UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Project and Professionalism

(6CS020)

A2: Project Report

Credit Card Fraud Detection

Student Id	: 1928584
Student Name	: Sunil Ghimire
Group	: C3G4
Supervisor	: Sachin Kafle
Cohort	: 3
Submitted on	: 12th June 2020

Declaration Sheet

Presented in partial fulfillment of the assessment requirements for the above award.

This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgments, references, and/or bibliographies cited in the work. I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the University a perpetual royalty-free license to do all or any of those things referred to in section 16(I) of the Copyright Designs and Patents Act 1988. (viz. to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make an adaptation of the work).

Student Name: Sunil Ghimire

Student ID Number: 1928584

Sunil Ghimire

Signature: ……………………………… Date: 12th June 2020

Acknowledgment

Technology made revolutionary changes to our lives.

With the regular support, cooperation, and coordination of **Mr. Sachin Kafle** of Herald College Kathmandu has provided me with the platform to enhance my knowledge of the project and its application. All the teaching and non-teaching staff of my college whose immense effort and gratitude made the project possible. They were who provided me with a friendly environment, valuable instructions, friendly behavior, and guidance and provide prompt comments and rectification necessary before finalization the report. I would like to thank Info Developer Team for providing me with the necessary information when needed.

I owe a great deal to my colleagues who directly and indirectly help us during the project and during making this report. I want to express my heartfelt and sincere gratitude towards all of you for understand and constant support in making this project successful.

Thank You!

Sunil Ghimire

Abstract

One of the big legal problems in the credit card business is fraud. The key goals of this research are, firstly to recognize the different forms of fraudulent credit cards, secondly, to explore alternative methods utilized in fraud detection. The sub-aim is to evaluate, present, and examine recent results in the identification of credit card fraud. The article sets out terms common in fraud involving credit cards and highlighting figures and key statistics in this field. Various measures such as Logistic Regression, Random Forest, Autoencoder and SMOTE can be taken and enforced based on the type of fraud faced by the credit card industry or financial institutions. In terms of cost savings and efficiency, the proposals made in this report are likely to have beneficial attributes. The importance of applying these techniques examined here in minimizing credit card fraud. Yet when legitimate credit card users are misclassified as fraudulent there are still ethical issues.

Keywords: Logistic Regression, Random Forest Classifier, Autoencoder, SMOTE

Table of Contents

Table of figure

Table

1. Introduction

1.1. General Introduction

When people in the old days decided to get a home, they needed to save the funds for it and build it on their own or employ someone else to do it. Without bank-like financial intermediaries' transactions like this were done in available funds. So, with the changing fields of technology and communication, financial institutions have become a hot industry among entrepreneurs which makes it easier for people with money to contact people who want to borrow money. So, financial institutions are one of the most important factors of the financial system of any country which plays a crucial role in assessing the efficiency and effectiveness of the financial system (Ghazi, 2019). A credit card is a part of a financial product that is issued by banks to make purchases on credit. In other words, a credit card also known as a charge card is defined as a plastic card having a magnetic stripe on the side of the card which contains the details of the critical cardholder and should therefore not be abused. The issuing bank assigns a different credit limit for each card where people can shop in stores, malls, and even online were using a card up to limits issued by the card provider. Also, the credit cards can be used by individuals every month for any amount up to the limit. When people use their card, the issuing company charges to the retailers on their behalf. So this the reason, people buy goods and services without having to pay directly out of their wallets (Salman & Salman, 2015).

Credit card frauds happen in different ways and have been a concern for years all over the world. There are various forms of credit card systems and programs but the illegal usage of a missing or stolen card is one of the simplest processes. Among the missing or stolen card, Account Takeover, Internet Fraud, Non-Receipt Fraud, and counterfeit credit card fraud are the major types involved in credit card fraud which not only affect the victims but also credit card companies and retailers are struck by the impact of this fraud (Barker & Sheridon, 2008). The amount of fraudulent behavior is increasing rapidly, with individuals and organizations at high risk. So, detection of fraud involves identifying a fraud that is either about to occur or after it has happened which can never be fully prevented. According to Experian, "Customers making an application want the best service but also want to be protected against fraudsters and identity theft. You have to balance protection while giving your genuine customers the best decision in the shortest possible time." (Al-Jumeily, et al., 2015)

Fraud is an illegal method of obtaining commodities and money and credit card fraud is a growing issue nowadays regarding payment cards as an illegal source of funds in transactions. The main goal of such illegal activity may be to get goods from an account without paying or obtaining an illegitimate fund. In the fraud detection system of the modern world, the investigator is not able to verify all the transactions. Also verifying all the transactions is a time consuming and costly process. So, identifying this kind of fraud is troublesome and may endanger companies and business organizations (Shirgave, et al., 2019).

1.2. Current Scenario

The existing traditional finance system is quite challenging because the traditional finance sector was focused on taking deposits for a fee from the depositors and lending this money at a price (interest) to interested borrowers. Nowadays, most account holders depend on plastic currency to withdraw money, debit card, instead of standing in lines to withdraw cash. Also, most of the depositors may not own a cheque book to avoid unnecessary restrictions that come with owning cheque books such as added fees and obligatory balance criteria. In a dynamic economy, Nepal Rastra Bank (NRM) plays a key role in directing network growth in a competitive environment, including structures such as payment systems. The first successful e-payment network device, Electronic Cheque Clearing System (ECCS), has been introduced and operational since November 2011 along with the efforts of banks and central bank that enables the clearing of the cheque of the same day irrespective of the location of the Banks and their branches (Giri, 2013).

1.3. Proposed System

Credit cards are currently one of the leading online and offline payment types and the use of credit cards has risen exponentially which means there is a high chance of misleading transactions. The making of detection is not possible, none of the standard programs monitor for credit card fraud. For example, if we want to accept the 'ABC' card as a real card or a fraud. There is no obligation to distinguish the actions of the card. And there is also no guarantee that the card is valid or counterfeit. Besides this, there is no fixed software or system that can provide a genuine, fraud log of transactions. People are more curious to figure out whether thirty people used their card or those used by other third-party apps. However, the main problem is the avoidance of potential misuse of credit cards.

Thus, Information technology is rapidly increasing with respect to automated systems. In such systems, people utilize computer-based expert systems to analyze and handle real-life problems such as Online Payment systems. The proposed Credit Card Fraud Detection (CCFD) is a solution for fraudulent identification that tracks the time and amount of money in the everyday transactions to determine if the credit card is accepted or not where dataset for this project of credit was obtained from Kaggle which includes input variables that arise from a PCA transformation. Unfortunately, the dataset cannot include the original features and further detailed details regarding the data because of confidentiality concerns. Features V1, V2 ... V28 are the key components obtained with PCA, Time and Amount are only the features that have not been transformed with PCA. The 'Time' features include the seconds between every transaction and the first transaction inside the dataset. The 'Amount' feature is the transaction amount which can be used for the example-dependent cost-sensitive learning. The feature 'Class' is the target variable which takes value 1 as a fraud transaction and value 0 for normal transactions. CCFD uses a machine-learning algorithm like Logistic Regression and Random Forest and deep learning algorithm like Autoencoder which finds a structured pattern of normal credit cards and fraudulent credit cards. At last, Tkinter is used which is the part of Python standard library that provides an object-oriented interface onto TK/TCL Also, Tkinter helps to develop to create a cross-platform GUI for FYP without more dependencies.

1.4. Project Scope

Machine learning and deep learning can hardly be claimed to be one of the best researches and growth opportunities. It is growing to uncharted height, innovating revolutionary technology dramatically changes computer science trends and studying across the globe. So, with time passing rapidly, any more development would be dampened with any second wasted. For to operate on these large amounts of data that need technologies that can speed the latest techniques and models. ML's and DL's techniques and models will simplify and visualize results, and reliably assess profound observations and trends from the training collection. Machine Learning and Deep

Learning is a new-generation technology that allows the newest technologies to be built from various mechanisms. The proposed application considers Logistic Regression and Random Forest as machine learning algorithms where logistic regression is used to minimize the wrong prediction and random forest which is used for classifiers which is used to handle the missing values and maintain the accuracy of a large proportion of data. And deep learning algorithm i.e. Autoencoder is used to denoise the testing dataset in the prediction process and classify the sample dataset. Along with ML and DL model techniques like SMOTE and K-fold cross-validation in data-preprocessing.

1.5. Aims and Objectives

1.5.1. Aims

The main aim of this report is to gain the ability to research various machine learning and deep learning algorithms along with its wrong mechanisms based on fraud credit cards and gain knowledge about the techniques which make complete algorithms.

1.5.2. Objectives

The objectives of this report are as follows:

- Getting information by proper research
- Understanding algorithms and its working mechanism
- Able to understand different algorithm based on CCFD
- Detect precision, recall, f1-score based on algorithm
- Understand the technique to handle imbalanced dataset
- Able to visualize the graph of dataset
- Create the report based on the project

1.6. Academic Questions

There are certain questions arose during the planning of the proposed algorithms which are listed below:

- What sort of problem is this project going to solve?
- How actual is fraud detected?
- What are the challenges involved in developing an algorithm to detect the fraud card?
- Are there any similar projects?

- Is the proposed research feasible to handle an imbalanced dataset?

1.7. Report Structure

The diagram for the report structure is shown below:

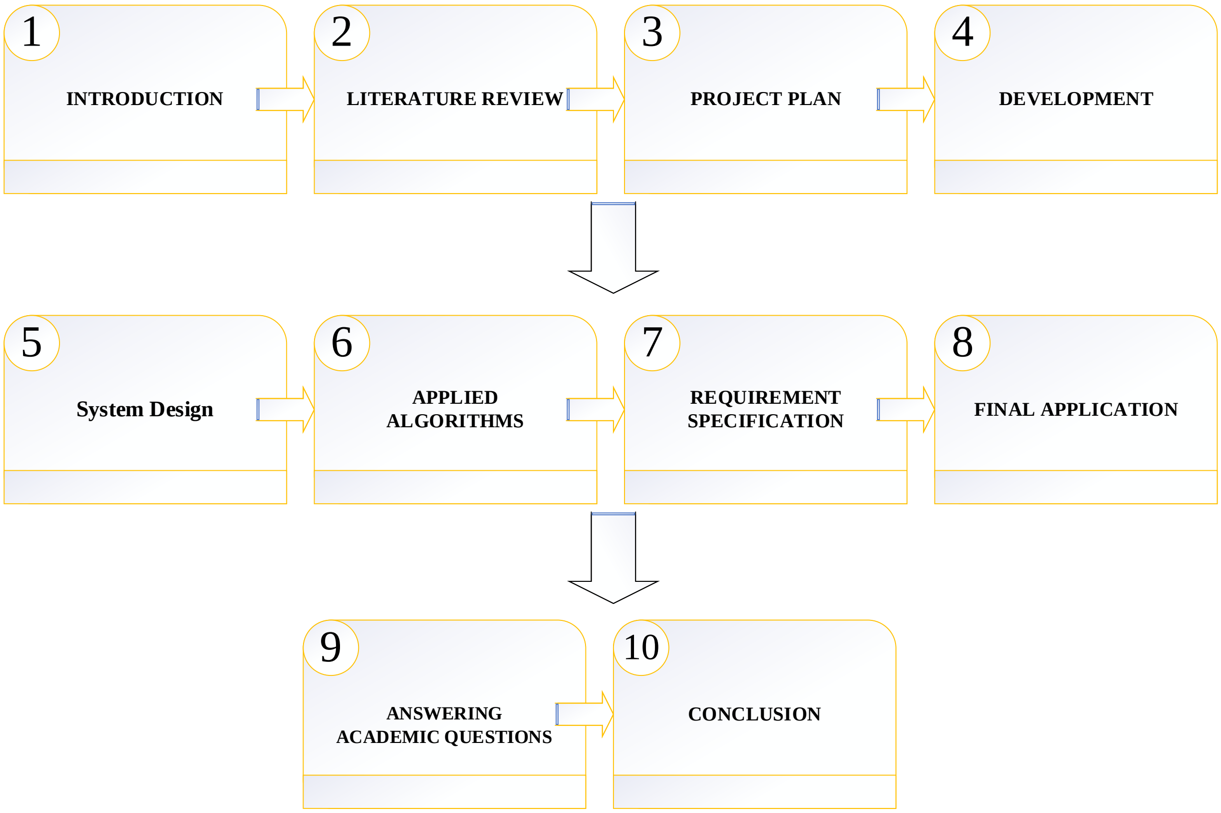


Figure 1: Structure of the report

The brief description of the structure of this report is as follow:

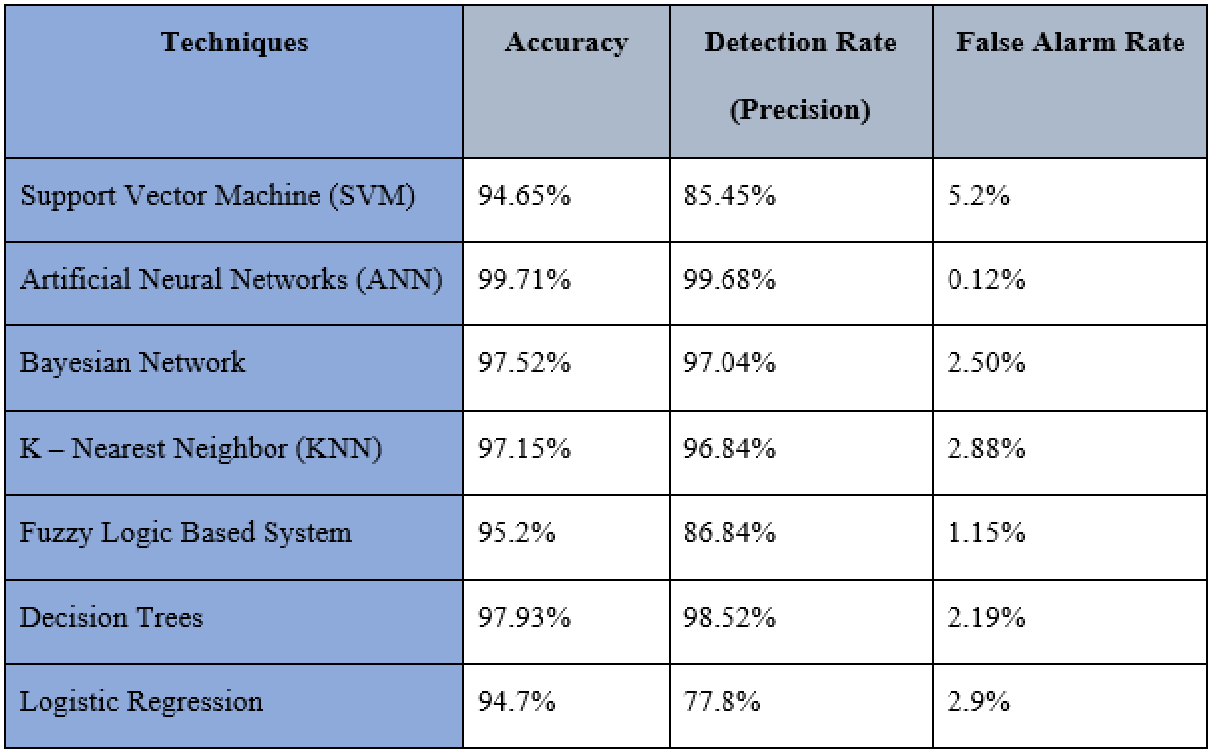
- Introduction: This section includes general information about the topic along with aims and objectives.
- Literature review: This section includes all the background research regarding the similar systems.
- Project plan: This section includes the detail plan to complete the project
- System Design: This section includes working flow to detect fraud card.
- Applied algorithms: This section includes the machine and deep algorithms used during development.

