card data is difficult. The existing laws prevent banks from disclosing customer data. It is difficult for academics to acquire credit card fraud datasets. ML in CCFD needs investigation. Therefore, there is a need for research on the application of ML in CCFD.

The main objective of this project is to apply ML algorithms appropriately to assess three important risks of banks. The specific aims of the work are as follows.

- (1) This project aims to develop a system capable of CCFD using ML algorithms. To achieve the goal, Credit Card fraud detector models based on Random Forest (RF), AdaBoost, and CatBoost classifiers will be built and trained with a credit card data set.
- (2) The classification accuracy of different ML-based CCFD schemes will be compared.

1.3 Organization of the report

The structure of this report is as follows: Chapter 2 Discussion on Credit Card Operations. Chapter 3 Discussion on ML Models. Chapter 4 The Proposed CCFD. Finally, the Conclusion is exhibited in Chapter 5.

CHAPTER 2

Discussion on Credit Card Operations

2.1 Guidelines on Credit Card

Cards as a method of payment for the acquisition of goods and services are expanding daily. The use of Credit Cards is on the rise due to their convenience, security, and the expansion of electronic point-of-sale (POS) terminals, as well as the incentives offered by Credit Card Issuers. In light of the urgent need to enhance the electronic payment mechanism and the expansion of domestic Credit Card operations, it is necessary to establish a framework of rules/regulations for Credit Card issuing Banks to manage the risks associated with their credit card business and protect the customers' interests. Bangladesh Bank (BB), which is the central bank of the country, may issue the following guidelines on Credit Card operations to ensure the safe, secure, and efficient use of Credit Cards as a payment instrument. [22].

Typically, a credit card refers to a card issued to a customer in order to enable or permit them to purchase products and online services. Each card has its own credit limit. Depending on the limit, users can advance cash withdrawals. In a short period, numerous attackers can obtain large sums of money, and the fraud is occasionally uncovered days later [1, 2]. Online or offline, fraudsters require sensitive information. Credit card numbers, expiration dates, SSNs, bank account details, etc. For offline purchases, a hacker must have a credit card. Fraudsters should get the customer identified via online payment (phone or Internet). Digital banking has increased credit card fraud [3]. Bangladesh Bank has already taken numerous measures and attempted

to implement highly secure systems for monitoring transactions and detecting fraud as quickly as possible.

There are some well-known instances of credit card fraud in the world. Approximately forty million credit card data records were compromised in 2012 due to an assault on Adobe Systems. [4]. In 2014, hackers stole around 56 million credit card details from the payment system of Home Depot [5]. There are a variety of CCFD methods. Each strategy aims to increase the identification rate while minimizing false alarms.

2.2 Issuing Authority, Stakeholders and Card Types

Credit Card Availability in Bangladesh All Scheduled Commercial Banks (SCBs) in Bangladesh may issue Taka. But Foreign Currency Credit Cards can only be issued by Authorized Dealer Banks [10-22]. Credit Card program participants may include: Cardholders are individuals authorized to use Credit Cards for payment of goods and services; Card Issuers are institutions that issue credit cards; Merchants are entities that agree to accept Credit Cards for payment of goods and services; Merchant acquirers are Banks that enter into agreements with merchants to process their Credit Card transactions; and Credit Card Associations are organizations that license Card Issuers to issue Credit Cards under their brand (i.e., Card Issuers and Merchant Acquirers) [22].

The phrase "Credit Card" generally refers to a plastic card issued by Scheduled Commercial Banks (SCBs) and allocated to a Cardholder with a credit limit that can be used to acquire cash advances or make purchases on credit. Credit Cards enable Cardholders to make purchases over time and carry a balance from one billing cycle to the next. Credit card transactions are often due following a grace period during which

no interest or finance charges are assessed. After the due date, interest is added to the unpaid balance of a charge card [22].

Credit cards fall into two categories:

General-use cards: Credit Card Associations (VISA, Master Card, JCB, AMEX) issue these cards, which are accepted by many merchants. Most card-issuing banks offer general-purpose credit cards. Banks categorise these cards as platinum, gold, or classic to differentiate services and income eligibility.

Few shops accept private label cards (e.g., a departmental store). Card Issuers may issue Corporate Credit Cards to corporate customers' employees. BB doesn't require credit card approval. Banks can issue credit cards with board approval [22].

2.3 Marketing strategies

Following guidelines can help Card Issuers create and implement marketing strategies:

All applicable Credit Card terms and conditions shall be communicated to the Customer in their preferred language (both English and Bangla) and in visible font size. These terms and conditions must be clear. These terms must define the Cardholder's responsibilities and liabilities, eligibility requirements, fees, and calculating method. Credit Cardholders cannot be charged beyond the contract's fees. Card Issuers must provide terms and conditions online. Card issuers must deliver monthly account statements unless no transactions or balances have occurred since the preceding statement. The card Issuer is accountable for all unlawful transactions after losing or stealing the card. Card issuers must publish their code of conduct/institutional policy on Credit Card operations on their website during the marketing process. The Card Issuer can't unilaterally upgrade a Credit Card's kind or limit without alerting cardholders. Any stipulation, caveat, clause, or restriction that could limit customers'

rights unreasonably is not allowed. The terms should state that credit cards cannot be used for criminal activities under Bangladeshi law. If a cardholder engages in criminal behavior, the card issuer must suspend the card and notify Bangladesh Bank. Classify and handle outstanding credit card debts according to applicable laws and regulations. [22].

