IMBALANCE DATA CLASSIFICATION TO IDENTIFY FRAUDULENT TRANSACTIONS

BY

Rafat Karim
ID: 161-15-867

Md. Rifat Mahmud
ID: 161-15-841

Maksuda
ID: 161-15-854

MD. Jannatus Saiyem
161-15-836

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Dr. S. M. Aminul Haque
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By
Mr. Ohidjjaman
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2019

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of Dr**. S. M. Aminul Haque, Associate Professor, Department of CSE,** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Dr. S. M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Mr. Ohidujjaman
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Rafat Karim
ID: 161-15-867
Department of CSE
Daffodil International University

Md. Rifat Mahmud
ID: 161-15-841
Department of CSE
Daffodil International University

Maksuda
ID: 161-15-854
Department of CSE
Daffodil International University

MD. Jannatus Saiyem
161-15-836
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our profound our indebtedness to **Dr. S. M. Aminul Haque,** Associate Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. S. M. Aminul Haque**, **Mr. Ohidjjaman**, and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Because of the expansion of social media and globalization now a days, peta byte scale of data is being generated in every second. Data mining is the process of extracting knowledge from this huge amount of data. Data mining applications are becoming more useful and key pre-requisite for any kind of business scenarios. However, for certain applications is supervised learning, lack of sufficient data for certain classes creates data imbalance problem. For example, in a credit card fraud detection application, most of the transactions are not fraud and few of them are fraud. In our research, we have applied some classification techniques on an imbalanced data set. We have tested synthetic data from a financial payment system because it is a great challenge to obtain real dataset. Synthetic data is artificially constructed which mimics real world events. We have tested Decision tree, Support Vector Machine, Artificial Neural Network and Adaboost algorithms to treat with class imbalance problem. Among these algorithms, we find promising accuracy from Adaboost compared of others. So in this paper, our main target is that for an imbalance dataset which classification algorithm performs better.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Data is considered to be the oil or fuel for the next generation. Data mining is one of the most widely used methods to extract hidden information from large datasets. The main goal of mining is knowledge discovery from databases, which in known as KDD. Mining and discovery is quite similar in the domain of data mining. We discover knowledge by doing mining from the databases.

Now the question is how to learn from the dataset. The answer is that there is some classification algorithm for the data mining filed and these are Support Vector Machine algorithm [1], Decision tree algorithm [2], Artificial Neural Network algorithm [3] and Adaboost algorithm, [4] etc. We train the dataset using these algorithms which classifies for us. Depending on the same or different scenarios, these algorithms' accuracy could be different. Most of the time this problem occurs for bi class datasets, and it also can occur for multi-class datasets as well. Another important term is supervised and unsupervised learning. In supervised learning class labels are known and at unsupervised learning class labels are unknown. And about our dataset this is supervised learning because which mentioned algorithm we have used those are best for supervised learning.

We have already mentioned that we are using imbalance data. So at first we need to know what is the imbalance problem is. Imbalance problem means that in data set class labels are not equal. Let an example to explain this imbalance problem. Suppose a dataset is about weather, and there are several attributes in this dataset. Depending on these a final attribute reflects a decision that a cricket match will take place or not that means this is class attribute. So suppose in dataset total instances are 500. 400 instances class attribute shows YES and rest of that mean 100 instances are NO. That mean the dataset is not balanced. This is called the imbalance problem. When we train any imbalanced dataset we get higher accuracy for the precedence class and get bad accuracy for the inferiority class. Most of the time the real-world data sets are

imbalanced. When we are going to predict by training those datasets there is a good chance of risk to get it wrong without knowing which algorithm to use and which algorithm got the best accuracy for imbalanced dataset. Suppose we are trying to detect cancerous cells by predicting from patterns from a medical dataset that is imbalanced. If the accuracy is not up to the mark, then it might get dangerous for the patients. In our paper, we are going to Decision tree algorithm , Support Vector Machine algorithm , Artificial Neural Network algorithm and Adaboost [4] classification algorithms to an imbalanced dataset to see which algorithm got the best accuracy for class imbalance problem.