- Requirement specifications: This section includes functional and non-functional requirements of the project.
- Final application: This section includes the final execution of all machine learning and deep learning algorithms with confusion matrix and classification report.
- Answering academic questions: This section includes the answer to the academic questions.
- Conclusion: This section includes the conclusion to the entire project and future escalations.

2. Literature Review

There are several published research papers related to detecting fraud credit card. A paper was published on "A Comparative Analysis of Various Credit Card Fraud Detection Techniques". Credit-card fraud has cost merchants and banks trillions of dollars worldwide. Even after various strategies for preventing fraud, fraudsters are actively finding different forms and techniques to commit fraud. Thus, in order to stop fraud, this paper proposes an effective system for detecting fraud, which not only identifies the fraud card but also identifies it before it occurs and adopts new methods for monitoring frauds cards in a specific manner also discussed numerous possible techniques for fraud detection such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Bayesian Network, K-Nearest Neighbor (KNN), Hidden Markov Model, Fuzzy Logic Based System and Decision Tree and such methods are used to build an effective, accurate and quick fraud detection program capable of identifying not just internet scams, such as phishing and site replication, but even the manipulation of credit cards themselves, i.e. the usage of a tempered credit card to cause a notice. The main drawbacks of this system are that they are non-guaranteed, they give the best results for dataset and poor results for other types of dataset. The algorithm ANN and Naive Bayesian gives high detection rates and high accuracy. ANN considers recent incoming transactions as fraud or genuine transactions based on previous records and Naive Bayesian consists of edges and nodes where nodes represent random variables and edges between nodes denote associations of other random variables and their probabilistic distribution, which are used to measure the minimum and maximum likelihood of fraud and legal transactions. For new incoming transactions if the probability of legal transactions is less than fraud transactions it is known as fraud transactions. Some other algorithms like KNN and SVM give excellent results for small datasets. The KNN model calculates the prevalent class for every new transaction and marks the transactions as belonging to the prevalent class. The SVM model generates a hyperplane that studies the activity of fraud and legitimate transactions and then classifies new transactions according to which class it belongs. For better results sampled and preprocessed data SVM and decision tree is used whereas logistic regression and fuzzy systems have greater precision for raw non-sampled results. Decision Tree used for classification and prediction contains internal nodes that represent an attribute test and each branch denotes the an outcome product and each leaf node has a class label that uses a technique called depth-first greedy approach or breadth-first greedy approach and stops until all the transactions are allocated to a certain class. Logistic regression is used for clustering, with the purpose of calculating the values of the parameter coefficients using the sigmoid function, and it tests the values of its parameters while the transaction is occurring and decides how transactions will continue. And the last proposed fuzzy logic is used for continuous data when there is an absence of discrete truth value in the dataset where Fuzzification, Rule-Based and Defuzzification are the three important components. Fuzzification is used for incoming transactions according to the numerical interest correlated with transaction small, medium, or large categories. Rule-Based regulations are expanded based on consumer behavior and last component Defuzzification, it is not permitted to proceed if a transaction does not conform with the predefined collection of laws. This is halted automatically, and then cross-checked with the consumer whether the permission to continue or to be terminated would be provided(Jain, et al., 2019).

Algorithms used to detect fraud card along with Accuracy, Detection Rate and False Alarm Rate is shown below:



Techniques	Accuracy	Detection Rate (Precision)	False Alarm Rate
Support Vector Machine (SVM)	94.65%	85.45%	5.2%
Artificial Neural Networks (ANN)	99.71%	99.68%	0.12%
Bayesian Network	97.52%	97.04%	2.50%
K – Nearest Neighbor (KNN)	97.15%	96.84%	2.88%
Fuzzy Logic Based System	95.2%	86.84%	1.15%
Decision Trees	97.93%	98.52%	2.19%
Logistic Regression	94.7%	77.8%	2.9%

Table 1: Comparison of different machine learning techniques (Jain, et al., 2019)

From the above table, ANN and Bayesian Network give higher accuracy, KNN, SVM and decision tree provides a medium level of accuracy and fuzzy logic-based system and logistic regression provides a low level of accuracy as compared with others.

Another paper was published on “Credit card fraud and detection techniques: A review”. The key goal of this paper is to recognize the different forms of credit card fraud and to explore the alternative methods used in detecting fraud. The sub-aim of this report is to present, evaluate and examine recent results in the identification of credit card fraud. This article defines specific terminology for fraud credit card, outlining the important facts and figures and minimization of credit card fraud faced by financial institutions or credit card companies. Techniques to counter fraud cards are checked and details are given from European markets where fraud occurs when a borrower cheats or fools a lender promising him / her transactions, assuming the borrower’s credit card account will compensate for these transactions. Also, the paper describes different forms of fraud including bankruptcy fraud, currency fraud, burglary fraud, computer fraud, and behavioral fraud and proposes different techniques for minimization of credit card frauds which can be carried out by Decision Trees, Genetic Algorithms, Clustering Techniques, and neural networks. Algorithms have presented the best result to detect fraudulent credit cards. The results showed that data mining techniques can be enough to detect fraud credit cards. This paper investigated different statistical techniques used by different countries to detect fraud credit cards. However, the most used technique is a neural network which is technically an online fraud detection system based on neural classifiers (Delamaire, et al., 2009).

Study	Country	Method	Details
Aleskerov (1997)	Germany	Neural Network	Card-watch
Bently (2000)	UK	Genetic Programming	Logic rules and scoring process
Brause and his team (1999)	Germany	Data mining techniques and neural network	Data mining application combined probabilistic and neuro-adaptive approach
Bolton and Hand (2002)	UK	Clustering Techniques	Peer group analysis and break point analysis
Ghosh and Reilly (1994)	USA	Neural Network	FOS (Fraud Detection System)
Dorronsoro (1997)	Spain	Neural Network	Neural Classifier
Leonard (1995)	Canada	Expert System	Rule-based Expert System for fraud detection (fraud modelling)
Zaslavsky and Strinzkak (2006)	Ukraine	Neural Network	SOM, algorithm for detection of fraudulent operations in the payment system
Kokkinaki (1997)	Cyprus	Decision Tree	Similarity tree based on decision tree logic.
Chan and his team (1999)	USA	Algorithms	Suspect behavioral prediction

Ezawa and Norton (1996)	USA	Bayesian networks	Telecommunication industry
Kim and Kim (2002)	Korea	Neural Classifier	Improving detection efficiency and focusing on bias of the training sample as in skewed distribution. To reduce “mis-detections”.

Table 2: Investigating different statistical techniques in credit card fraud (Delamaire, et al., 2009)

Another research paper was published on “Credit Card Fraud Detection System Based on Machine Learning Techniques”. The paper proposes machine learning and data mining techniques like Naive Bayesian, Decision tree, and SVM used for the evaluation terms which is the initial phase because they require fewer assumptions and deliver higher analytical accuracy also known as “Standard model” to detect the fraudulent credit card. The machine learning tree-based ensemble model is especially popular for bagging and boosting where the hierarchical tree model is capable of modeling non-linear relationships which are usually used for classification and regression and big independent variables are expected to do a better performance. Therefore, the random forest is an assembly machine learning technique that utilizes bagging to use multiple trees as classifiers and the random forest incorporates information across all trees to reveal variable value after majority voting across all classifiers. The second phase introduces the proposed algorithm (hybrid methods) based on “AdaBoost and Gradient Boosting Machine ” known as Boosting methods which is another form of ensemble method to increase the accuracy and strongly relatedness of any given learning algorithm and closely related to the random forest that uses model efficacy to evaluated only use of the credit card collection which is publicly available that may be concerning credit risk, customer profit, stock prices and automated trading (Kumar, et al., 2019).

The below table shows the accuracy after using different model where the proposed algorithm has higher accuracy as compared with other

	KNN	**Random Tree**	**Proposed Algorithm**
Accuracy	0.9691	0.9432	0.9824
Sensitivity	0.8835	0	0.9767
Specificity	0.9711	0	0.9824
Limitations	Cannot detect the fraud at the time of transactions	No suitable for Randomness dataset	Not applied for non-linear data

Table 3: Comparison of algorithms (Kumar, et al., 2019)

Application offered by Jumio i.e. NetVerify is a web application to detect and prevent credit card fraud by using computer vision, biometric facial recognition, and machine learning where human reviews are needed to see recognized patterns. The web application also detects some effort to modify the ID, such as cropping a portion of the picture or identity. Consumers must provide valid identification card registration, identification checking and authentication of documentation where software can catch fake IDs. Companies add another feature i.e. biometric facial recognition device with eyeball monitoring and most minute face moments identifying fraudsters. The company verified mobile transactions including KYC and more than 120 million identities through the web where algorithms are used by computer vision to enable recognition of trends (Jesus, 2019).

3. Scope Identification

3.1. Fact-Finding Techniques

Fact-finding is a method gathering evidence and knowledge based on strategies that involve existing documents sampling, analysis, observation, questionnaires, interviews, prototyping, and joint requirements planning. Using successful fact-finding methods, system analyst constructs and implements existing programs. The compilation of necessary information is very critical for the implementation in the System Development Life Cycle (SDLC) because the system or software is unable to be utilized accurately and reliably without sufficient abstraction from the evidence. During the early stages of the SDLC, fact-finding strategies are implemented during the process of system analysis, configuration, and post-implementation evaluation. Facts used in any information of the system will be evaluated according to three measures (Tulasi, 2020):

a. **Data-facts**: Used to making useful information system

b. **Process-function**: To perform the objectives

c. **Interface-design**: Interacting with users.

Following are the fact-finding techniques:

3.1.1. Questionnaires

Questionnaires are the most popular and widely used method to collect knowledge and information which contains a variety of standard questions. The aim of the questionnaires is to collect detailed information regarding CCFD in a more efficient way. Questionnaires greatly serves to obtain insight on the operation of the program from experienced professionals and helps to gain practical guidance (tutorialspoint.com, 2020). Therefore, while doing CCFD projects also have a part to contribute to the practical operating scenario of the program and to get feedback about how to be successful in the method to produce meaningful outcomes via questionnaires. Questionnaire section is mentioned in the appendix section [Question answer session with AIHUB].

Questions to be asked are as follows:

a. How do you handle the imbalanced dataset?
b. How do you classify the feature and target class of the dataset?
c. On What basis you remove the null values from the dataset?
d. On what basis is the visualization allocated?
e. What is the curse of dimensionality and what are some ways to deal with it?

f. The dataset obtained from Kaggle contains only numerical input variables which are the result of PCA transformation. So, why is PCA needed in Machine Learning?
g. What are the types of data mining techniques that can detect the actual card and fraudulent card?
h. How does one choose which algorithm is best suitable for the dataset at hand?
i. How to apply machine learning in fraud detection?
j. What are the factors I must consider before comparing the performance of two-meta algorithms applied to a problem?
k. Why do we need a validation set and a test set?
l. On what basis is k-Fold cross-validation allocated?
m. What are some factors that explain the success and recent rise of machine learning and deep learning?

3.1.2. Observation

Observation is one of the most effective fact-find methods used to gather data and information in social research (Kawulich, 2012). The field research has been carried out in many specific observing types, three aspects of observing types are listed below:

a. Participant observation

Participant observation was an important research strategy for this project, involving a relatively unstructured and flexible combination of informal interview data and information being gathered (Kawulich, 2012). In this type of observation, particular sorts of activities like knowing about credit card, the working mechanism of the ATM system, the working mechanism of online transaction even knowing what type of database are using by a financial institution are observed and recorded by participates directly in the study of the activities of the population, possibly accompanying and assisting a member of bankers to collect information. This enables understanding of the study population from the perspective of their own activities. At the same time, it is possible to check the difference between what people are saying and what they are doing. So, in participant observations, knowledge of the local language is important. A knowledge and correct banking information like ATM, credit cards, type of online transaction helps greatly in understanding the theme of people. The System takes a long time to develop. With studies of financial institutions such as the description of system, techniques, or marketing, this may be less of a problem than for more sensitive topics such as government law, politics, though the usage of CCFD system may be part and parcel of these sensitive issues.

b. Nonparticipant observation

Nonparticipant observation also was an important research strategy for this project to study different kinds of research papers and systems to detect and prevent fraud credit card. The research can be organized, at least to the point of choosing to concentrate on a particular task such as finding the best machine learning and deep learning algorithms and technique used to balance the imbalanced dataset. During this type of observation, visualization of data is equally important so that one can know about the best algorithm and techniques with a confusion matrix. In the case of forming theories from research paper to be checked, or covering up unresolved interactions, there can be a deliberate structuring of findings. So, it is important to be careful not to enforce preconceived ideas and should stay agile and accessible to new understanding. And the main thing observe during this observation is that honesty is the best policy for both ethical and practical factors, the people who are observing CCFD should have the right to know about the purpose and scope of the system.

c. Time allocation studies

An important aspect of the CCFD survey is how much time is spent on tasks like collecting data, understanding data, understanding machine learning and deep learning, and other related activities like handling imbalanced data, feature reduction technique like PCA. It may be difficult to get precise information because an algorithm involved in detecting fraud and genuine credit is hard to understand and take more time. The biggest problem with our data is that large data size, recalling one of our initial warnings on avoids extra data collection. Most time is spent studying algorithms and train our data.

3.1.3. Research

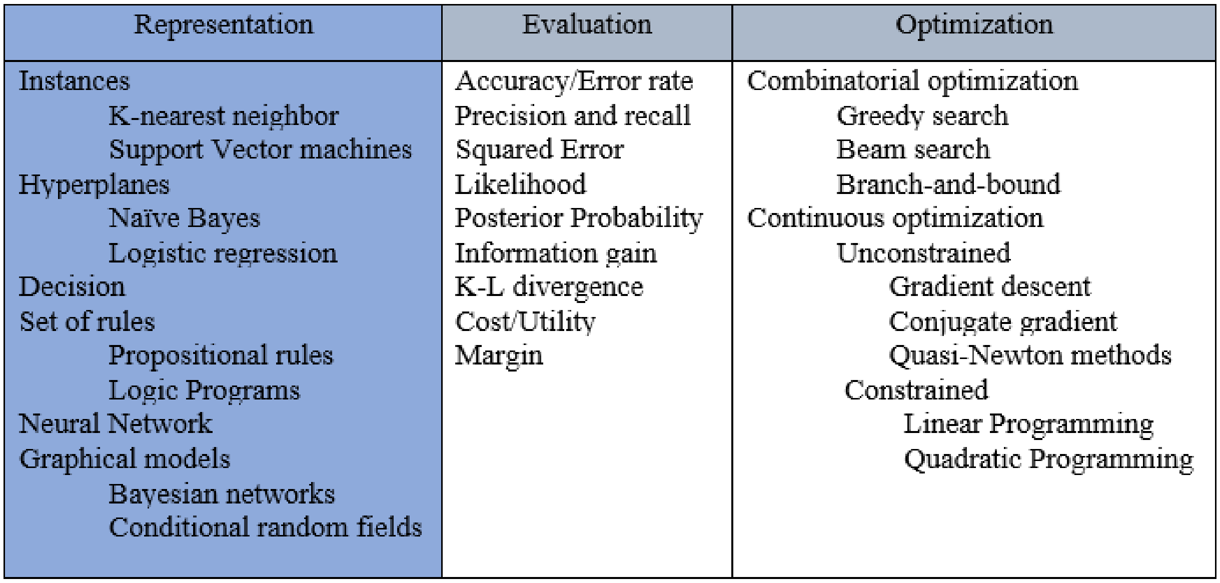
Research is the most significant process of evaluating the problems that other sources(documents or human) have already solved.(tutorialspoint.com, 2020). The research's primary aim is to guide practice, prove a hypothesis, and lead to the advancement of expertise in a field of analysis.