2.4 Consideration for Card Issuance

This section discusses the factors for issuing principal credit cards [10, 11, 22]. Only 18-year-old Bangladeshi citizens or residents who can pay their own bills can get a credit card. Both TINs are legitimate. Card Issuers must follow "Know Your Customer" requirements. A card issuer must receive a properly completed, signed application and all required evidence from a customer. Before providing a credit card, banks should analyze the applicant's credit risk and do PRM checks. Card Issuer collects credit information from the CIB (CIB). Card Issuers may not issue new Credit Cards to customers with the same income information without first collecting their outstanding credit balances. Banks should prevent multiple cardholders from overspending. Bangladeshis can only receive Taka/foreign-currency credit cards. Foreign Exchange Regulation Act, 1947, Guidelines for Foreign Exchange Transactions (vol-1), 2009, and any further Foreign Exchange Policy Department directives control credit card issuing to non-residents/foreign nationals. Maximum-limit cards must be collateralized. Terminals should require a PIN for card security. Before issuing cards, issuers check the spending limit and card count. The limit will not be surpassed no matter how many cards issuers offer. Credit limit and creditworthiness are based on a cardholder's income.

2.5 Wrongful Billing and Recovery

The Card Issuer should ensure that customers do not get incorrect bills. In the event that a customer disputes a bill, the Card Issuer should give an explanation and, if required, verifiable evidence to the consumer within sixty (60) days in an effort to resolve the dispute amicably [22]. At the time of Credit Card issuance, the Card Issuer must inform the Customer of the recovery method in the event of delinquent payments. However, they should not engage in any action that is contrary to the public interest, and they should adhere to the values of honesty and good faith. Recovery letters shall be sent to the Customer's last known address, bearing the designation, contact number(s), and office address of the responsible party. iii. Any verbal or physical harassment or threats towards Customers, their family members, references, or friends shall not be used during the collection process [22].

2.6 Protection of Customer Rights and Dispute Resolution

The Card Issuer shall not disclose any customer information collected at the time of account opening or Credit Card issuance to any other person or organization without the express agreement of the Customer. Unsolicited loans or other facilities shall not be available to Credit Card Customers [22]. Card Issuers must have dispute resolution and service procedures. The Card Issuer must escalate unresolved calls. Online details should be posted. The website should list important executives and the dispute resolution officer. Even phone complaints should be acknowledged and tracked. The Card Issuer should resolve the Customer's disputed transaction quickly and according to VISA, Master Card, or any other international Card company/association, considering the transaction nature, distance, time zone, etc. [22].

2.7 Credit, Market and Operational Risks

Investors are influenced by a bank's risk management. A lack of risk management can diminish a bank's profitability due to loan losses, even if it generates high revenues. Value investors are more likely to invest in a profitable, low-risk bank. Governments can support responsible management and decision-making by recognizing bank risks.

Banks' largest risk is the credit risk which happens when borrowers or counterparties breach contracts. When borrowers miss a loan payment, for example. Mortgages, credit cards, and fixed-income investments can default. Derivatives and guarantees can also be breached. Due to their business strategy, banks can't be totally insulated from credit risk, but they can reduce it. Since industry or issuer decline is typically unpredictable, banks diversify. By doing so, banks are less likely to be overexposed during a credit slump. To reduce risk, they can lend to persons with good credit, deal with high-quality counterparties, or use collateral.

Operational risk is the risk of loss caused by people, systems, or processes. Simple corporate activities like retail banking and asset management have low operational risk, while sales and trading have higher risk. Internal fraud or transaction errors are human-caused losses. Large-scale fraud can result from bank cybersecurity breaches. It allows hackers to steal client information and money from banks and blackmail them. Banks lose capital and customer trust in this situation. A damaged bank's reputation makes it harder to attract deposits or business. Most operational losses are frequent and minor, whereas less frequent/high-impact losses concern regulators and risk officers. Equipment failures, inadequate management practices, personnel errors, internal and

external fraud, IT system interruptions, and natural disasters are common operational risk events. Internal fraud encompasses conduct intended to defraud, misappropriate property, or circumvent regulations, the law, or business policy. Financial fraud includes bank personnel embezzling funds. External fraud occurs when a third party defrauds, steals, or breaks the law. Credit card fraud is an example. Internal and external fraud may coexist if external fraud is performed in conjunction with corporate workers. Most fraud is committed by external third parties, although fraud detection technologies have been employed to mitigate operational risk. IT disruptions are IT systems that fail and cause substantial losses, which can affect an institution or the financial system. The MasterCard computer virus stole client data for fraud. This loss may be external fraud.

Market risk is the risk of losses in on- and off-balance sheet positions of a FI due to adverse fluctuations in market rates or prices, such as interest rates, stock prices, foreign exchange rates, commodity prices, and credit spreads. Interest rate risk occurs when a change in interest rates affects a FI's cash flows. This risk stems from mismatched cash flow repricing dates (including final maturities). It affects FI earnings and net asset values. The amount of risk depends on interest rate movements and the mismatch's maturity structure. Equity risk is the financial risk of holding investment equity. It frequently relates to common or preferred stock purchases. Equity risk could develop if the FI's listed shares lose value.

2.8 Internal control, monitoring, and Fraud Control

Card Issuers need good management, accounting, and control methods to minimize financial and non-financial risks. Before introducing new products or services, the card issuer must do a risk and feasibility analysis. The card Issuer must have enough skilled

people to operate the system. Card Issuers must have well-documented operational and technical procedures to ensure dependability, honesty, and punctuality. Card Issuers must have a reasonable, effective, well-documented, and proven business contingency plan. Every six months, banks must prepare a Credit Card Review Report. Card Issuers should have due diligence and monitoring mechanisms for outsourced connections that could affect system functioning. Card Issuers self-evaluate employee compliance with regulations, guidelines, and code of conduct. Compliance officials or independent assessors may examine self-assessment auditors. Approval power and functions should be separated. Internal audits and controls avoid mismanagement and fraud. Credit card issuers must maintain a database and data recovery system [22].