We are mainly going to work with imbalanced data and it is our main focus. So we have selected synthetic data from a financial payment system that shows the fraud activity of a transaction system. Because it is so tough to get the real world dataset. Based on some criteria the dataset will show if fraud is made or not so it makes it a bi class dataset. As the data set is imbalanced so it is hard for the algorithms to classify objects properly and accuracy for the minority class will be poor. Now it is our work to find out which algorithm that we have chosen performs well for this imbalanced dataset. We have already mentioned the algorithms that we have used here. These algorithms are playing vital roles in machine learning. Data mining is the part of a machine learning where we use data to create patterns and then taught patterns to the machines for future prediction and then the machine can take its own decision by using the knowledge it has gained.

Classification algorithms are the core thing of the data mining field. These algorithms can work fine when the datasets are balanced, but the classification becomes poor when the dataset is imbalanced. In our paper, we will try to learn which algorithm has better performance when it comes to imbalanced datasets and also try to figure out why the algorithm gives better than other algorithms. Then we will recommend the algorithm in the resulting part for using it to classify for imbalanced dataset because most of the realistic datasets are not balanced that mean those are imbalanced. In data mining, if instances of data set are not equilibrium that means the data set ratio is not equal than it is called imbalance problem. That means instances from one class is momentously greater than another class.

So, if we see that still the imbalance problem is a big fact for data mining. That mean it is so tough to predict from imbalance data. So, our main focus is to find the algorithm that is best for imbalance data classification. For that we have used an imbalance dataset and then apply the classification algorithms those we have mentioned above.

After applying the mentioned algorithms for our imbalanced dataset we found the accuracy of these algorithms. But we found out that Adaboost had the best accuracy over our imbalanced dataset compared to the rest of the mentioned algorithms. Considering this result we can reach our target of finding the best algorithm for our imbalanced dataset.

In this paper we tried to cover the following parts: Related work describes in part2. Methodology that mean algorithms are in part 3. Part 4 contains the result. At the last, part 5 concludes the paper

1.2 Motivation

Frauds in money transaction can be a great threat for the economic condition of a country. So, to prevent fraud in money transaction fraud detection is a very effective way. But the matter of regret is that there is a few number research on this topic of classify imbalance data for fraud detection. And one of the most obvious reason is that as a matter of privacy and confidentiality there is also lack of real transactional data. So, our aim is to classify the imbalance data to detect the fraud of financial payment system. And we hope that this analysis will be very helpful to classify the imbalance data to detect fraud and to prevent fraud from financial payment system.

1.3 Rationale of the Study

Financial payment system is one of the most common form of money transaction. But fraud in this transaction is a growing concern which can harm the economic condition of our country. So, fraud detection can be a productive way to prevent fraud in money transaction. A large number of researches on fraud detection can help to accomplish this goal. But there is a lack of real transactional data because of privacy of the user and as a result of that the amount of research is also very poor. So, our objective is to classify the imbalance data to detect fraud in financial payment system using data mining. We also did not find any real data, so have used a synthetic dataset to perform our analysis.

After dividing the dataset into two parts we looked forward to find the accuracy of the test set and based on that accuracy we can predict the possibility of fraud in new data and that can be very effective to prevent fraud. And we hope that the prevention of fraud will protect our economic condition from being decreased.

1.4 Research Questions

1. What will be the accuracy of the test data set?

2. How to classify the imbalance data to detect fraudulent behavior in transaction?

3. How can we predict a fraud in money transaction from imbalance data?

1.5 Expected Outcomes

We are trying to classify the imbalance data to detect fraud from previous dataset and for this we have divided our dataset into two parts such as training set and test set. If the accuracy of the test dataset is impressive then we can apply the model to the new data and can predict the possibility of fraud in that new transactional data. So ultimately our expected outcome would be finding the accuracy of the test data set using Decision Tree, Support Vector Machine and Artificial Neural Network algorithm and to understand the behavior of imbalance data to fraudulent and non-fraudulent transactions using these data mining algorithms.

1.6 Report Layout Chapter

In this paper chapters are oriented as follows: After the introduction we have discussed the background study in chapter two that mean the existing work about this and summarize of them. Then the research methodology in chapter three. After that with this methodology what results we have got those are in Experimental Result & Discussion. And then the conclusion and implementation for future scope are described in chapter five. At last there are some references at chapter six.