Research in detecting fraud credit card helps to be more familiar with the terms like:

Representation: A classifier needs to be in a structured language that the computer is able to handle.

Evaluation: A method of evaluation that is also regarded as an analytical feature or ranking system used to distinguish good classifier from weak ones.

Optimization: The optimization technique is used to look for the highest score one of classifiers in the languages.



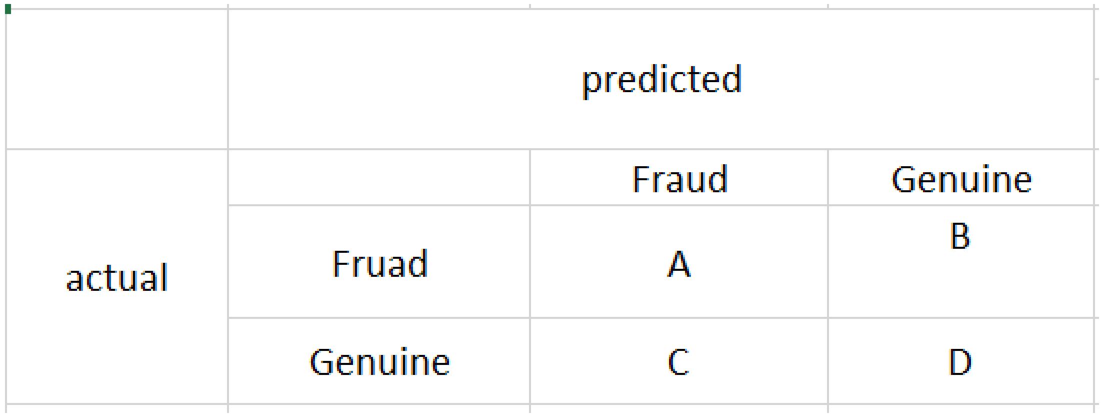
Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support Vector machines	Squared Error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naïve Bayes	Posterior Probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision	K-L divergence	Gradient descent
Set of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic Programs		Constrained
Neural Network		Linear Programming
Graphical models		Quadratic Programming
Bayesian networks		
Conditional random fields		

Table 4: Component of algorithms (Domingos, 2012)

How does a confusion matrix work?

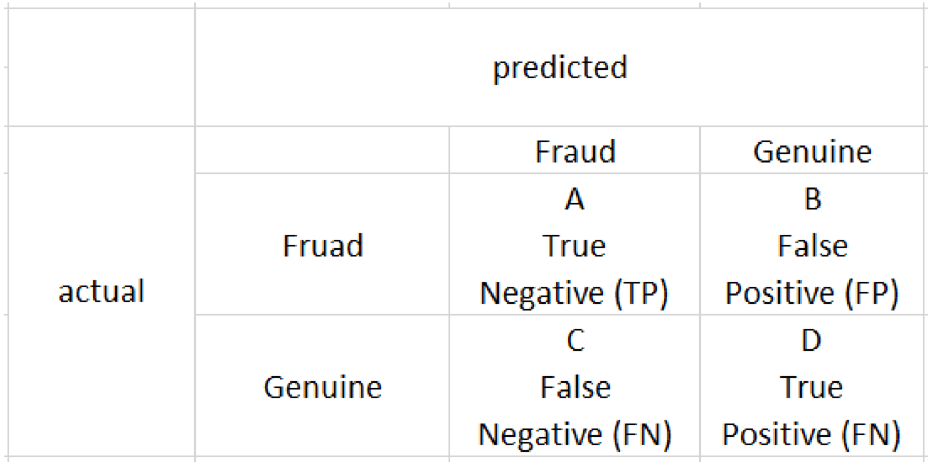
A confusion matrix is usually computed for computing the cross-tabulation of observable(true) and predicted class (model) in any machine learning algorithms like logistic regression, decision forest, Naïve Bayes and may more. There are many matrices such as precision and recall which help the model’s accuracy and pick the best model.

Suppose a 2-class case confusion matrix with Fraud and Genuine is shown below where a row represents the instances of an actual class and each column represents the instances of a predicted class.



	predicted		
actual		Fraud	Genuine
	Fruad	A	B
	Genuine	C	D

The fields in the matrix state the following



	predicted		
actual		Fraud	Genuine
	Fruad	A True Negative (TP)	B False Positive (FP)
	Genuine	C False Negative (FN)	D True Positive (FN)

Now, we can describe important performance measures in machine learning

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

The accuracy is not acceptable performance measure in machine learning. Let's say we have 1000 samples of data where 995 were fraud class and only 5 of them are genuine class. When we use classifier for this sample data, the accuracy would be a remarkable 99.5%, also classifier is unable to classify any positive samples (Visa, et al., 2011).

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

$$\text{True Negative Rate} = \frac{FP}{TN+FP}$$

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

3.2. Business Process Model (BPM)

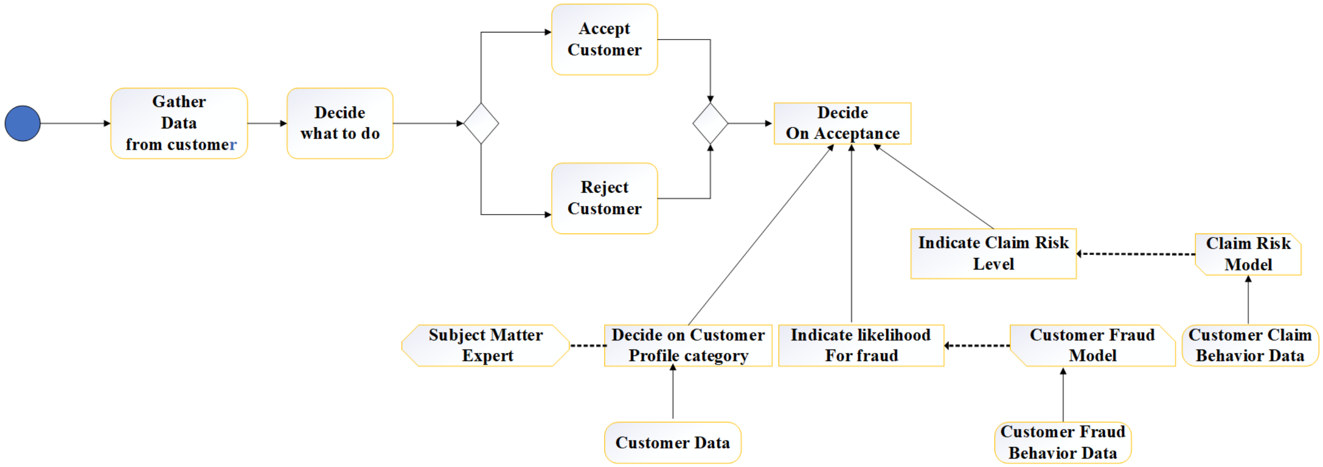

Gather Data from customer
Decide what to do
Accept Customer
Reject Customer
Decide On Acceptance
Indicate Claim Risk Level
Claim Risk Model
Subject Matter Expert
Decide on Customer Profile category
Indicate likelihood For fraud
Customer Fraud Model
Customer Claim Behavior Data
Customer Data
Customer Fraud Behavior Data

Figure 2: Business Process Model

4. Software Requirement Specification (SRS)

SRS is a paper or collection of documents documenting the features and action software application or actions of a program. In case of some confusion and disagreement SRS has been established for future reference. SRS provides of the requirements, behaviors, constraints, and performance of the system (tutorialspoint.com, 2020).

4.1. Requirement analysis

Requirement analysis has used the transform of an operational need into software configuration, software description and software performance parameter by using standard, iterative, analytical process and trade-off studies for understanding what the customer wants to analyze needs, assessing feasibility, negotiate a reasonable solution validating the specification and managing the requirements.

4.1.1. Purpose of SRS

The aim of SRS is to identify the criteria for the identification of fraud in credit cards also include a broad outline of our project, including user requirements, product perspective and requirement overview, and general constraints. It will also include the functionality needed for this project and specific requirements for this project such as interface, functional requirements, and performance requirements.

4.1.2. Scope of SRS

The scope of this SRS document like functionality, performance, constraints, interface, and reliability which persists for the entire life cycle of the project which defines the final state of the software requirements agreed upon by the customers and designers. Finally, at the end of the project, all tasks can be tracked from the SRS to the component.

4.1.3. Overview

The system specification document for the software requirements covers the following two sections i.e. General Descriptions and Specifications Requirements where general description provides a general project description including description of the production function, user features and general constraint and specification requirement describes both functional and, non-functional requirement of the project.

4.1.4. General Descriptions

A fraud in credit cards has been developed to alert the customer to their credit card fraud. After the payment method, the transactions carried out are checked if the transaction out is a true transaction or a false transaction and reduce the false alarm by applying a machine learning and deep learning algorithms.

4.1.4.1. Product Function

The project is expected to provide consistent outcomes and the functionality of the product to detect a number of fraud transactions effectively and offering flexibility to the customer in a safe and reliable manner.

4.1.4.2. User Characteristics

Customers and administrator are classified as the user of the system.

- Customer are those who make the transaction through any means
- Administrator are those who compute on the transaction and reports about the fraud usage

4.1.4.3. General Constraints

- **Audit Functions:** No audit functions are need
- **Interfaces to other applications:** No interfaces are needed
- **Control Functions:** No control functions are needed
- **Hardware Limitations:** There are no hardware limitations
- **Parallel Operations:** Required parallel operations

4.1.5. Functional Requirements

The interaction between input and output of the system is defined by SRS functional criteria.

4.1.5.1. Technical Issues

Many software projects have failed due to incomplete or incorrect analysis, including technical problems. Technical issues are a crucial factor in designing the software program.

4.1.5.2. Risk Analysis

Project risk estimation is for expense estimation with defined precision and cost for capital investment projects. The key task is to decide how to model and visualize the complex relationships between risks, to identify and track the effectiveness of risk, to assess the probability of risk incidence, to minimize the negative influence of risks, and to control the success of the project with risks and uncertainties.

4.1.6. Interface Requirements

The performance of the system is adequate. Mostly vendor deals for the user internet access, 60 percent is up to the client-side.

4.1.6.1. Hardware Requirements

- **Processor type**: Pentium III-compatible processor or faster.
- **Processor speed: Minimum**: 1.0 GHz, Recommended: 2.0 GHz or faster
- **RAM**: 512 MB or more
- **HARD DISK**: 20GB or more
- **Monitor**: VGA or higher resolution 800x600 or higher resolution
- PC/laptop/Server

4.1.6.2. Software Requirements

- **Operating System:** Windows XP Professional or more or Linux
- **Back End:** SQL server
- **Application Software Framework:** Python
- **Library:** Scikit Learn

4.1.7. Performance Requirements

Following are the performance requirements of the project:

- The key criterion is that the no-fault situation forces a sudden exit of the project.
- Any mistake that occurred at any step would send a simple indication of mistake.
- The response /answer should be fairly simple, and the participants in the experiment should not be uncertain about the operation taking place at any point in time.
- The performance of the system is adequate.

4.1.8. Non-Functional Requirements

- Secure access of confidential data (user’s details)
- 24 X 7 availability and should be efficient
- Effective configuration of the modules to maximize efficiency at peak time.
- For future extension, flexible service-based architecture will be highly desirable
- In case of failure avoiding system breakdown, the system must display the necessary information

4.2. Feasibility Study

4.2.1. Economic Feasibility

The project requires a high-end graphics processing unit (GPU) and CPU for creation however for CPU for creation however for the user to use the product a decent GPU and CPU will be enough. Hence, the system we are going to develop does not require an enormous amount of money so it will be economically feasible.

4.2.2. Operational Feasibility

The project requires the general user-friendly environment to store the dataset then predicts the frame using trained with dataset.

4.2.3. Technical Feasibility

This project requires a large amount of database space to stores credit card transaction details and processing power for processing real-time data to recognize fraud transactions, and genuine transactions. For the administrator to use the product a decent GPU and CPU will be enough.

5. Approach

5.1. Sampling Technique

Learning from class imbalanced data in a supervised learning algorithm remains to be a popular also consider as difficult problem as traditional classifications algorithms for handling balanced class distributions are structured. Although there is various approach to solve this issue, methods generating artificial data for achieving a balanced distribution of classes are more stable than classification algorithm modifications. These methods, like over-samplers, changing outcomes of training so that each classifier can be used on class-imbalanced datasets. For this problem, many algorithms or methods have been suggested, but most of them are difficult to understand, and seem to generate unnecessary noise. The research provides a clear and efficient form of oversampling focused on SMOTE oversampling, which prevents noise production and essentially overcomes imbalances between classes and within them. Empirical findings of detailed studies from the credit card fraud transaction dataset indicate that over-sampled training data for the suggested approach increases classification performance where an implementation is supported in the python programming language (Last, et al., 2017).

5.1.1. SMOTE

Synthetic Minority Oversampling Technique also known as SMOTE is a technique based on machine learning for classification of data where Kaggle data for this research is trained using the SMOTE technique to solve data imbalance which is specifically to distinguish fraud transactions from initial cardholders transactions. The transaction is initially performed in the form of a confluence. Thus, the SMOTE method trained the confluence data to synthesize illegal (fraud) transactions from non-fraud (genuine) transactions (Sahayasakila.V, et al., 2019).

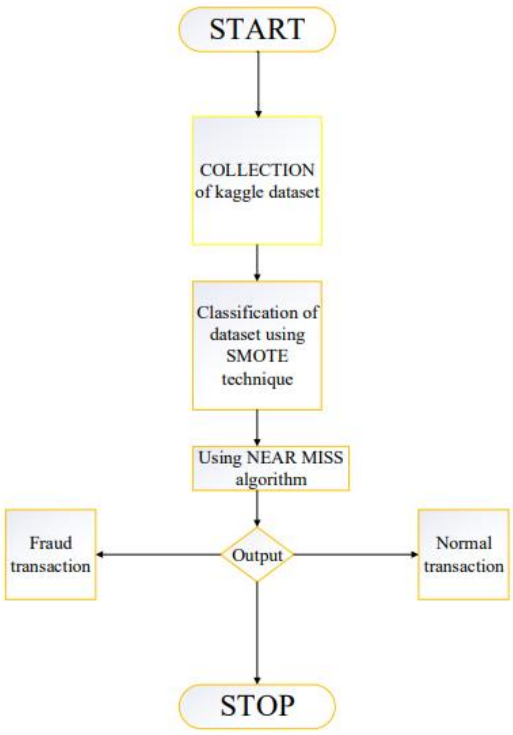

START
COLLECTION of kaggle dataset
Classification of dataset using SMOTE technique
Using NEAR MISS algorithm
Fraud transaction
Output
Normal transaction
STOP

Figure 3: Flow Chart of SMOTE Technique (Sahayasakila.V, et al., 2019)

More details on how SMOTE algorithms operates are listed below:

Step 1: For SMOTE algorithms, the initial step is to set minority class let's say set A, by calculating Euclidean distance between x and each other sample in set A, the KNN of x are obtained for each x ∈ A

Step 2: The sampling rate N is calculated by the imbalanced proportion, for each x ∈ A, N (i.e. x_1, x_2, ... x_n) are randomly selected from k-nearest neighbors, and the set A_1 is generated.