Card Issuers should set up internal control systems to combat fraud, participate in fraud prevention committees/task forces, and conduct proactive fraud control and enforcement measures. card-issuing Banks should block a missing card as soon as a customer reports it and notify BB. Banks may provide consumers insurance to cover lost card obligations. Only Cardholders willing to pay the premium should be covered for lost cards. Fraud monitoring and investigation specialists should work in risk management [22]. Bangladesh Bank retains the power to penalize card-issuing Scheduled Banks under the Bank Company Act, 1991 (Amended up to 2013) for violating these standards. [22].

CHAPTER 3

ML Algorithms for CCFD

This chapter discusses the necessary theoretical knowledge for understanding the methods. First, a relevant discussion about ML algorithms, including Random-Forest (RF), AdaBoost, and CatBoost classifiers, is provided in detail. Finally, it is explained how these disciplines are combined in ML.

3.1 ML

ML allows computers to learn from data without being explicitly programmed [32]. ML and computational statistics both use computers to make predictions. Mathematical optimization provides methods, theory, and application domains. Regression and classification are supervised problems. Clustering, visualization, dimensionality reduction, and association rule learning are unsupervised tasks. Supervised or unsupervised ML algorithms are common [32]. Semi-Supervised and Reinforcement Learning are others. Supervised Learning uses labeled datasets for Classification and Regression. Supervised algorithms require ML skills to give input, output, and feedback during training. Data scientists choose certain variables or traits to investigate and predict. The trained algorithm learns fresh data [8, 32].

Unsupervised Learning involves unlabeled data and unknown results. Clustering and dimension reduction are two methods. Unsupervised algorithms don't need outcome data. Deep learning is used to examine data and draw judgments iteratively. Unsupervised learning methods are used for picture recognition, speech-to-text, and natural language processing. Semi-supervised learning is a blend of labeled and unlabeled data. Labeled data includes names, types, and numbers. Unlabeled data lacks tags. Reinforcement Learning has no training data sets and is used to play games. This system combs through millions of training data instances to detect minor relationships.

These algorithms are only possible in the age of big data because they require vast training data. [8, 32].

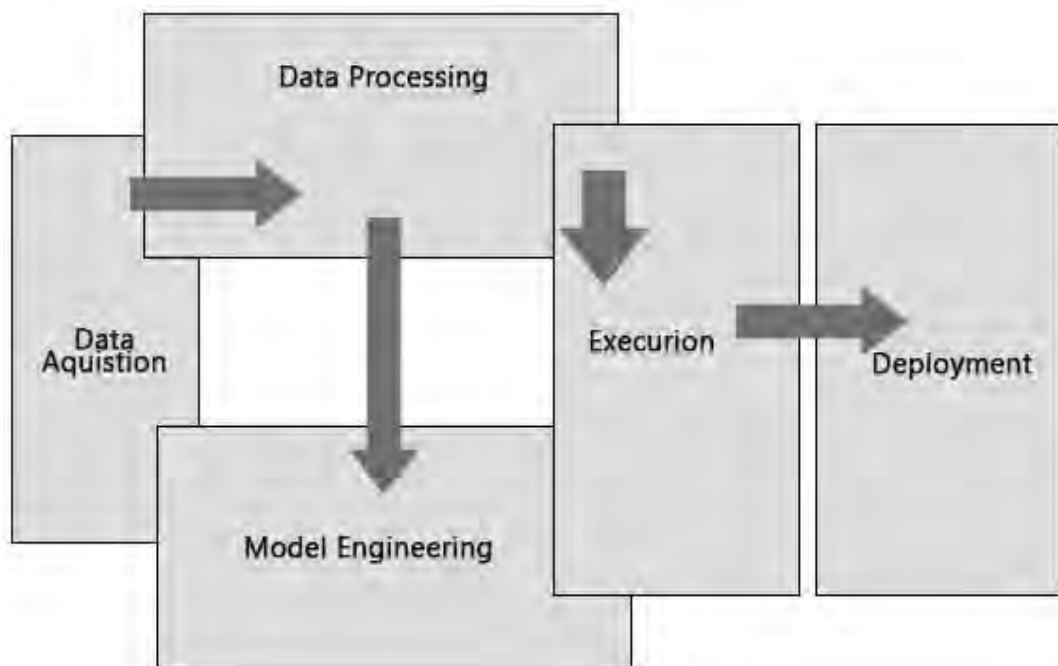


Figure 3.1: Block diagram of decision flow architecture for ML systems [28]

3.2 ML effectiveness

ML is the study of algorithms that improve with experience. [1]. It is considered AI. ML algorithms generate a model using "training data" to make predictions or judgments without being explicitly programmed. [2]. ML techniques are utilized in email filtering and computer vision, where traditional algorithms are difficult or impossible to create. Not all ML is statistical learning, although a subset is. Computational statistics focuses on making predictions with computers. Mathematical optimization provides ML with methods, theory, and applications. Data mining focuses on unsupervised exploratory data analysis. [25, 26]. In business, ML is called predictive analytics [27].

3.3 Different ML Models

Random forests or random decision forests are ensemble learning methods for classification, regression, and other applications. Random forests output the class most trees choose for classification problems. For regression tasks, the average tree forecast is returned. [29, 30]. Random decision forests remedy decision trees' training set overfitting [31]. Random forests outperform decision trees but are less accurate than gradient-boosted trees. Data qualities affect performance [32, 33]. In 1995, Tin Kam Ho [29] developed the first algorithm for random decision forests using the random subspace method [30]. Ho's formulation provides a way to put Eugene Kleinberg's "stochastic discrimination" approach to classification into practice [34, 35, 37]. Leo Breiman and Adele Cutler [37] created an expansion of the algorithm, and in 2006, they registered [39] "Random Forests" as a trademark (as of 2019 it is owned by Minitab, Inc.) [40]. In order to create a collection of decision trees with controlled variance, the extension combines Breiman's "bagging" concept and random selection of features, which were first suggested by Ho [29] and later independently by Amit and Geman [41].