CHAPTER 2

BACKGROUND STUDY

2.1 Related Works

Yanmin et al. [1] increasing the accuracy of imbalanced data classification by exploring meta-techniques like Adaboost boosting algorithm and used imbalanced dataset for training by the classification algorithms. Ahmed et al. proposed LIUBoost for imbalanced data classification. LIUBoost mainly uses for under-sampling. LIUBoost is more suitable over RUSBoost. Subudhi et al. [3] work in fraud detection and applies data mining techniques which are Support Vector Machine, Decision Tree and Multi-Layer Perceptron. In the auto insurance sector, this methodology helps with an adaptive over sampling of various techniques. Brown et al. [4] tried to boost or increase the rate of predicting the credit loan data which was imbalanced and also used five real-life credit data sets.Sun et al. [5] work for both learning time reduction and accuracy improvement. AdaBoost algorithm for developing the classification performance. Dhankhad et al. [6] applies the multiple supervised machine learning algorithms that are used on real-world datasets to detect credit cards in fraudulent transactions. Further, find out accuracy and check performance for the supervised machine learning algorithm. Xuchun Li et al. [7] work with both SVM and Adaboost and combined it together and named it AdaboostSVM. SVM doesn't easy as a classifier to train but AdaboostSVM became easy to train and it worked just fine like SVM.

Thanathamathee et al. [8] worked based on the location of separating function and a new technique proposed which related to data boundary of each sub-cluster ans also used the concept of bootstrapping along with the use of Adaboost algorithm. Iain Brown et al. [9]worked with imbalanced date to find the good payer and bad payer among the people. Used five real world datasets and Gradient boosting, Random forest has the best result. Dech Thammasiri et al. [10] used real-world imbalanced dataset then tested three balancing techniques—oversampling, under-sampling and synthetic minority over-sampling (SMOTE)—along with four popular classification methods—logistic

regression, decision trees, neuron networks and support vector machines. The results indicated that the support vector machine combined with SMOTE data-balancing technique achieved the best classification performance with 90.24%. Guo Haixiang et al. [11] collected 527 papers that are related to imbalanced learning and reviewed all papers from both a technical and a practical point of view.

Above we have discussed the existing work that have already done with imbalance data. In [12] they have used Artificial Neural Network, Support Vector Machine and Decision tree and get the following precision 77.8, 77.0, 73.5 and recalls are 75.8, 76.1, 72.3. In [13] they used several algorithms for three types of financial datasets. From those they got 85.70% on the decision tree over on a dataset which is better from boosting technique. For different dataset different algorithm performs better. In [14] they used real dataset and used support vector machine on that dataset. They got the accuracy 55.1% over that dataset. Over these existing works if we want to compare these with our work than the main point is that we have used synthetic dataset because to obtain the real-world dataset is so difficult. But our main target is that in an imbalance data which algorithm performs better. In this paper we have already mentioned some algorithms that we have used on our dataset. Though our dataset is imbalanced, we get the best accuracy on Adaboost technique and the accuracy is 99.44%.
From this we can say that over an imbalance dataset Adaboost technique performs better from another classification algorithm.

2.2 Research Summary

The summaries of research papers which we have read during our research are given below: The application of data mining techniques for classify the imbalance data to financial fraud detection: A classification framework and an academic review of literature.

Objective of the paper: This paper represents some technique of data mining algorithm which is discussed about imbalance data classification to identify fraudulent transactions. For financial fraud detection 49 journal articles which published from 1997 to 2008 are analyzed. They identify four types of financial fraud and applied six classes of data mining technique. For research they followed methodological

framework, classification framework for application and analyzed imbalance data classification to identify fraudulent transactions.

Algorithm/Method used by the article: Classification, Clustering, Regression, Visualization, Prediction, outlier detection.

Result: In this paper they worked on four categories of fraud (bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud). And they get different types of problems. And find out some limitations which creates several types of problem.

Future work: In this paper has two main limitations. Firstly, we used several keywords which published between 1997 and 2008. A future review could be expanded in scope. Secondly we write it in English and in future try to convert it different language.

Learned lessons in credit card fraud detection from a practitioner perspective. Objective of the paper: Due to fraud in credit card transaction there is caused a loss of billion dollars every year. So to reduce the losses, designing efficient algorithms for fraud detection can be an effective way. But the designing of these algorithms is very difficult for some reasons such as, dynamic distribution of data, incompatible distribution of classes and dynamic flows of transactions.
And there is also a lacking of real data for confidentiality and privacy matters. As a result we cannot be able to identify which is the most effective algorithm to handle them. So the objective of this paper is to generate some answers from the point of view of a practitioner by considering three critical issues: incompatibility, non-stationary and assessment.