Step 3: Take an example, x_k ∈ A_1(k = 1,2,3 ... N) the below formula is used to create new sample (geeksforgeeks.org, 2020).

$$x' = \text{x} + \text{rand } (0,1) * | \text{ x - } x_k \text{ } |$$

Where,

rand (0,1) = random number between 0 and 1.

In the above figure (3), synthesized transactions are re-sampled to test the consistency of the records. NearMiss algorithm is used to configured synthesized transactions for fraud this aims at balancing distribution by randomly eliminating samples of the majority class. If two separate class instances are very similar to each other, the majority class instances are excluded to maximize the difference between the two class samples that helps in the process of classification. Also, it is used to prevent information loss of the dataset.

The dataset contains a high volume of majority class than minority class where the majority class is the genuine transaction and minority class is the fraud transactions of the dataset.

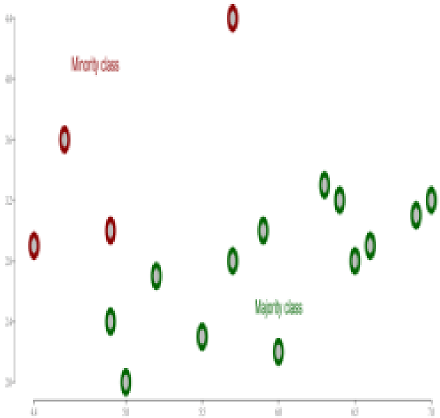


Figure 4: SMOTE Technique (Sahayasakila.V, et al., 2019)

Following are the fundamental theory on the function of near-neighbor method:

Step 1: Firstly, the distance between all the majority class cases and minority class cases are identified by the method. That is where they tend to under-sample the largest class.

Step 2: Second. the majority class of N instances which have the lowest difference to those of the minority class are then chosen.

Step 3: Third, if the minority class includes k instances, the nearest approach in the majority class would result in k * n instances.

Several variations in the NearMiss algorithm exist to find n closest instances in the majority class are listed below:

Version 1 – NearMiss: When choosing the minority class of increasing average size, the k closest instances of the minority class are smaller.

Version 2 – NearMiss: The k farthest minority class instances are smallest when selecting a majority class sample for each mean distance.

NearMiss – Version 3: It is working in two steps. First, their nearest M-neighbors will be stored for every minority class example. Finally, the majority class instances for which the average distance to the nearest n-neighbors is the highest are chosen (geeksforgeeks.org, 2020).

The below figure is the fraud transactions of SMOTE synthesis from the initial non-fraud transactions:

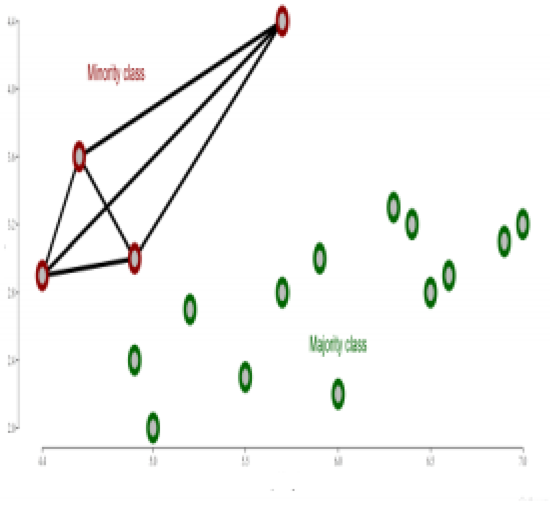


Figure 5: Synthesis of minority class (Sahayasakila.V, et al., 2019)

.

The dataset consists of credit card transactions. Out of the total transactions i.e. 284,804, the review includes 492 fraud transactions which means 0.172% of transactions are considered to be fraud. After applying the SMOTE algorithm, we get below output as our fraud transactions

```
>>> from imblearn.over_sampling import SMOTE
>>> sm = SMOTE(random_state= 42)
>>> x_Sampled,y_Sampled = sm.fit_sample(xData,yData.values.ravel())
>>> Source_data_no_fraud_count = len(data[data.Class==0])
>>> Source_data_fraud_count = len(data[data.Class==1])
>>> print('Percentage of fraud counts in original dataset:{}%'.format
...      ((Source_data_fraud_count*100)/(Source_data_no_fraud_count
...                                        +Source_data_fraud_count)))
>>> Sampled_data_no_fraud_count = len(y_Sampled[y_Sampled==0])
>>> Sampled_data_fraud_count = len(y_Sampled[y_Sampled==1])
>>> print('Percentage of fraud counts in the new data:{}%'.format
...     ((Sampled_data_fraud_count*100)/(Sampled_data_no_fraud_count
...                                        +Sampled_data_fraud_count)))
```

Output:

```
Percentage of fraud counts in original dataset:0.1727485630620034%
Percentage of fraud counts in the new data:50.0%
```

5.2. Algorithm Used

There has been a lot of research done on prevention of fraudulent credit cards. The credit card system’s history and non-technical knowledge information about the credit card can be learned from the introduction part of this report. While selecting a specific resource for literature review, majority of the credit card fraud detection is based on supervised learning algorithms like neural networks. The results of classifier models developed using a deep learning model and well-known logistic regression and random forest with an imbalanced dataset are compared in this analysis.

5.2.1. Logistic Regression

A logistic regression is well known statistical method for predicting binomial or multinomial outcomes. The multinomial logistic regression algorithm can generate models if the target fields is set field with two or more possible values. And the binomial regression algorithm is limited to those models where the target area is a binary or flag area.

We use logistic regression for classifying fraud detection. Logistic regression is a form of statistical classification probabilistic model which uses the logistic curve to detect fraud. The below is the formula to calculate univariate formal logistic curve:

$$\text{p} = \frac{e^{(c_o+c_1x_1)}}{1+e^{(c_o+c_1x_1)}} \text{ ----------- } (\text{I})$$

The value between 0 and 1 is given by the logistic curve so that it can be interpreted as a probability of class membership. To execute the regression, the logistic function can be applied as shown below: The logistic function can be applied to execute regression as shown below:

$$log_e\left(\frac{p}{1-p}\right) \text{ ------------- } (\text{II})$$

In the above equation (II), 1 – p is the probability that tuple will not be in class and p is the probability that tuple will be in class. The algorithm however selects, C_o and C_1 coefficient which optimizes the probability of incoming transactions. The logistic regression flow is shown below, where the confusion matrix is used to illustrate the classifier's performance (Sahin & Duman, 2011).

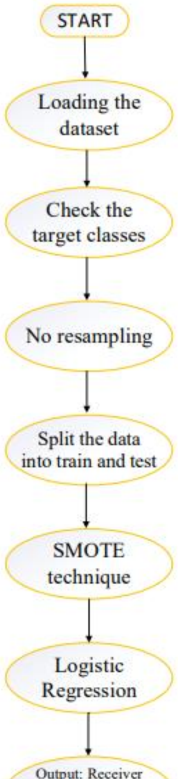


Figure 6: Algorithm flow of Logistic Regression

5.2.2. Random Forest

Random forest is the part of the supervised learning technique used to solve classification as well as regression problem whereas random forest also known as the collection of decision tree classifiers. In this research, we have not used the decision tree as our classifier model because the random forest has benefited over the decision tree as it tries to correct the training set with the habit of overfitting. A subset of the training set is randomly sampled in such a way that each tree is trained and then the decision tree is constructed, each node then splits into a feature selected from a random subset of the complete feature set. In the random forest, training is highly necessary particularly for a broad set of several features and data instances, and since each tree is trained independently of the others. It has been observed that the random forest algorithm gives a reasonable approximation of the error of generalization and is resistant to overfitting (Xuan, et al., 2018).

$$\underset{\text{Posterior Probability}}{\underset{\downarrow}{\mathrm{P}\,(\mathrm{c}\mid \mathrm{x})}} = \frac{\overset{\text{Likelihood}}{\overset{\uparrow}{P\,(x\mid c\,)\,P(c)}}}{\underset{\text{Predictor Prior Probability}}{\underset{\downarrow}{P(x)}}} \longrightarrow \text{Class Prior Probability}$$

Random forest lists the significance of variables in a natural way in classification and regression problems. The random forest uses the below pseudocode to perform the prediction of a fraudulent transaction.

Step 1: Extract the incoming transaction test features and apply the rules of each decision tree that is randomly created to generate the target(outcome) and store the generated target(outcome).

Step 2: Calculate the votes for every target output predicted

Step 3: Evaluates the highly voted projected goal as the final statistical performance from a specific decision tree.

5.2.3. Autoencoder Neural network

An autoencoder is a form of artificial neural network used in an unsupervised way to learn effective data coding. For more concrete understanding, autoencoder is a form of neural network that is learned to recreate whatever is fed as input as output. There is a series of hidden layers in between the input layer and the output layer. There is a layer right in the middle, containing fewer neurons than the input. The output of that layer is the result of the autoencoder's so-called encoder part.

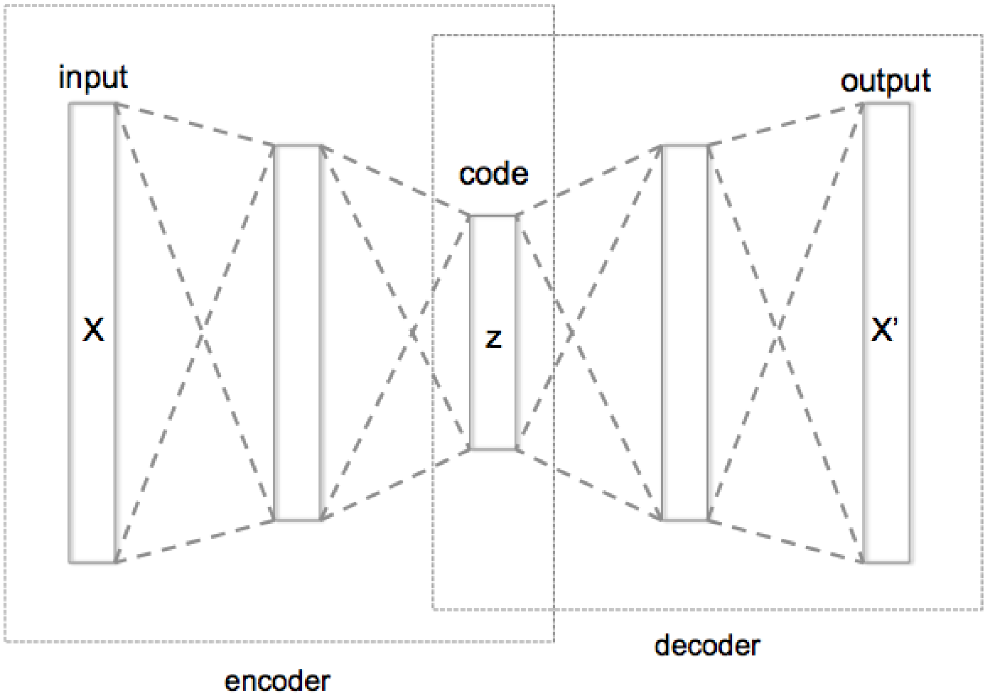


Figure 7: Architecture of autoencoder neural network

The reasonable behind utilizing autoencoders is that in a pleasant way the secrete layers map the data to a vector space. The autoencoder in this research map our input from a high dimensional space to one with fewer dimensions by using the layer with few neurons. In the above figure 9, the network structure has layer-to-layer connections but does not have a connection within each layer, where X is the input sample and X' is the output feature. Autoencoder neural network training in this project is designed to reduce reconstruction error by using the samples received. The cost function defined in the project for the autoencoder neural network is shown below:

$$J_{A,E} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\ ||X' - X||^2\right) ----- \quad (\text{I})$$

where m represents a number of input samples.

Gaussian noise, and salt and pepper noise are the widely used. The below is the cost function for denoising autoencoder neural network:

$$J_{D,A,E} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\ ||X' - X||^2\right) ----- \quad (\text{II})$$

Where, **X' = f $((\sum \boldsymbol{wX} + \boldsymbol{b}))$, b = bias and w = weights of function**

There are a range of new autoencoders known as denoising autoencoders that will enable autoencoders to learn how to remove noise and to replicate uninterrupted inputs as much as possible.

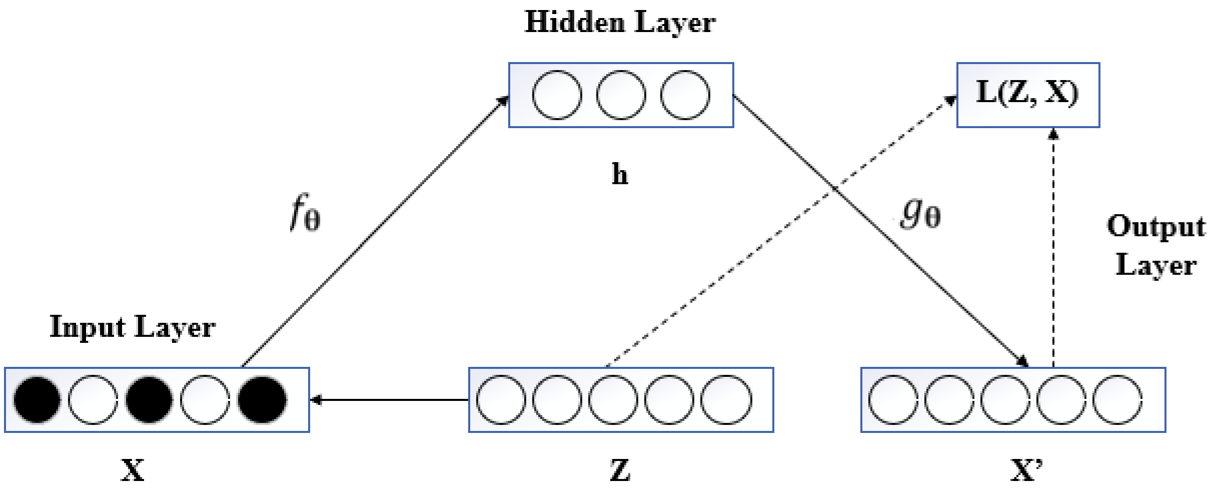


Figure 8: Denoising autoencoder neural network

The original data X and Z as shown in the figure is the data corrupted with noise. The output X' is the complete denoising autoencoder process. The aim of the loss function is to minimize the gap between the resulted output and the original data in a such way that the autoencoder can eliminate the influence of the noise and recover the features of the distorted results. Hence, the features produced by learning distorted input with noise are more robust, which enhanced the autoencoder neural network model's ability to generalize data to input. Entropy is a measure of the content of information and may be described as the impermissibility of occurrence. And the higher the probability, the smaller the impermissibility which means that the importance of the information is also very high. If an occurrence eventually happens with a 100% chance, so the unpredictability and 0 is the value of information. The entropy equation functionality is the advantage of cross-entropy loss function. So, cross-entropy is used for multiclassification problems with the SoftMax activation function. So, the SoftMax activation function in our autoencoder neural network model used in the last layer of our neural network that first determines the additive value of each component, then normalizes all the component and causes the output components total to be equal to 1. The Cross-entropy loss function can determine the efficiency of a classification model, is shown below (Yuan, et al., 2019).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} 1(y_i = j) log \frac{e^{\theta^T{}_j{}^{x_i}}}{\sum_{i=1}^{k}\theta^T jx_i} \quad ----- \quad (III)$$

5.3. Comparative analysis

In order to compare various techniques, the CCFD research measures the true negative (TN), false negative (FN), true positive (TP) and false positive (FP) produced by the method or algorithm and uses them to analyze and compare the output of different systems in quantitative measurements.