Businesses typically employ random forests as "blackbox" models because they produce accurate predictions across a variety of data sets with minimal preparation.

Yoav Freund and Robert Schapire, who shared the 2003 Gödel Prize for their work, created the statistical classification meta-algorithm known as AdaBoost, short for Adaptive Boosting. Its performance can be enhanced by combining it with a variety of other learning methods. A weighted sum that represents the final output of the boosted classifier is created from the output of the other learning algorithms (also known as "weak learners"). Because future weak learners are adjusted in favor of those instances that prior classifiers misclassified, AdaBoost is adaptive. It may be less prone to the

overfitting issue than other learning algorithms in particular situations. The final model may be shown to converge to a strong learner even if the individual learners are weak as long as their performance is somewhat better than random guessing [42].

CatBoost is a technique for decision trees that uses gradient boosting. It is the replacement for the MatrixNet algorithm, which was created by Yandex researchers and engineers and is widely used within the company for task rating, forecasting, and recommendation-making. It is ubiquitous and adaptable, and it can be used in a number of contexts and situations [43].

3.4 Data Acquisition and Processing in ML

As machine learning is dependent on the data supplied to the system, the initial phase in the design is data collecting. This entails collecting data, compiling and separating case scenarios depending on particular characteristics of the decision-making cycle, and sending the data to the processing unit for additional categorization. This level is sometimes referred to as the data pre-processing stage. The data model anticipates dependable, quick, and elastic data that may be discrete or continuous. The data is subsequently transferred to stream processing systems (for continuous data) and batch data warehouses (for discrete data) prior to being sent to the data modeling or processing phases [28].

The data is sent to the data processing layer from the data acquisition layer. It undergoes sophisticated integration and processing, which includes data normalization, data cleansing, transformation, and encoding. The processing of data also depends on the type of learning employed. For instance, if supervised learning is employed, the data must be separated into many phases of sample data required for training the system; the resulting data is referred to as training sample data or training data. In addition, the data processing is dependent on the type of processing required and may involve options ranging from action on continuous data, which requires the use of specific function-based architecture, such as lambda architecture, to action on discrete data, which may require memory-bound processing. The data processing layer specifies whether data in transit or at rest will undergo memory processing [28].

This layer of architecture entails the selection of several algorithms that could adjust the system to address the challenge for which learning is being designed. These algorithms are evolving or inheriting from a collection of libraries. The algorithms are employed to model the data appropriately; this prepares the system for execution [28].

In this phase of ML, experiments are conducted, tests are conducted, and adjustments are made. The overarching objective is to optimize the algorithm in order to obtain the necessary machine output and maximize system performance. The output of the phase is an improved solution that can provide the computer with the necessary facts for making judgments [28]. Like any other software output, machine learning outputs must be operationalized or routed for further exploratory processing. The output is a nondeterministic inquiry that must be further integrated into the decision-making system. It is recommended to move the output of machine learning (ML) directly into production to enable the machine to make decisions based on the result and avoid reliance on extra exploratory procedures [28].

CHAPTER 4

The Proposed CCFD

4.1 Overview

This chapter refers to the feature importance and classification algorithms required for the recognition of credit card frauds. The correctness of classification is calculated using several metrics, including classification accuracy, precision, recall, and area under the receiver characteristics curve (AUC-ROC).

4.2 Experimental setup

In this chapter, a discussion is provided about the software and hardware required to complete the project and the environment setup for the object detection model. This project made use of several software libraries, packages and programs to utilize ML. Python was the choice of programming language. Anaconda ide consisting of Jupiter notebook or colab (online) is used to implement the idea easily. The following were considered in the experimental setup:

- 1) Prerequisite tools:
 - Python 3.6 (or above)
- 2) Install Anaconda (python 3.6 or 3.5)
 - If python is not installed, using Anaconda is recommended (python 3.6 64bit)
 - Go to <https://www.anaconda.com/download/>
 - Choose python 3.6 (64bit) and install
 - Check the "add python to the PATH" option during the install

- 3) Install the other necessary packages and libraries by issuing the following commands:

```
C:\> pip install jupyter
```

```
C:\> pip install numphy
```

```
C:\> pip install matplotlib
```

```
C:\> pip install pandas
```

```
C:\> pip install catboost
```

```
C:\> pip install AdaBoost
```

4.3 Explanation of the Dataset

This section describes the datasets for CCFD considered in this research. It can be noted that the publicly available datasets do not disclose detailed information or even the meaning of some of the columns/attributes/features for security purposes. Usually, in banks, a number of information is used to detect credit card fraud. Some of these columns/attributes are: Last Payment Date, Guarantor Amount, Monthly Installment Amount, Minimum payment, Interest Charge, Grace Period, Credit limit, Outstanding Amount/Balance, Sanction Limit, Remaining Amount, Overdue Amount, Number of Overdue, Late Payment Fee, Date of next installment, Currency Name, Borrower's Income, and Borrower Income Source. In this project, a dataset is considered that consists of transactions performed with credit cards by European cardholders in September 2013 [10]. There are 492 fraudulent transactions, and the total number of transactions is 284,807. This dataset has 31 columns where Time, Class, and Amount are three columns and V1, V2, V3,,V28 are the 28 columns. The definitions of V1 to V28 are not disclosed in the dataset. Moreover, it is a highly imbalanced dataset. About 0.172% of all the transactions belong to the positive class (i.e., fraud class). Numerical input variables are presented here due to the principal component analysis

(PCA) transformation. Because of confidential reasons, the unique features and more background data about the dataset have not been offered [15]. V1, V2, V3,..., V28 are PCA applied features and the rest i.e., 'time', 'amount' and 'class' are non-PCA applied features. In this case, the 'Class' is the response variable having a value of 1 for the case of a fraud and a value of 0 in a normal case. Due to the massive amount of imbalance dataset, the classification algorithms need to avoid overfitting the classes.