Source of dataset used by the article: The dataset used in this article is a genuine charge card dataset which they have got from their modern accomplice and that is an installment specialist co-op situated in Belgium. This dataset holds the logs of a subset of exchanges from the first of February 2012 to the twentieth of May 2013.

Algorithm/Method used by the article: Neural systems [12], Rule-based strategies (BAYES [13], RIPPER [14]), Tree-based calculations (C4.5 and CART), RF, SVM, NNET, inspecting technique (Under, SMOTE, Easy Ensemble, Incremental methodology (Static, Update, Forget).

Result: The paper exhibits the fraud discovery issue and proposes AP, AUC and Precision Rank as right execution measures for an extortion recognition task. The last best system executed the overlooking methodology together with Easy Ensemble and day by day update.
Future work: The programmed determination of the best unequal strategy on account of web based learning [2].
Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study.

Objective of the paper: Maximum time credit card is an easy way for fraud which takes short time and less risk. In this paper, supervise machine learning algorithm is used on real world datasets for detecting credit cards fraudulent transaction. Credit card datasets are very imbalanced dataset because it carries more allowed fraudulent transactions. The main goal of this paper is find out the accuracy and check the performance for the supervised machine learning algorithm.

Tools/Platform they used to analysis: In this paper, they applied ten machine learning models and compare their Accuracy, TPR, FPR, G-mean, Recall, Precision, Specificity and F1-Score. All machine learning algorithm is used for identity fraud or non-fraud transaction. The main purpose of this paper is to apply supervised machine learning algorithm to real world data sets.

Algorithm/Method used by the article: Supervised and unsupervised machine learning, imbalanced data, Fraud Detection Classifier.
Result: Described all models are giving better result in overall performance. So, top ten features can be used to find-out the accuracy, Recall, Precision, Confusion matrix and compare it to the old result.

Future work: In future they apply voting classifier and compare performance with other machine learning algorithms. They are also thinking to increase the training and testing dataset. Later on they use all the learning algorithms of machine learning and try to find out one of the best outcome [3].

Applying simulation to the problem of detecting financial fraud. Objective of the paper: This thesis is for applying a monetary reproduction of two economic estate, financial payment and retail stores systems. Because in every transaction there is a big problem called fraud, which can cause a failure in the economy. But because of the lacking of transaction data there is a poor amount of experimentation in fraud detection. So, the ultimate objective of this research is to apply a simulation in the detection of fraud and its application in the economic services. But as there is a lacking of real data, so they developed two simulators like financial payment simulator (PaySim) and retail store simulator (RetSim) for generating synthetic transactional data and this data present both normal customer behavior and fraudulent behavior. They are also working on another simulator called *Banksim*which can be used for detecting money laundering cases.

The principle objective of building up these test systems is that it creates and share sensible and various fraud information with the exploration network.

Existing similar works & their objectives:

The work by Gaber et al. presents another comparative procedure to produce manufactured logs for misrepresentation recognition.

Episode Response Sim by Gorton is a reenactment instrument to help the appraisal of danger of web based financial administrations.

The work by Rieke et al., Zhdanova et al. on fraud recognition in portable

Zhdanova et al. is a continuation of the work done by Rieke et al. and utilizes the test system created by Gaberetal. To assess the outcomes.

Malekian and Hashemi dealt with a fraud detection technique that handles the idea float on e-installments

Alexandre and Balsa present a technique to identify extortion utilizing wise specialists that play out the errands that physically a security officer ought to do without anyone else over a restricted measure of information.

Source of dataset used by the article: The dataset they have used for their research is a real transactional dataset which they have got from their exploration partner. This dataset was adjusted to coordinate the conduct of staff and clients utilizing accumulated exchanges from a store of one of the greatest shoe retailers in Scandinavia (Paper I).

Tools/Platform they used to analysis:

1. Retail Store Simulator (RetSim).
2. Mobile money Payment Simulator (PaySim).
3. Multi-Agent Based Simulation toolkit, called MASON.

Algorithm/Method used by the article: Verification and Validation.

Result:

Quantification and measurement of the quantity of loses committed by their noxious operators, this is particularly useful for estimating the expense.

Effectiveness of threshold detection**.**

Future work:

1. With the help of real data we want to improve the accuracy of the payment simulator PaySim.
2. Identifying complex kinds of frauds, for example, illegal tax avoidance
3. Modeling and improving BankSim by accessing real data sets.
4. Developing a multi-simulator by coordinating all three simulators that shares a typical reference to clients and can monitor the exchanges of a solitary operator over all test systems [4].