TP is the situation in which the amount of transactions that were illegal and classified or listed by the system as fraudulent transactions. TN is the situation in which the amount of transactions that were legal and classified or listed by system as genuine transactions. FP is the situation in which the amount of transaction that were legal and classified or listed by the system as fraudulent transactions. FN is the situation in which the amount of transactions that were fraudulent and classified or listed by the system as legal transactions.

The following are the different evaluation metrics:

i. Accuracy: Accuracy is the number of transactions that have been correctly classified which is also known as detection rate is most efficient and widely used performance metrics that is

Accuracy (ACC) = $\frac{TN + TP}{TN + TP + FP + FN}$

ii. Precision: Precision which is also known as positive predicted value is the number of transactions that were properly classified as genuine or fraudulent.

Precision / Positive Predicted Value = $\frac{TP}{TP + FP}$

iii. Recall: Recall which is also known as sensitivity or probability of detection or true positive rate which measures the fraction of records properly classified by the system means the record which has the maximum chance of being fraudulent.

Recall / Sensitivity / True Positive Rate: $\frac{TP}{TP + FN}$

iv. Specificity: Specificity which is also known as true negative rate measures the percentage of normal records properly classified by the system implies the records that have a minimal probability of being a fraud.

Specificity / True Negative Rate = $\frac{TN}{TN + FP}$

v. False alarm rate: False alarm rate calculates from total cases reported as fraudulent how many were incorrectly listed.

False Alarm Rate = $\frac{FP}{FP + TN}$

vi. Cost: Cost tells us that the costs of our system are effective.

Cost = 100 * FN + 10 * (FP + TP)

vii. F1-Score: Harmonic Mean of precision and recall

F1 – Score = $\frac{2 * Precision * Recall}{Precision + Recall}$

We have discussed the credit card dataset on all machine learning and deep learning strategies referred to in the previous section using the measurement metrics.

Techniques	Accuracy	Recall	Precision	F1-score
Logistic Regression	95%	94.5%	94.5%	95%
Random Forest	97%	94%	97.5%	97%
Autoencoder with balance dataset	89%	89%	90%	89%
Autoencoder with imbalance dataset	98%	89%	0.525 %	54.5%

Table 5: Algorithm Comparison

It is clear that from the table above that autoencoder without balance dataset gives the highest accuracy with a low detection rate which is only able to detect the genuine transaction log. So, machine learning algorithms i.e. random forest give the highest accuracy. While comparing logistic regression and autoencoder with the random forest, logistic regression gives a medium level of accuracy and autoencoder gives a low detection rate. Similarly, along with accuracy, a high detection rate i.e. precision and high f1-score value is offered by random forest. But is also interesting that random forest and logistic regression perform well at all conditions and they are not expensive to train while autoencoder is expensive to train at all because it involves using an optimization algorithm to find a set of weights to best map inputs to outputs.

Following are the major steps involved in current CCFD models:

i. The unavailability of full credit card details because it is private property and neither customer nor banks can disclose their information resulting in insufficient and under-trained models.

ii. A single powerful algorithm that can work reliably is impossible in any environment and can outperform any other algorithm.

iii. There is a lack of an efficient and competitive evaluation criterion that not only describes the efficiency of the system but also offers better comparative outcomes across several approaches.

iv. The Inability of the program to successfully respond to growing changes because modern misleading (fraudulent) technique and genuine improvements are made in user purchasing habits.

6. Interface Design

i. Welcome Page

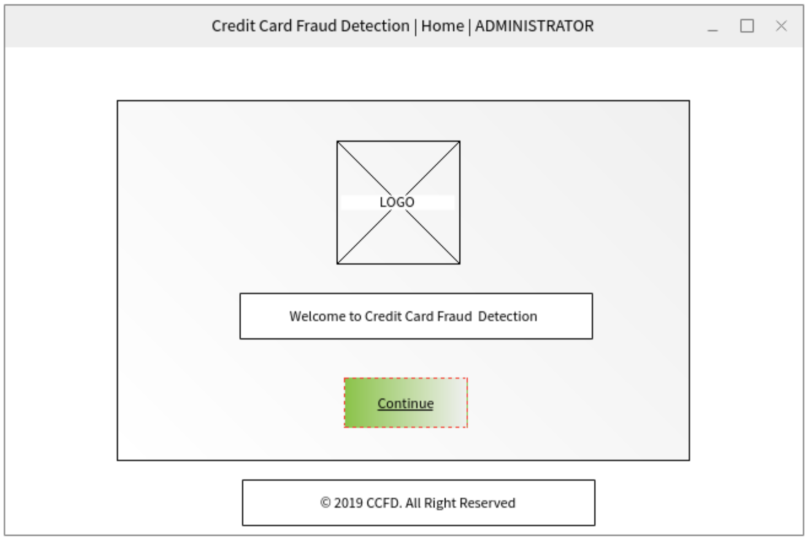

Credit Card Fraud Detection | Home | ADMINISTRATOR
LOGO
Welcome to Credit Card Fraud Detection
Continue
© 2019 CCFD. All Right Reserved