Fig. 4.1 illustrates the time density plot of credit card transactions. There are two types of classes. These are: "Fraud" and "Not Fraud". "Fraud transactions" and "Not Fraud i.e., real transactions" classes can also be defined as Class 1 and Class 0, respectively. The distribution of these transactions is distributed uniformly. This distribution includes the fact that the number of transactions is very low late at night.

Table 4.1 describes the statistical report of the classes.

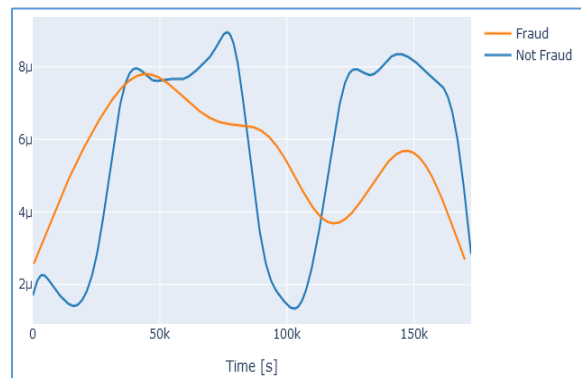


Figure 4.1: Time density plot

Table 4.1: Statistical Analysis of Classes

Class	count	mean	Standard Deviation	Minimum Value	25%	50%	75%	Maximum Value
0	284315	88.291	250.105	0.000	5.650	22.000	77.050	25691.160
1	492	122.211	256.683	0.000	1.000	9.250	105.890	2125.870

4.4 Methodology

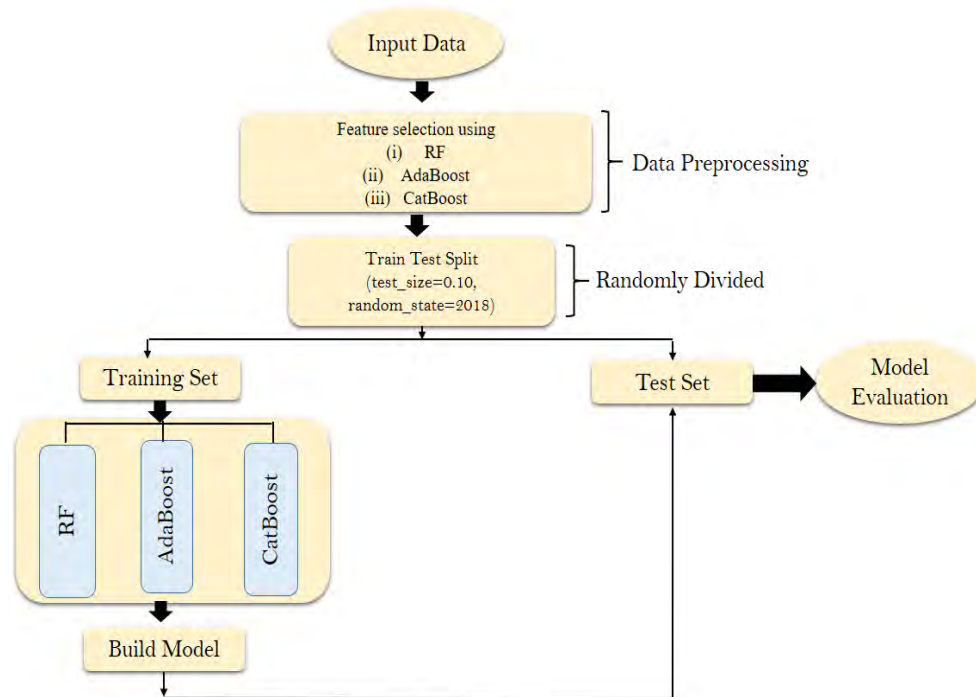


Figure 4.2: Work flow diagram of the methodology

Fig. 4.2 depicts a flow chart of the proposed work. It can be seen that feature importance is calculated in the pre-processing stage. Next, the *train_test_split()* function is conducted to split the dataset such as testing and training data samples where the training data size is 90%, and the size of testing data is 10% of the full dataset. Next, three classification algorithms, including RF, AdaBoost, and CatBoost are implemented for evaluating the dataset. After data-processing, the important features are considered for classification using RF, AdaBoost and CatBoost respectively. Next, we have evaluated the dataset according to various performance metrics reported in the literature [16-19].

4.5 Performance of RF Classifier

For training, the selected classification model is executed using the training dataset. After that, a validation set is used for validation. Gini is used as a validation criterion. Note that random forest calculates the relevance of each feature using Gini importance or mean reduction in impurity (MDI). Gini significance is often referred to as the complete decline in node impurity. This is the decrease in model fit or precision when a variable is removed. The variable is more significant the larger the drop. Here, the mean decrease is a key variable selection criteria. The number of estimators is 100 and the number of parallel jobs is 4. From Fig. 4.3, it can be shown that very important or priority-based features are V17, V12, V14, V16, V11, V10. Fig. 4.4 and Fig. 4.5 illustrate the confusion matrix. The ROC-AUC score obtained with RF Classifier is 85.29%.