A Comprehensive Survey of Data Mining-based Fraud Detection Research.

Objective of the paper: The main objective of this paper is to identify challenges in different types of large data sets and streams. Then categories, compares and summaries relevant data mining-based fraud detection methods. This survey paper has been sampled by the last 10 years review paper articles. And also compare all related reviews on fraud detection which helps to take proper decision about FFD.

Source of dataset used by the article: Though this is a survey paper so they discussed about the different dataset used on the papers and analyses their attributes. They select four types of fraud (telecommunications, credit card, and insurance, internal) and made two.

Algorithm/Method used by the article: There are different types of algorithm mentioned in this paper which is used in many papers. They discussed which technique or method is given better result. They mentioned about four approaches,

1. Supervised Approaches on Labelled Data:
    - The neural network and Bayesian network
    - Decision trees, rule induction, and case-based reasoning have also been used
    - The cross validated decision tree
    - Two-stage rules-based fraud detection system
    - Case-based reasoning
    - Statistical modelling such as regression
2. Hybrid Approaches with Labelled Data
    - Supervised hybrid
    - Supervised/Unsupervised hybrid
3. Semi-supervised Approaches with Only Legal (Non-fraud)
    - Kim et al (2003) applied in five steps fraud detection method on a novel
4. Unsupervised Approaches with Unlabeled Data
    - Applied unsupervised neural network method
    - Use cluster analysis for outlier detection, spike detection, and other forms of scoring
    - Peer group analysis for inter account behavior
    - Point analysis for inter account behavior over time
    - Experimental real-time fraud detection system based on a Hidden Markov Model (HMM).

Existing similar works & their objectives: In this paper they applied different type of algorithm to find out fraud detection fraud such as credit card and telecommunications, and related domains such as money laundering and intrusion detection. Then outline techniques from credit card, telecommunications, and intrusion detection. Next neural networks, recurrent neural networks and artificial immune systems for fraud detection Result: This survey paper has covered almost all related studies about fraud detection. All types of fraud, methods and techniques are discussed here. After discovering the limitations about methods and techniques of fraud detection, this paper shows us that this field can benefit from other related fields.

Future work: In future work we want to work on the credit application fraud detection [5].

A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers.

Objective of the paper: Credit scoring and behavioral scoring are the two procedures by using which associations take decisions about to allow or to not allow the credit to shoppers who appeal to them in the hope of getting credit The objective of this review is to give an outline of the targets, systems and difficulties of credit scoring as an application of estimating. It also defines the method of changing the systems from determining the possibility of a buyer defaulting to deciding the profit a buyer will prompt the loaning association. It additionally brings up how effective has been this under-inquired about zone of determining financial hazard.

Tools/Platform they used to analysis: Demo-graphically based segmentation tool, graphical network tools.

Algorithm/Method used by the article: Algorithm/Methods used in Credit scoring:

1. Linear regression
2. Recursive partitioning algorithm
3. Logistic regression and classification trees
4. Neural systems
5. Expert frameworks
6. Genetic calculation
7. Nearest-neighbor techniques.

Algorithm/Methods used in Behavioral scoring:

1. Bayesian Method.
2. Markov chains.

Algorithm/Methods used in Profit scoring:

1. Proportional hazards models.
2. Accelerated life models.

Result: Credit and behavioral scoring are the absolute most significant divining systems utilized in the retail and buyer finance territories As an unadulterated diving instrument instead of a basic leadership one, credit scoring has principally been utilized as a method for anticipating future awful obligation so as to set aside suitable provisioning. With the associations being made between scoring for default and scoring for focusing on potential deals, these scoring procedures will plainly be utilized to figure the offers of items just as the profit an organization will make later on [6].

7.A review of risk in banks and its role in the financial crisis.

Objective of the paper: The objective of this paper is to analysis the role of operational risk in the 2007/2008 financial crisis and to provide recommendations regarding the improvement of operational risk management to assist in the prevention of future crises.

Source of dataset used by the article: The dataset used by the article Esterhuysen at al. (2010).

Algorithm/Method used by the article: credibility theory, Value at Risk (VaR),

Peak over threshold (POT), Hill's method.

Result: This research describes the 2007-8 financial crisis and Role of operational risk in the financial crisis and how should we act in any financial task. It also tells how we can improve operational risk management [7].