Figure 9: Welcome page

ii. Login Page

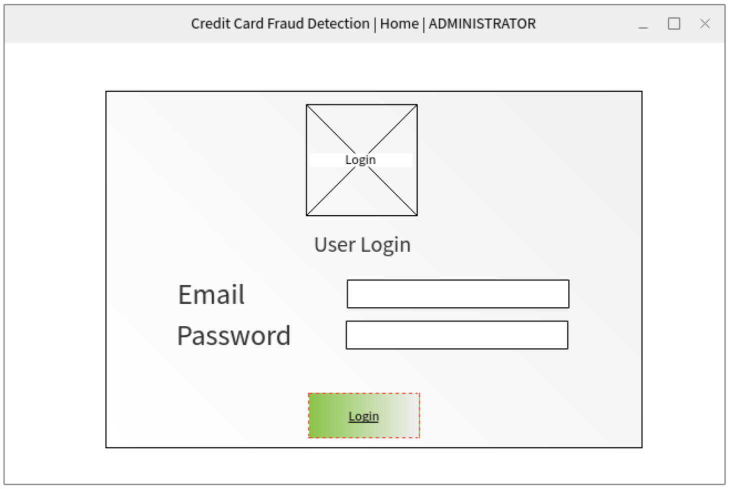


Figure 10: Login Page

iii. Dashboard Page

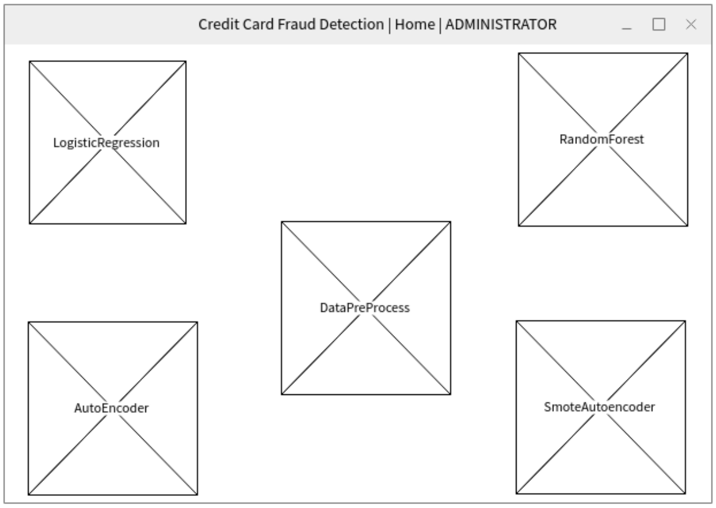


Figure 11: Dashboard page

iv. Logistic Regression Window page

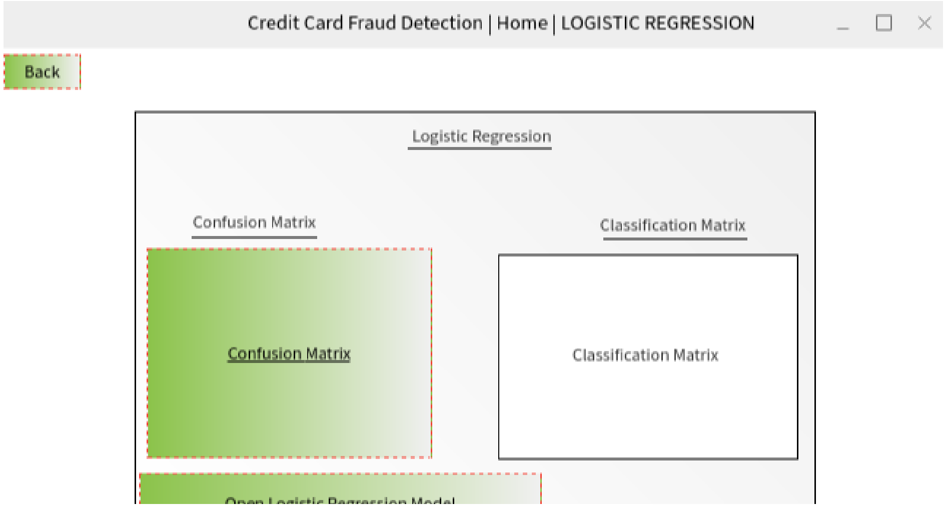


Figure 12: Logistic regression window page

v. Random Forest Window page

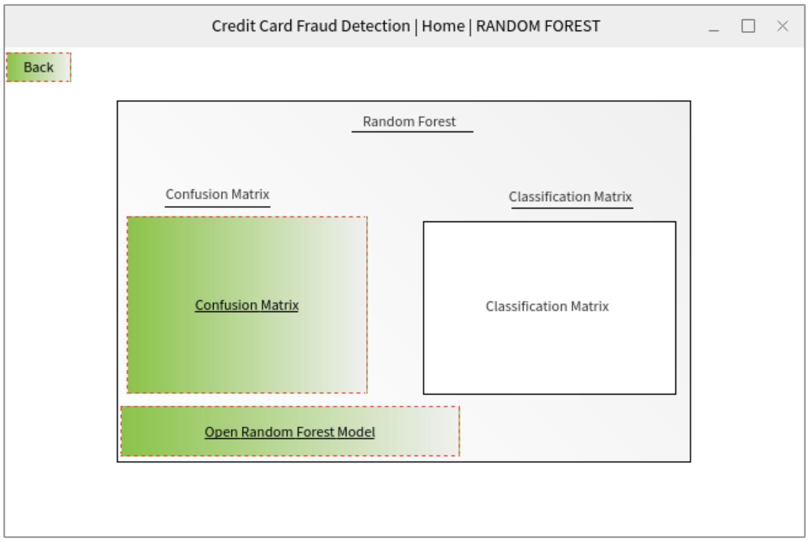


Figure 13: Random Forest Windom page

vi. Autoencoder window page

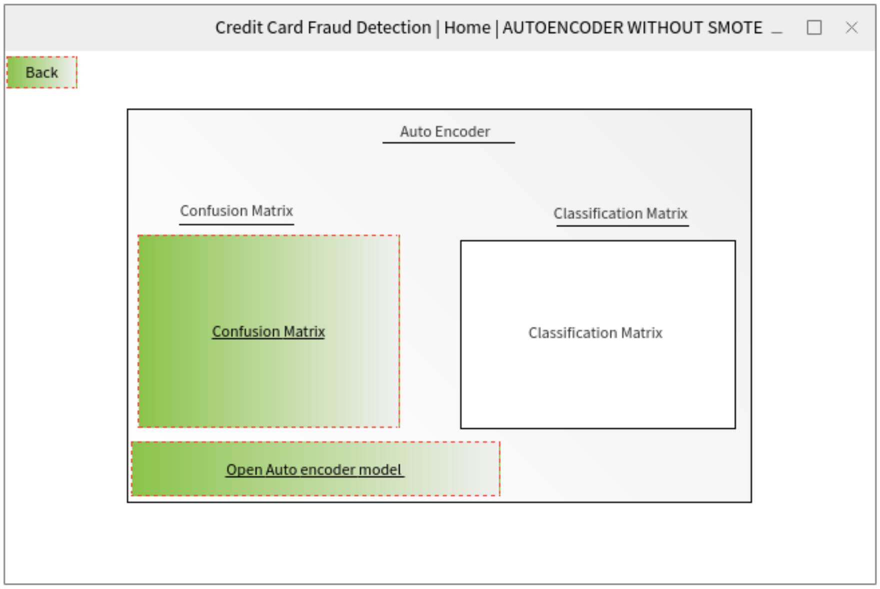


Figure 14: Autoencoder window page

vii. Smote Autoencoder window

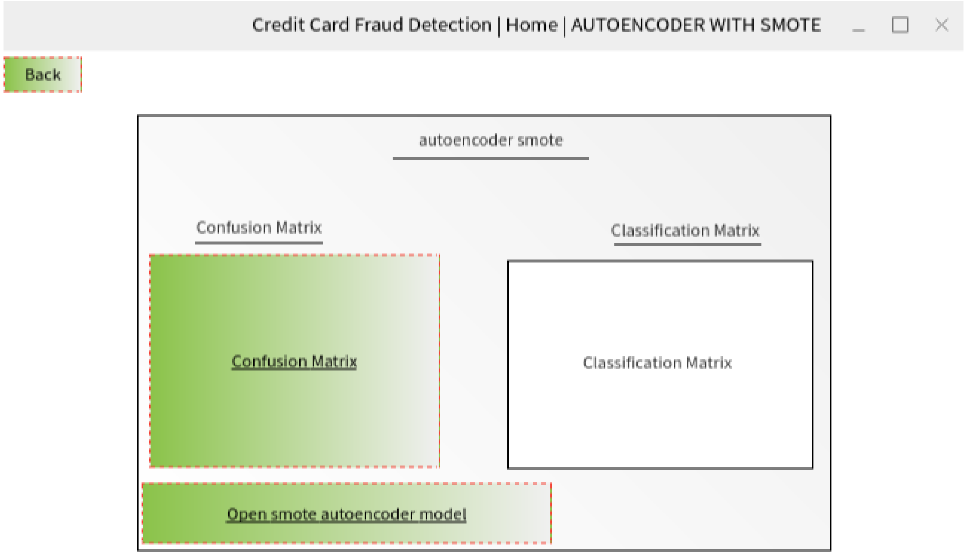


Figure 15: Smote Autoencoder page

viii. Data Pre-Process window

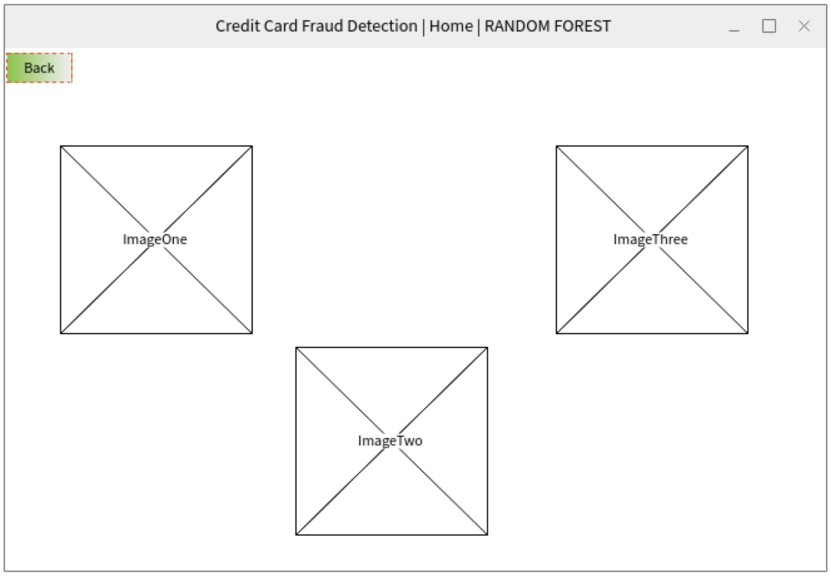


Figure 17: Data pre-process page

7. System Design and Architecture

7.1. System Design

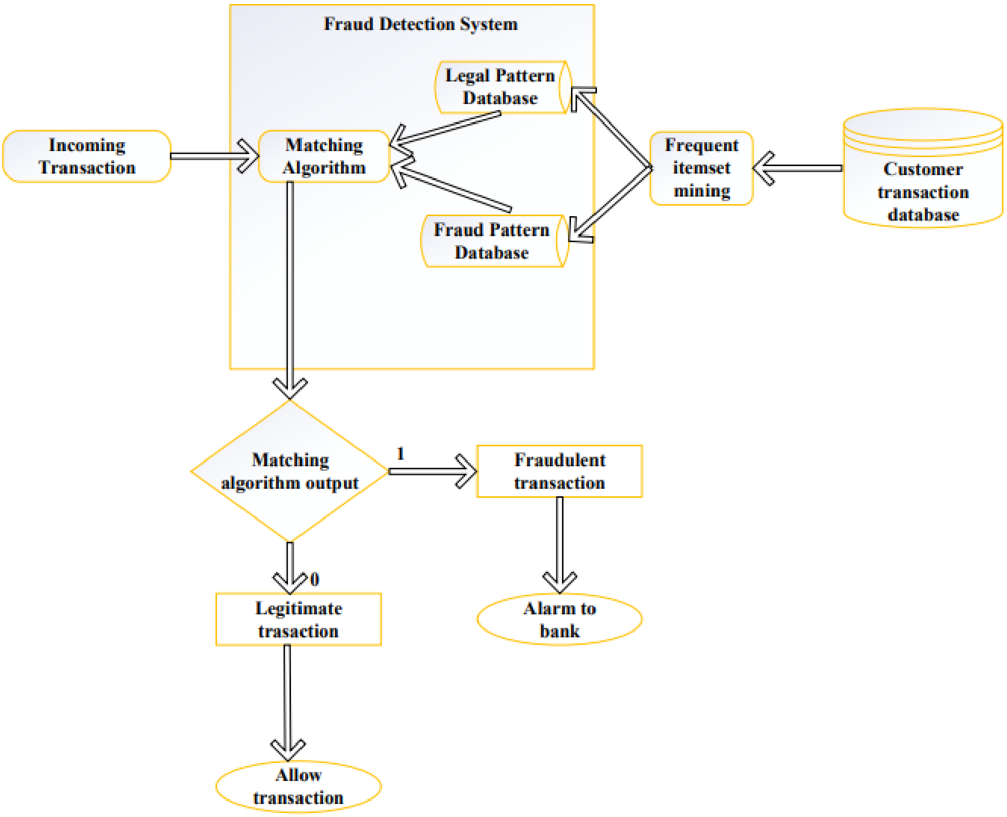


Figure 18: System Design

The above system design transverse to the legal and fraud pattern database from the customer transaction database which is far bigger in size than the legal and fraud pattern as pattern database includes just one few records for a customer to find legal and fraud transactions. For a connection with the incoming transaction and pattern database, the research suggests a matching algorithm (Logistic Regression, Random Forest Classifier, and Autoencoder) to prevent fraud. If the matching algorithm returns 0 then it will green signal to allow the customer as legal transaction and if the matching algorithm returns 1 then it will give a red signal which gives an alert message to the bank to stop the transaction.

7.2. Use Case Diagram

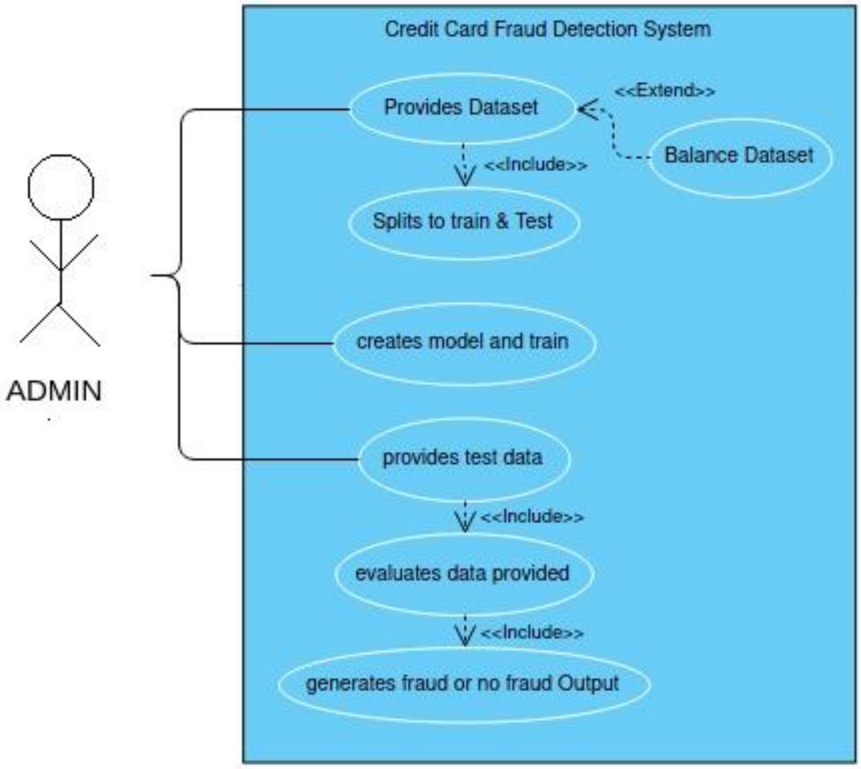


Figure 19: Use case diagram

8. Testing

Artificial intelligence (AI) technique and data-driven machine learning techniques is one of the most leading fields in both academic and industry communities. The reason that AI and machine learning has been on the cheery of the top is due to wide applications in several fields and it also combines knowledge from other major scientific fields. So, AI software testing plays an important role in building quality software system. Thus, AI software testing applies to numerous quality assurance practices for AI-based information system utilizing well-defined model, methods and techniques for quality validation which main objective is to validate system functions and features of machine-based learning models, technologies and techniques developed (Gao, et al., 2019). Some of the testing that involved while performing CCDF model are listed below:

8.1. Data Pre-Processing

i. Imbalance dataset

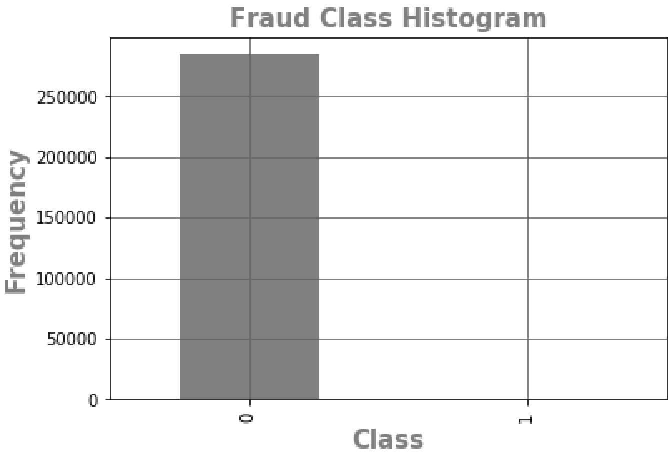


Figure 20: Imbalance Dataset

ii. Fraud transaction based on Time

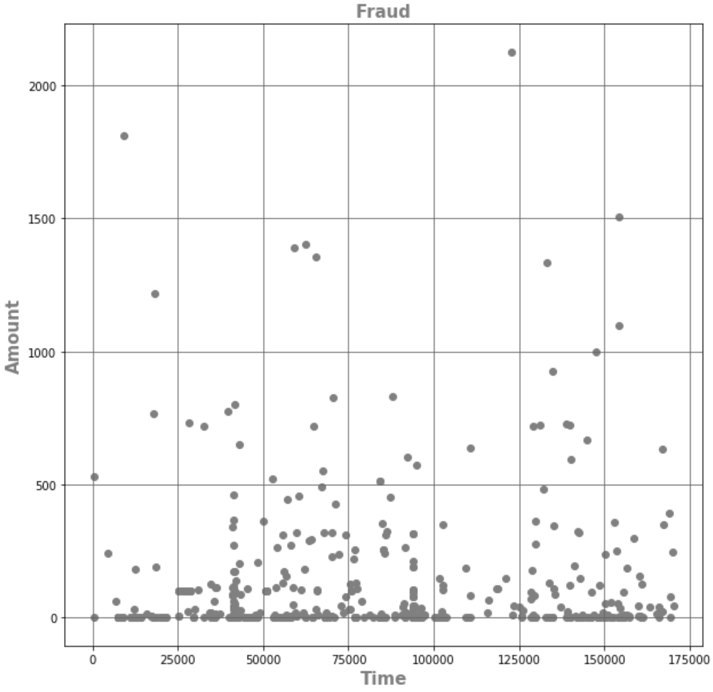


Figure 21: Graph of fraud transactions between Amount with respect to Time

iii. Normal transaction based on time

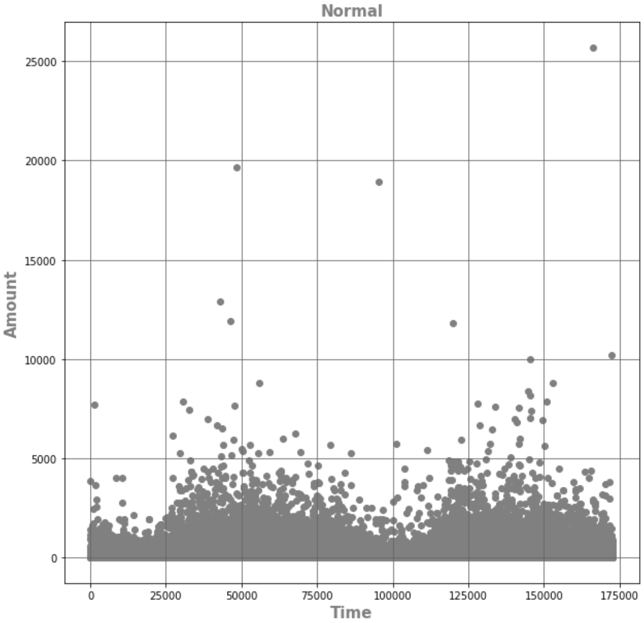


Figure 22: Graph of normal transactions between Amount with respect to Time

8.2. Machine learning model

i. Logistic Regression model at threshold = 0.36821

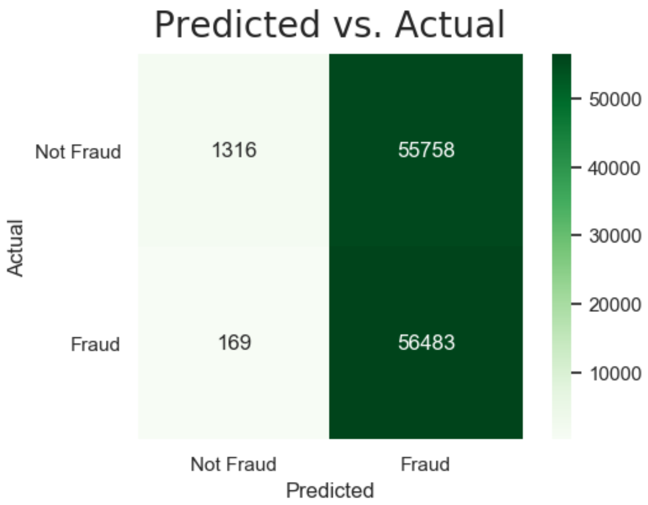


Figure 23:Logistic Regression at threshold = 0.36821

ii. Random Forest Classifier model at threshold = 0.421053’