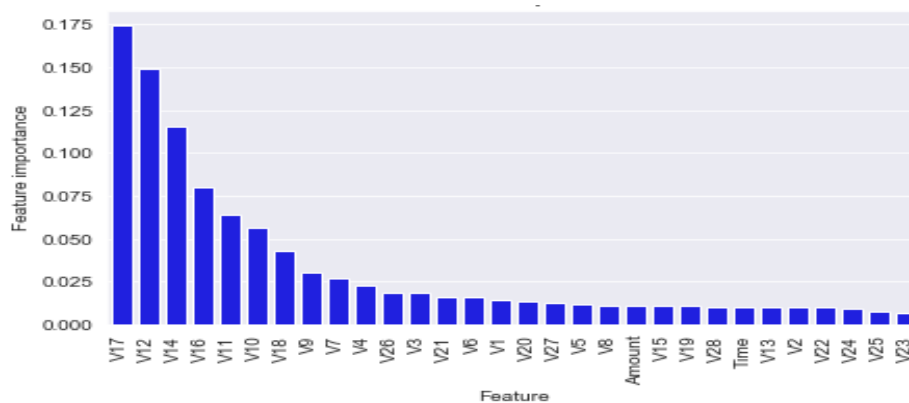


Figure 4.3: Important feature selection using RF

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, True Positive Rate (TPR) and False Positive Rate (FPR). The term TPR is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

On the other hand, the term FPR is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

	Positive (1)	Negative (0)
Positive(1)	True Positive (TP)	False Positive (FP)
Negative (0)	False Negative (FN)	True Negative (TN)

Figure 4.4: Basics of a confusion matrix

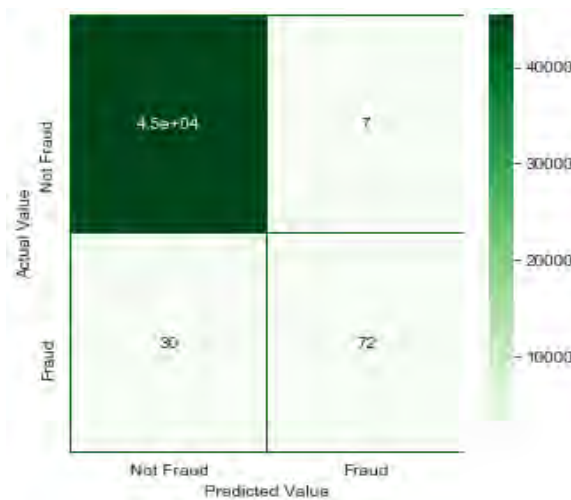


Figure 4.5: Confusion matrix using RF

4.6 Performance of AdaBoost

This section shows the results obtained for the case of AdaBoost. Experiment results show that the ROC-AUC score obtained with AdaBoost Classifier is 83.32%. From Fig. 4.6, it can be seen that the most important or priority-based features are V12, V14, V4,

V10, V18, V6. Fig. 4.7 illustrates the confusion matrix. "SAMME.R" algorithm is used to obtain very low testing errors with only a few boosting iterations. In this case, the learning rate is 0.8.

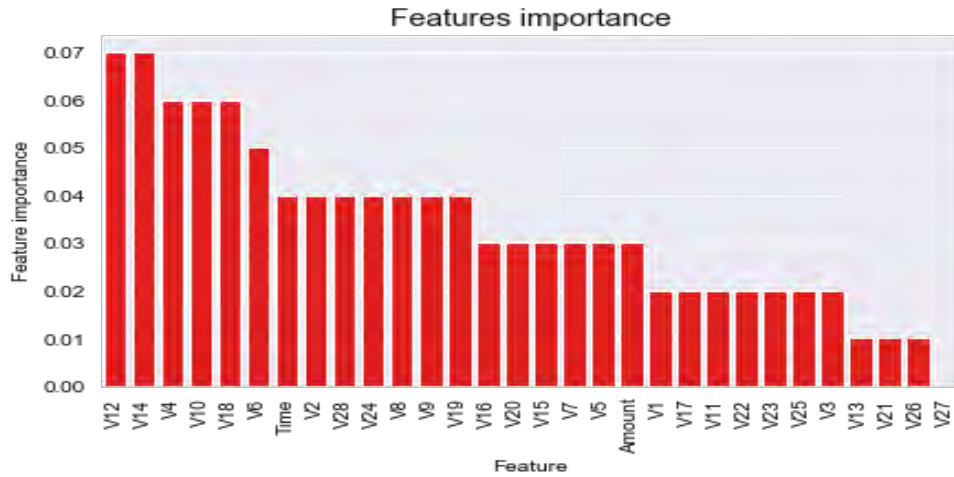


Figure 4.6: Important feature selection using AdaBoost

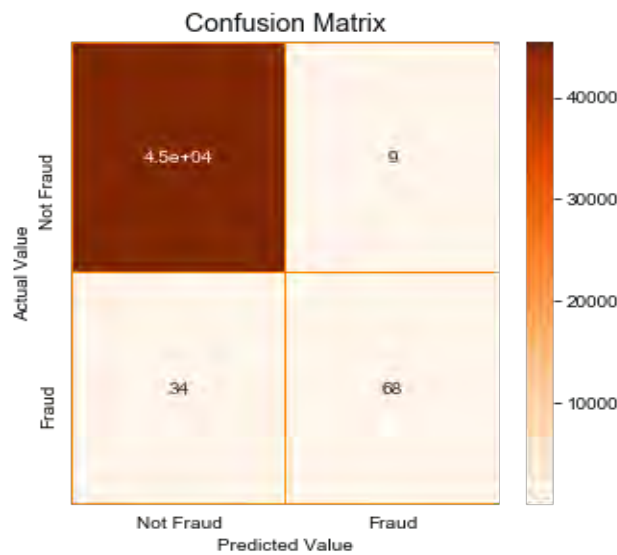


Figure 4.7: Confusion matrix using AdaBoost

4.7 Performance of CatBoost

This section describes the results obtained for CatBoost algorithm. The number of iteration rate is 500 and learning rate is 0.02 when bagging temperature is set to 0.2 at the time of executing the CatBoost classifier. The ROC-AUC score obtained with CatBoost Classifier is 85.77%. From Fig. 4.8, it can be seen that very important or

priority-based features are V26, Time, V8, V4, V17, V6, amount, V24. Figure 4.9 illustrates the confusion matrix for the case of CatBoost.