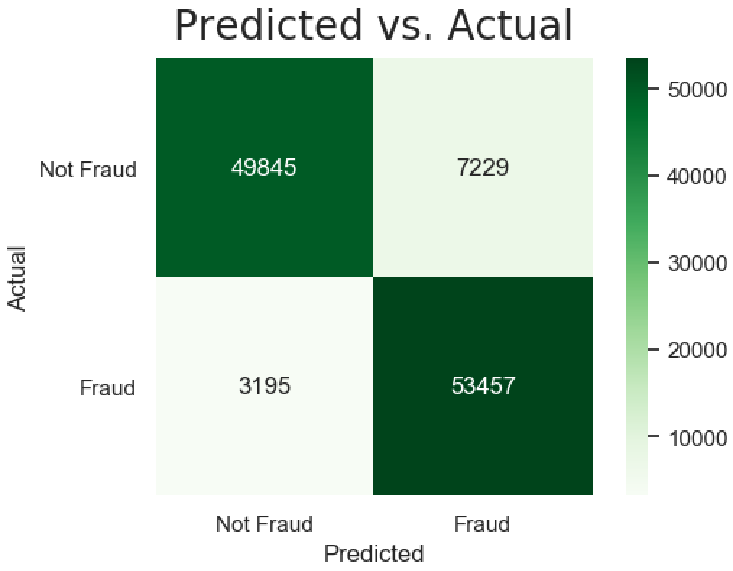


Figure 24: Random Forest Classifier model at threshold = 0.421053

iii. ROC curve of logistic regression and random forest

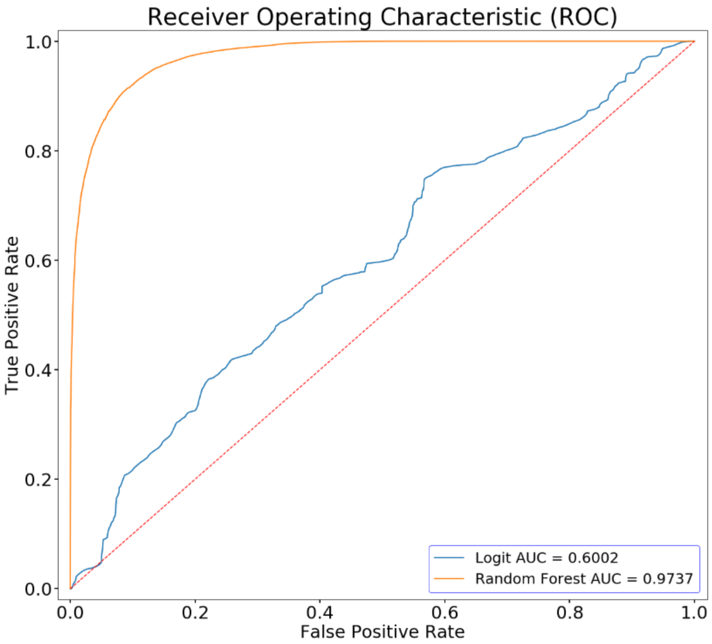


Figure 25: ROC curve of Logistic Regression and Random Forest

iv. Confusion matrix of logistic regression

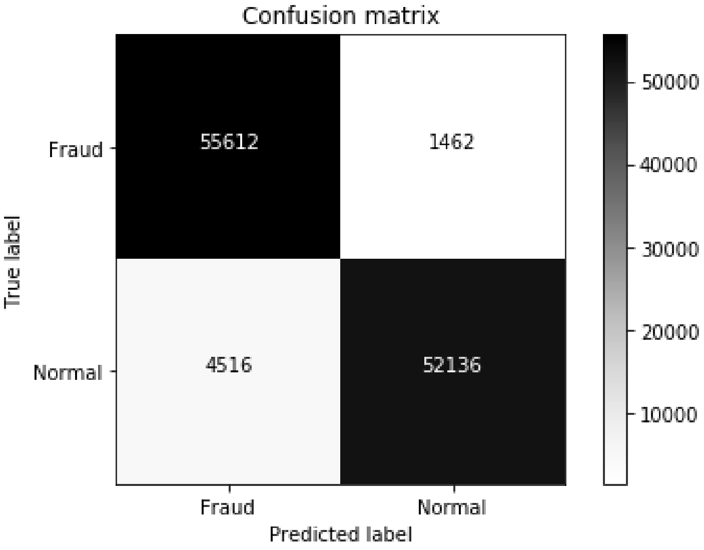


Figure 26: Confusion matrix of Logistics Regression

v. Confusion matrix of Random Forest Classifier

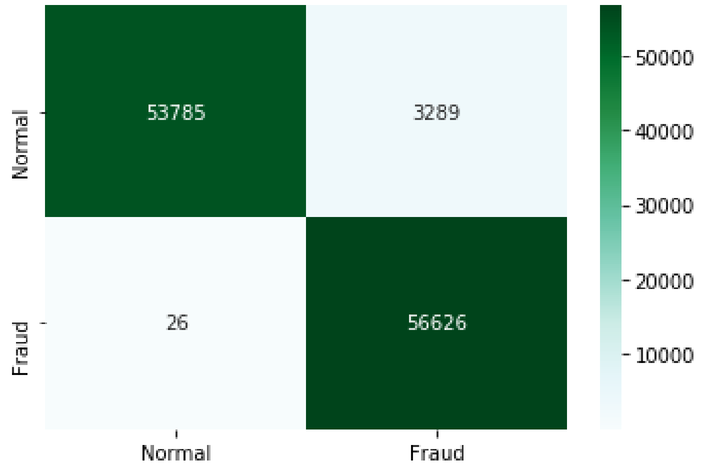


Figure 27: Confusion matrix for random forest classifier

8.3. Neural Network

i. Graph-based on model loss and epoch

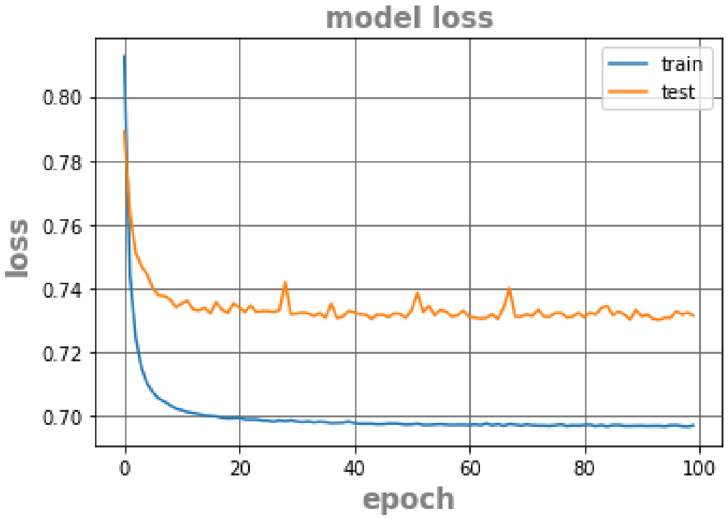


Figure 28: Graph-based on Model Loss and Epoch

ii. ROC curve of autoencoder

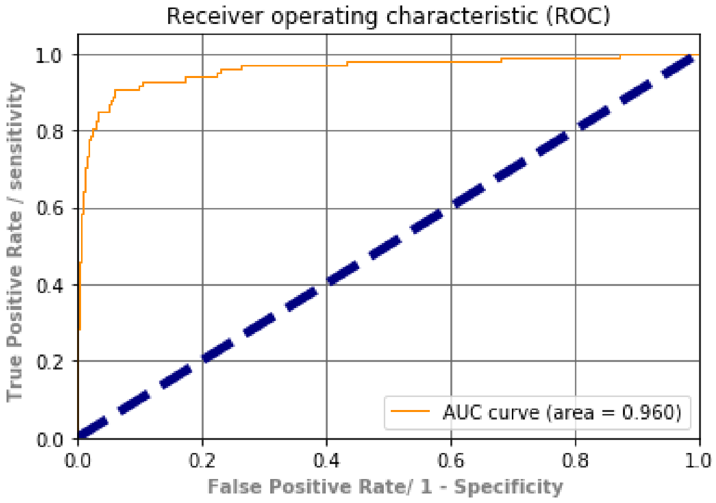


Figure 29: ROC curve of autoencoder

iii. Graph-based on recall and precision

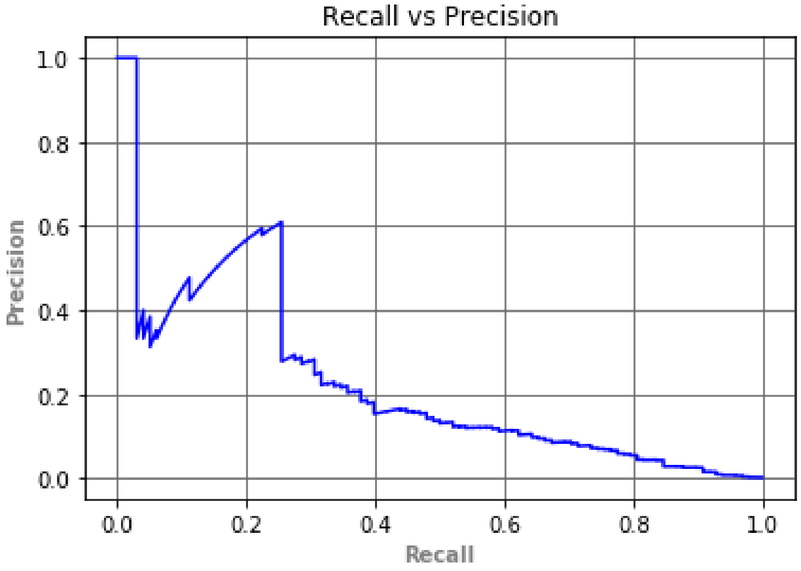


Figure 30: Recall and precision of autoencoder

iv. Precision for different threshold value

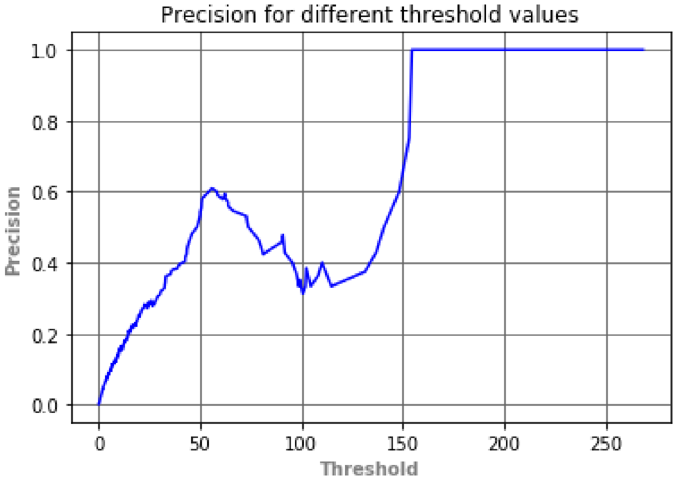


Figure 31: Precision for different threshold value

v. Recall for different threshold value

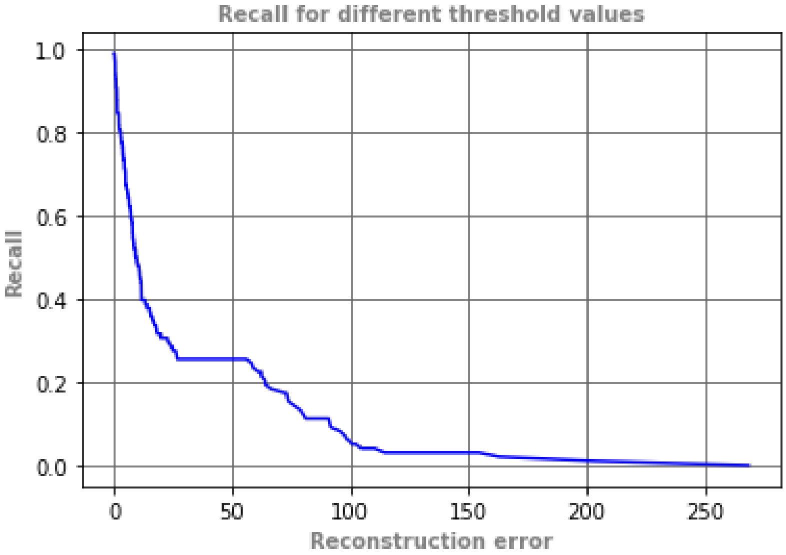


Figure 32: Recall for different threshold value

vi. Reconstruction error for different for normal and fraud cases

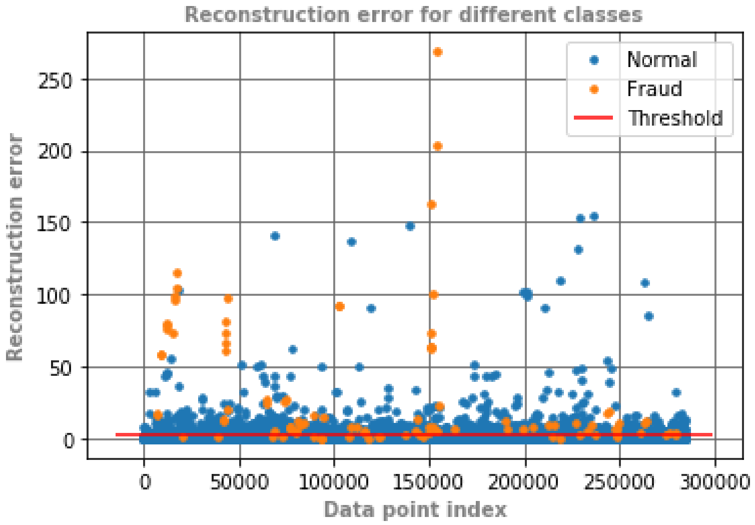


Figure 33: Reconstruction error for different cases

vii. Confusion Matrix for autoencoder

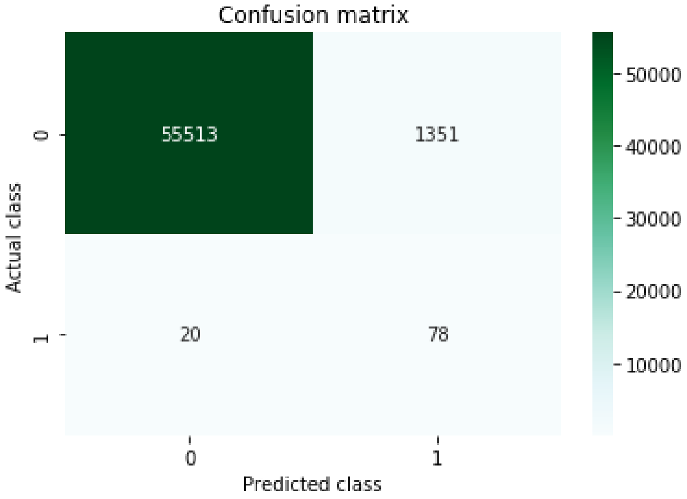


Figure 34: Confusion matrix for autoencoder

9. Answer of Academic Questions

i. What sort of problem is this project going to solve?

Answer:

Today’s era is the era of technology. So, as the increase in technology, the number of fraudulent transactions is rapidly increasing. The proposed research is based on detecting fraud cards and genuine card using different machine learning and deep learning techniques. During training and pre-processing, the machine learning and neural network is used to find user behavior and predict fraud card from data points.

ii. How actually fraud is detected?

Answer:

AI removes the time-consuming tasks and allows full data preprocessing within milliseconds and identifies complicated patterns in the most effective way to detect fraud credit card.

iii. What are the challenges involved in developing an algorithm to detect the fraud card?

Answer:

While doing research, the problem occurs where legal transactions appear just like illegitimate transactions, owing to certain circumstances. Illegitimate transactions could appear as legal transactions in another way. Most of the features have categorical data when analyzing the credit card data. So, handling categorical data is the most difficult part which means most of machine learning is not suitable to handle categorical data. The most difficult problem is to feature selection and choice of an algorithm that can detect fraud credit cards where an algorithm cannot detect the new type of data as fraud or normal transaction.

iv. Are there any similar projects?

Answer:

Yes, there are many similar researches regarding detecting fraud credit card. According to (Jain, et al., 2019), algorithms like SVM, ANN, KNN, Bayesian

Network, Decision tree, and Logistic Regression are used to detect fraud credit cards. According to (Kumar, et al., 2019), algorithms like KNN, Random Forest, and Proposed Algorithm are used to identify the complex fraudulent pattern of credit cards. Similarly, NetVerify is a web application to prevent and detect fraud credit card which use machine learning, computer vision, and biometric facial recognition to see recognized patterns of human reviews.

v. Is the proposed research feasible to handle an imbalanced dataset?

Answer:

Yes, the proposed research is feasible to handle imbalanced datasets because a technique algorithm known as SMOTE helps to produce minority synthetic samples and used to train the classifier.

10. Conclusion and Recommendation

10.1. Summary

Though there are many strategies available today to identity fraud but none of them can completely detect any fraud before it currently happens, usually they recognize it after the fraud has been performed. This is because a limited number of transactions forming the overall transactions have been found to be fraudulent transactions. Therefore, we need modern technology at a minimum cost that can identify the illegal transactions as it takes place so that it can be avoided. So, today's key challenge is to develop a precise, accurate, and fast identification of credit fraud systems that can identify internet fraud also credit card fraud i.e. when using the manipulated credit card, it is signaling an alarm.

The biggest drawback to all the technology is that in all conditions they are not likely to produce the same results. For a particular form of dataset, they offer great results and bad or unsatisfactory results for other forms. Some techniques like autoencoder which is part of the neural network with an imbalance dataset gives a high detection rate and gives a high accuracy for the genuine transaction which means this technique detects the genuine transactions and costly to train. Some techniques like Logistic Regression and Random Forest Classifier offer outstanding results on sample and preprocessed data but are not scale to broad datasets.

10.2. Future Scope

It is evident from the above comparative study of the different approaches for detecting credit card fraud that autoencoder with imbalance dataset gives high detection rate and high accuracy for the genuine transaction but this technique is unable to detect the fraud transaction and it is expensive to train and can easily be overstrained. We need to set up a hybrid autoencoder neural network that can detect both fraud and genuine transaction log with some optimization technique to minimize their experience. Genetic algorithm, Artificial Immune System, and Case-based Reasoning are optimization techniques that can be efficiently integrated with a neural network. The Genetic Algorithm helps by choosing the optimal edge weight in the neural network (Patidar & Sharma, 2011). Artificial Immune System lowers costs by eliminating weights that cause the highest mistake (Castro & Timmis, 2003). Based on direct correspondence with the user's profile, Case-Based Reasoning first attempts to predict the result (Wheeler & Aitken, 2000).

11. Critical Evaluation

After the completion of research and development of artifacts proposed different techniques logistic regression, random forest classifier and autoencoder that can be used to detect fraud credit card and SMOTE technique to handle imbalanced dataset. Consider the positive aspect a series of mathematical-statistical graphs of credit card transactions can be represented to show the user behavior. And the fraud detection techniques are used to find the complex pattern of data points the relationship between legal transaction or fraudulent transaction. The evaluation parameter i.e. Recall, Precision, Accuracy, and F1-score are used to check which techniques are more better comparing with others.

Some of the negative aspects while doing research is that some of the fraud detection algorithms are unable to handle the categorical data. And some of fraud detection classify fraudulent transactions as the legal transaction and legal transaction as the fraudulent transactions which not only affect the system but also affect whole county economies.

12. References

Al-Jumeily, D., Hussain, A. J., MacDermott, Á. & Tawfik, H., 2015. The Development of Fraud Detection Systems for Detection of Potentially Fraudulent Applications. 12, pp. 7-13.

Barker, K. J. & Sheridon, J. D. P., 2008. Credit card fraud: awareness and prevention. *Journal of Financial Crime,* Volume 15, pp. 398-410.

Castro, L. D. & Timmis, J., 2003. Artificial Immune Systems as a Novel Soft Computing Paradigm. *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* Volume 7, pp. 526 - 544.

Delamaire, L., Abdou, H. A. & Pointon, J., 2009. Credit card fraud and detection techniques: A review. *Banks and Bank Systems,* Volume 4.

Domingos, P., 2012. *Communications of the ACM.* Seattle, WA 98195-2350, U.S.A.: Communications of the ACM.

Gao, J., Tao, C., Jie, D. & Lu, S., 2019. *Invited Paper: What is AI Software Testing? and Why.* s.l., s.n., pp. 27-2709.

geeksforgeeks.org, 2020. *ML | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python.* [Online]
Available at: https://www.geeksforgeeks.org
[Accessed 10 05 2020].

Ghazi, R., 2019. Financial institutions and their role in the development and financing of small individual projects. 07.

Giri, P., 2013. PAYMENT AND SETTLEMENT SYSTEM DEVELOPMENT IN NEPAL: HURDLES AND WAY OUT. *Economic Journal of Development Issues,* Volume 15 and 16 , pp. 1-2.

Jain, Y., Tiwari, N. & Jain , a., 2019. A comparative analysis of various credit card fraud detection techniques. *International Journal of Recent Technology and Engineering,* Volume 7, pp. 402 - 407 .

Jesus, A. d., 2019. *Machine Learning for Credit Card Fraud – 7 Applications for Detection and Prevention.* [Online]

Available at: https://emerj.com/ai-sector-overviews/machine-learning-for-credit-card-fraud/ [Accessed 06 05 2020].

Kawulich, B., 2012. Collecting data through observation. In: C. Wagner, B. Kawulich & M. Garner, eds. s.l.:McGraw Hill, pp. 150 - 160.

Kumar, 1. S., Naidu, M. V. & Sujatha, D., 2019. Credit Card Fraud Detection System Based On Machine Learning Techniques. *IOSR Journal of Computer Engineering (IOSR-JCE),* 21(3), pp. 45-52.

Last, F., Douzas, G. & Bação, F., 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE. 02 11.

Mathers, N., Fox, N. J. & Hunn, A., 2000. Research Approaches in Primary Car. In: *Using Interviews in a Research Project* . s.l.:Radcliffe Medical Press/Trent Focus, pp. 113 - 134.

Patidar, R. & Sharma, L., 2011. Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE),* Volume 1, pp. 32-38.

Sahayasakila.V, Monisha, D. K., Aishwarya & S. V., 2019. Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm. *International Journal of Engineering and Advanced Technology (IJEAT)* , 8(5), pp. 190 - 192.

Sahin, Y. & Duman, E., 2011. Detecting credit card fraud by ANN and logistic regression. *nternational Symposium on INnovations in Intelligent SysTems and Applications.*

Salman, A. & Salman, A., 2015. Relationship between the Incentives Offered on Credit Card and its Usage. 01, p. 30.

Shirgave, S. K., Awati, C. J., More, R. & Patil, S. S., 2019. A Review On Credit Card Fraud Detection Using Machine Learning. *International Journal of Scientific & Technology Research,* Volume 8, pp. 1217 - 1220.

Tulasi, 2020. *Fact Finding Techniques || Fact-Finding Techniques for Requirements Discovery || Bcis Notes.* [Online]
Available at: https://bcisnotes.com/thirdsemester/system-analysis-and-design/fact-finding-techniques-fact-finding/#:~:text=Fact%2Dfinding%20techniques%20are%20a,prototyping%2C%20and%20joint

%20requirements%20planning.
[Accessed 03 05 2020].

tutorialspoint.com, 2020. *Software Requirements.* [Online]
Available at: https://www.tutorialspoint.com/software_engineering/software_requirements.htm
[Accessed 06 05 2020].

tutorialspoint.com, 2020. *System Analysis & Design - System Planning.* [Online]
Available at:
https://www.tutorialspoint.com/system_analysis_and_design/system_analysis_and_design_planning.htm#:~:text=Information%20Gathering%20Techniques,precise%20SRS%20understood%20by%20user.&text=be%20complete%2C%20Unambiguous%2C%20and%20Jargon,tactical%2C%20and%
[Accessed 03 05 2020].

Visa, S., Ramsay, B., Ralescu, A. & Knaap, E. v. d., 2011. Confusion Matrix-based Feature Selection.. *CEUR Workshop Proceedings,* Volume 710, pp. 120 - 127.

Wheeler, R. & Aitken, S., 2000. Multiple Algorithms for Fraud Detection. *Knowledge-Based Systems,* Volume 13, pp. 93-99.

Xuan, S., Liu, G., Li, Z. & Zheng, L., 2018. Random forest for credit card fraud detection. *IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),* pp. 1 - 6.

Yuan, F.-N., Zhang, L., Shi, J.-T. & Xia, X., 2019. Theories and Applications of Auto-Encoder Neural Networks: A Literature Survey. *Jisuanji Xuebao/Chinese Journal of Computers,* Volume 42, pp. 203 - 230.

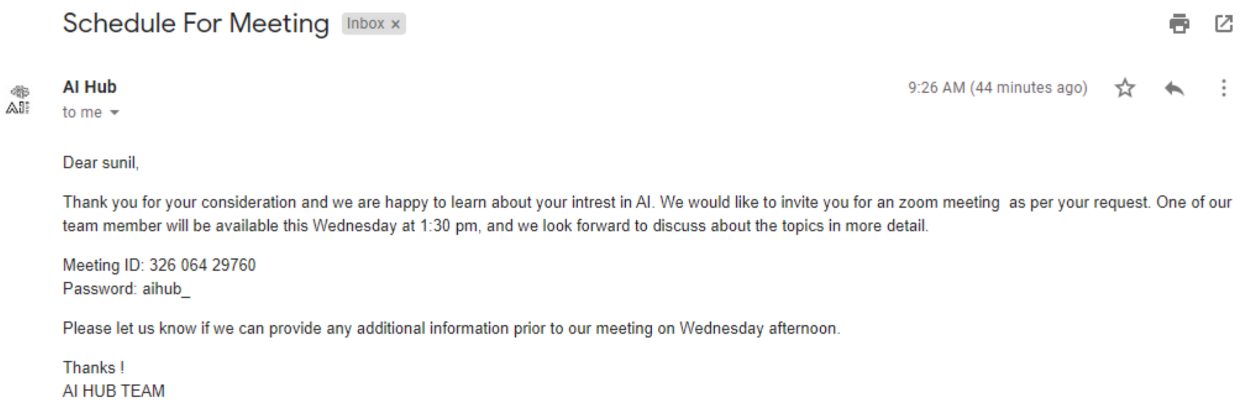
13. Appendix

13.1. Title of final project report and declaration sheet

Student Name	Sunil Ghimire
Student ID	1928584
Location/site	https://canvas.wlv.ac.uk/
Module Code	6CS020
Module Name	Project and Professionalism
Project Title	Credit Card Fraud Detection
Supervisor Name	Mr. Sachin Kafle
Submission Date	12th June 2020

13.2. Question answer session with AIHUB

During the pandemic period of time, we (Sunil Ghimire and AIHUB team) have set our meeting via a zoom call to collect standard information regarding CCFD.



Schedule For Meeting

AI Hub
to me

9:26 AM (44 minutes ago)

Dear sunil,

Thank you for your consideration and we are happy to learn about your intrest in AI. We would like to invite you for an zoom meeting as per your request. One of our team member will be available this Wednesday at 1:30 pm, and we look forward to discuss about the topics in more detail.

Meeting ID: 326 064 29760
Password: aihub_

Please let us know if we can provide any additional information prior to our meeting on Wednesday afternoon.

Thanks !
AI HUB TEAM

Figure 35: Question and Answer session with AIHUB

Following are the conclusion what we got from the meeting.

Questions:

1. How do you handle an imbalanced dataset?

➔ The obvious difficulty of addressing the class imbalanced is that one of the classes lacks records. Most machine learning algorithms with imbalanced datasets do not work very well. So, we prefer you to use SMOTE oversampling technique to handle an imbalanced dataset which helps to align the dataset by increasing the unusual sample size.

2. How do you classify the feature and target class of the dataset?

➔ The dataset for your research is known as supervised data where the target class variable is dependent to feature class variables and feature class variable is independent variables. So, class 1 which is regarded as fraud class, and class 0 is regarded as normal transactions is the target variable and, the rest of the other are your feature class variables.

3. On What basis you remove the null values from the dataset?

➔ One important step in data wrangling is the removal of null values from the dataset because it adversely affects any machine learning algorithm's performance and accuracy. So, before implementing any machine learning algorithms to the dataset, it is really important to delete null values form the dataset.

4. On what basis is the visualization allocated?

➔ Data visualization involved graph or map to facilitate the identification of trends, patterns, and outliers of broad datasets and it is important to promote the interpretation and analysis of human brain data. In your research visualization is allocated to find the distribution of fraud class and normal class, the relation of time with fraud and normal class, data after handling imbalanced dataset, the graph of features from v1 to v28.

5. What is the curse of dimensionality and what are some ways to deal with it?

➔ The curse of dimensionality applies to the anomalies that arise while classifying, storing, and evaluating high-dimensional data not found in low-dimensional spaces, especially the problem of data “closeness” and data “sparsity”. Dimensionality reduction is used to solve the curse of dimensionality by reducing the feature space.

6. The dataset obtained from Kaggle contains only numerical input variables which are the result of PCA transformation. So, why is PCA needed in Machine Learning?

➔ PCA is an unsupervised and non-parametric mathematical method used in machine learning for predictive models and used mainly to reduce the dimensionality of a dataset with several variables compared with each other through maximizing accumulation with variations present in the dataset. Some of the applications of PCA in machine learning are listed below:

   i. In lower-dimensional space, we can visualize the large complex data.
   ii. We can use it as a technique for selecting features.
   iii. For supervised learning issues, we will use the key components as data.

7. What are the types of data mining techniques that can detect the actual card and fraudulent card?

➔ Logistic Regression, Bayesian Network, Hidden Markov Model, Decision Tree, Random Forest classifier are the types of data mining techniques that can detect the actual card and fraudulent card.

8. How does one choose which algorithm is best suitable for the dataset at hand?

➔ To choose algorithm we are looking for precision and recall, specificity and sensitivity which are the accuracy measurement metrics of algorithm. Also, the ROC curve, TPR, and FPR can be used for algorithm selection.

9. How to apply machine learning in fraud detection?

➔ To identify the fraud transactions, data collection in the machine learning model is the initial step and analyzes all the collected data, segments it, and extracts the features it requires. And the model finds the complex pattern of the training dataset which is consider as a fraud card.

10. What are the factors I must consider before comparing the performance of two-meta algorithms applied to a problem?

➔ The speed of the convergence and convergence rate with the detection rate is one of the factors.

11. Why do we need a validation set and a test set?

➔ In fact, the validation set is used to build a model and neural network which is considered as an integral part of the training set. Similarly, the training set is used for evaluating the performance of the model and neural network.

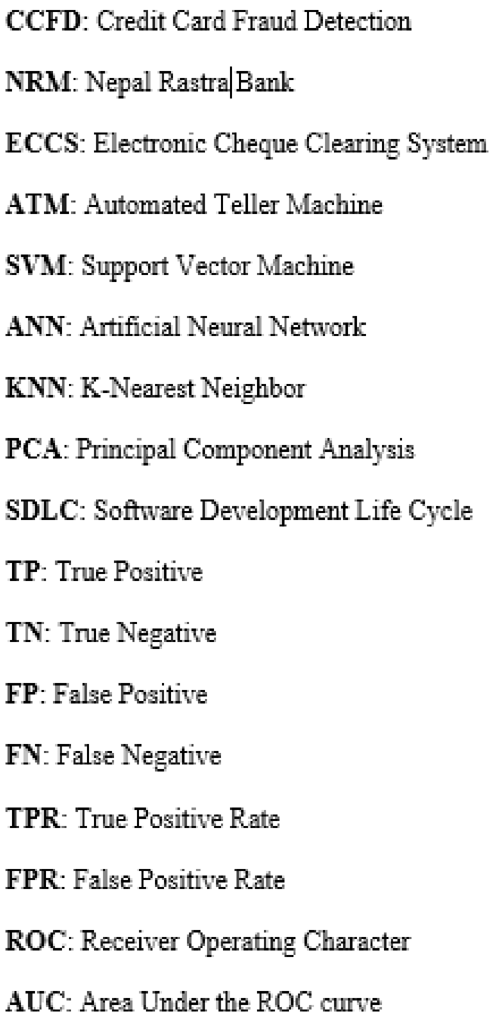
12. On what basis is k-Fold cross-validation allocated?

➔ Cross-validation is a resampling process that is used on a small dataset to validate machine learning models. In another word, to use a limited sample to predict how the model is going to act normally and to draw conclusions regarding data not used during model testing.

13. What are some factors that explain the success and recent rise of machine learning and deep learning?

➔ A large number of data accessible and strong processing capacity allows big business to spend massive capital in this technology. Rather than seeing this as the advent of emerging technologies, it is the product of significant corporate participation. Ultimately generate further jobs because of broad business participation and massive expenditure in science means that more individuals are drawn into it.

13.3. List of Symbols and Abbreviations



CCFD: Credit Card Fraud Detection

NRM: Nepal Rastra Bank

ECCS: Electronic Cheque Clearing System

ATM: Automated Teller Machine

SVM: Support Vector Machine

ANN: Artificial Neural Network

KNN: K-Nearest Neighbor

PCA: Principal Component Analysis

SDLC: Software Development Life Cycle

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

TPR: True Positive Rate

FPR: False Positive Rate

ROC: Receiver Operating Character

AUC: Area Under the ROC curve