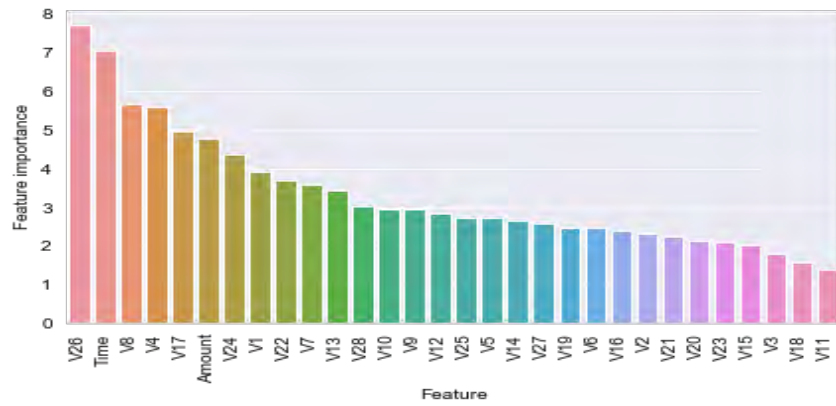


Figure 4.8: Important feature selection using CatBoost



Figure 4.9: Confusion matrix using CatBoost

Table 4.2(a). Performance evaluation of classifiers

Classifier	TP	FP	FN	TN	TPR	TNR
RF	45000	7	30	72	99.933%	91.139%
AdaBoost	45000	9	34	68	99.924%	88.312%
CatBoost	45000	6	29	73	99.936%	92.405%

Table 4.2(b). Performance evaluation of classifiers

Classifier	PPV/ Precision	Sensitivity/ Recall	NPV	FDR	FOR	FNR	Accuracy
RF	99.98%	99.93%	70.58%	0.02%	29.41%	0.07%	99.92%
AdaBoost	99.98%	99.92%	66.66%	0.02%	33.33%	0.08%	99.91%
CatBoost	99.98%	99.94%	71.56%	0.01%	28.43%	0.06%	99.92%

Table 4.2 illustrates the comparative analysis of the different classification algorithm for CCFD where the comparison is done in terms of TP, FP, FN, TN, TPR and TNR. From Table 4.2(a), the highest value of TPR is 99.936% achieved by using CatBoost classifier. The TPR rate of RF is closer to CatBoost classifier. The highest value of TNR is also for CatBoost being 99.36%. Table 4.2(b) shows the algorithms' precision, recall, accuracy, and false detection rate (FDR) values. It can be seen that an FDR value 0.02% is achieved by RF and AdaBoost classifier, while CatBoost has an FDR of 0.01%. Table 4.2b) also indicates that the three classifiers have almost similar precision, recall and classification accuracy levels. All the of these have a precision value of 99.98%. Moreover, RF, Adaboost and CatBoost have recall values of 99.93%, 99.92% and 99.94%, respectively. Furthermore, RF and CatBoost exhibit the highest value of 99.92%, whereas AdaBoost has 99.91% accuracy level.

A comparative study is reported in [20], where naïve Bayes provides higher accuracy compared to other algorithms, including logistic regression and k-nearest neighbour. The accuracy of naïve Bayes reported in [20] is 97.92% which is lower than the 99.92% accuracy obtained by the proposed RF and CatBoost classifiers in our work. Another study [21] considers majority voting and AdaBoost algorithms where the highest accuracy is close to the accuracy values of our study. However, the work in [21] uses a different dataset and hence cannot be directly compared with the results of our work.

CHAPTER 5

5.1 Conclusion

This study examines the use of a variety of ML classifiers to the problem of predicting the authenticity of credit card transactions. In the first step of the process, the data of the datasets are examined to look for any imbalanced data. The next step is to identify the most essential features. After that, three distinct kinds of classifiers are put into use. In order to fulfill this objective, the holdout technique partitions the dataset into testing and training sample sets. In this particular instance, ninety percent of the data samples are collected for the purpose of training, while the remaining ten percent are utilized for examination. For the CatBoost algorithm, the most important features are V26, time, V8, V4, V17, V6, amount, and V24. Results show that RF and CatBoost achieve an accuracy value of 99.92%, while AdaBoost exhibits an accuracy value of 99.91% for CCFD. One of the difficulties that the CCFD faces is that there are not too many real datasets available because financial firms do not disclose the transactions they make. The effectiveness of the algorithms, on the other hand, will fluctuate depending on the changes made to the dataset. In order to determine which algorithms produce the best results, it is necessary to apply the classification and feature significance algorithms to a variety of distinct real-world datasets.

5.2 Future work

Although ML approaches can improve CCFD accuracy, there are still some challenges that need to be addressed in the future. To avoid data imbalance, we need huge datasets to train the model. Real-time datasets can provide additional data, but privacy is a concern. In the future, there will be a need for large real-time datasets to train the model

while protecting privacy. Federated learning may help improve ML model fraud detection. Strategies should be developed to enable financial institutions and banks to use real-time datasets by collaborating to establish an effective CCFD system. There are some deployment restrictions. Banks and financial institutions have tight laws and regulations. Adapting ML methods will be difficult because banks and financial institutions have their own restrictions and rely on internal resources rather than a centralized approach. It will be necessary to resolve these concerns in the near future.

It is expected that the ongoing research on the development of ML approaches can change CCFD using real-life datasets, giving the banking and financial industry a new horizon.

References

1. S. Kotsiantis, D. Kanellopoulos, P. Pintelas (2006). Handling imbalanced datasets: A review. *International Transactions on Computer Science and Engineering*.
2. Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, 165, 631-641.
3. R.J. Bolton, D.J. Hand (2001). Unsupervised profiling methods for fraud detection. In *Conference on credit scoring and credit control*, Edinburgh.
4. <https://www.wsj.com/articles/no-headline-available-1387459173?tesla=y> [Last access 01 November 2022]
5. McCurry, Justin (23 May 2016). "100 thieves steal \$13m in three hours from cash machines across Japan". *The Guardian*. Retrieved 23 May 2016.
6. D. Kibler, D.W. Aha, M. Albert (1989). Instance-based prediction of real-valued attributes. *Computational Intelligence*, Vol(5); 51-57.
7. Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, 165, 631–641. doi:10.1016/j.procs.2020.01.057
8. P.K. Chan, W. Fan, A.L. Prodromidis, S.J. Stolfo (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems*, pp 67–74.
9. G. Potamitis (2013). Design and Implementation of a Fraud Detection Expert System using Ontology-Based Techniques. A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences.
10. <https://data.world/raghu543/credit-card-fraud-data> [Last access 01 November 2022]
11. Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)* (pp. 152-156). IEEE..
12. Jha.Sanjeev, G. Montserrat, J.C.Westland (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert system with application*, 39: 12650-12657

13. Priscilla C.V., Prabha D.P. (2020) Credit Card Fraud Detection: A Systematic Review. In: Jain L., Peng SL., Alhadidi B., Pal S. (eds) Intelligent Computing Paradigm and Cutting-edge Technologies. ICICCT 2019. Learning and Analytics in Intelligent Systems, vol 9. Springer, Cham.
14. Masoumeh Zareapoora, Pourya Shamsolmoal, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier", *Procedia Computer Science*, Volume 48, 2015, Pages 679-685.
15. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In *Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE, 2015.
16. Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. *IEEE Transactions on dependable and secure computing*, 5(1), 37-48..
17. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*, 29(8), 3784-3797..
18. Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J. M. (2009). Association rules applied to credit card fraud detection. *Expert systems with applications*, 36(2), 3630-3640..
19. Bharati, S., Podder, P., & Mondal, M. R. H. (2020, June). Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms. In *2020 IEEE Region 10 Symposium (TENSYP)* (pp. 1486-1489). IEEE.
20. Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-9). IEEE.
21. Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. *IEEE access*, 6, 14277-14284.
22. Guidelines on Credit Card Operations of Banks: https://www.bb.org.bd/aboutus/draftguinotification/guideline/cc_operation_bank.pdf [Accessed: 2021-02-08].
23. Top Machine Learning Algorithms: https://medium.com/@webadmin_46735/top-machine-learning-algorithms-for-predictions-a-short-overview-5ed1ff6942ff Accessed: 2021-02-08.
24. Early History of Machine Learning: <https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html>

25. ^ Machine learning and pattern recognition "can be viewed as two facets of the same field." [3]:vii
26. ^ Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". *Computing Science and Statistics*. 29 (1): 3–9.
27. https://en.wikipedia.org/wiki/Machine_learning Accessed: 2021-02-08.
28. <https://www.educba.com/machine-learning-architecture/> Accessed: 2021-02-08.
29. Ho, Tin Kam (1995). Random Decision Forests (PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995*. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
30. Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20 (8): 832–844. doi:10.1109/34.709601.
31. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning* (2nd ed.). Springer. ISBN 0-387-95284-5.
32. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". *Journal of Transportation Engineering, Part B: Pavements*. 146 (2): 04020022. doi:10.1061/JPEODX.0000175.
33. Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021-02-01). "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling". *Journal of Infrastructure Systems*. 27 (2): 04021005. doi:10.1061/(ASCE)IS.1943-555X.0000602. ISSN 1076-0342.
34. Kleinberg E (1990). "Stochastic Discrimination" (PDF). *Annals of Mathematics and Artificial Intelligence*. 1 (1–4): 207–239. CiteSeerX 10.1.1.25.6750. doi:10.1007/BF01531079. Archived from the original (PDF) on 2018-01-18.
35. Kleinberg E (1996). "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition". *Annals of Statistics*. 24 (6): 2319–2349. doi:10.1214/aos/1032181157. MR 1425956.
36. Kleinberg E (2000). "On the Algorithmic Implementation of Stochastic Discrimination" (PDF). *IEEE Transactions on PAMI*. 22 (5): 473–490. CiteSeerX 10.1.1.33.4131. doi:10.1109/34.857004. Archived from the original (PDF) on 2018-01-18.

37. Breiman L (2001). "Random Forests". *Machine Learning*. 45 (1): 5–32. doi:10.1023/A:1010933404324.
38. Liaw A (16 October 2012). "Documentation for R package randomForest" (PDF). Retrieved 15 March 2013.
39. U.S. trademark registration number 3185828, registered 2006/12/19.
40. "RANDOM FORESTS Trademark of Health Care Productivity, Inc. - Registration Number 3185828 - Serial Number 78642027 :: Justia Trademarks".
41. Amit Y, Geman D (1997). "Shape quantization and recognition with randomized trees" (PDF). *Neural Computation*. 9 (7): 1545-1588. CiteSeerX 10.1.1.57.6069. doi:10.1162/neco.1997.9.7.1545.
42. <https://en.wikipedia.org/wiki/AdaBoost> [Last access 01 November 2022]
43. <https://affine.ai/catboost-a-new-game-of-machine-learning/> [Last access 01 November 2022]