2.3 Scope of the Study

Most of the papers we have read through our research have given only the review of some other papers. They didn't have done experimental work and didn't apply any algorithm to reach any decision. But in our research, we have directly applied algorithm to classify the imbalance ta to detect fraud in financial money transaction. We have determined the accuracy of test set to make decision further on new data and visualize a decision tree that will help a lot to detect fraud and to prevent fraud.

2.4 Challenges

We tried to find real transactional data for our research of fraud detection, but we did not find any real data because as a matter of confidentiality transactional data are not available. So we started working with synthetic dataset. But the dataset is so huge that we have faced many problems to work with that dataset. At first, we tried to work with Weka (Waikato Environment for Knowledge Analysis is a suite of machine learning

software written in Java). But as the dataset is huge Weka could not load the whole dataset. Then we started working with python. When we tried to visualize our decision tree, the tree was so huge that it was not possible to display clearly. After that, we tried to discretize all the attributes whose values are numeric but faced problems too because of the huge size of the dataset.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Classify imbalance data for Fraud detection in financial payment system is a well-known problem. It's growing up day by day. Financial payment system is very much needed in developing countries where banking systems are not so much available. In many review papers researchers talked about different types of fraud and applied different algorithms in their individual research purpose. For our research we collected synthetic datasets from the source of Kaggle. There are ten types of attributes. We use some data as training set and some for testing set. We used python software. Then applied transformation and reduction methods for the sake of our work. We applied classification for decision tree algorithm.

3.2 Research Subject and Instrument

The title of our research is "Imbalance Data Classification To Identify Fraudulent Transaction" In this paper we use python as a programming language, 'Scikit-learn' a free software machine learning library for the Python programming language to implement our algorithm, 'Pandas' a software library written for the Python programming language for data manipulation and analysis and we use Jupyter Notebook as a platform.

3.3 Data Collection Procedure

Source of dataset: The source of dataset from Kaggle. Kaggle is a place that based on machine learning. Here's a discussion of data mining, datasets, data science along with machine learning. Many times, we get the training dataset and testing data set from the Kaggle to show us the kernel. Our dataset is a synthetic dataset that used for financial fraud detection. In the datasets the total fraud transaction is 7200. And not-fraud transaction is 587443. The size of datasets is 46.7 MB.

Describing different attributes: In this dataset there are ten types of attributes. There are given bellow:

Step: Step refers to the total number of data that can be passed through one medium in a single hour. It's like a unit of time.
customer: This attribute represents the client who began the exchange/ transaction.
age: This attribute represents the age of the clients.

gender : It represents that the client is male or female or third gender.
zipcodeOri: This represents the area of customer.
merchant: It represent that who is receiving the money.
zipMerchant: This attribute represents the money transection area code.

Category: This represent the type of payment.

Amount: Amount column presents the amount that the customer is going to transac.

fraud: This attribute identifies whether the transaction is fraudulent or not. It contains 0 and 1. 1 represent fraud and 0 represent not fraud.

Size of the dataset: This dataset contains huge amount of data. Total 594643 data are present here. And total number of fraud 7200 and total number of not fraud 587443
Pre-processing: Data preprocessing is used to simplify the process of data processing. The main target of data processing is to find out the target or knowledge.
It is seen that there is a huge amount of data, there are some things we do not need, so data preprocessing is required. Besides, there are many problems in the data such as the data is inconsistent, incomplete and noisy.

Noisy means that data that you want to process in not accurate, bears miss information and is not complete. Data incomplete means I am adding a feature to forty students in my class, from where 40 students will have the first name last name. Many have seen the first name but did not give the last name and many have given the last name but did not give the first name. Due to this the data feature is not full.so data is incomplete.

Due to these reasons data preprocessing is required. Data preprocessing is our data mining fastener. And the algorithm works well.

There are four steps of data preprocessing. These are cleaning, integration, transformation and reduction. We have used only two pre-processing steps for our work convenience. Those are transformation and reduction.

Transformation: Transformation means to transform data from one format to another format.

Normalization is another part of transformation. There are three types of normalization.

There are min-max normalization, Z-score normalization and decimal scaling normalization.

But in our dataset, we have to convert categorical data into numeric data using standard spreadsheet model.

For our datasets we used transformation for an attribute. 'Type' attribute is changed from categorical to numeric form. Though the dataset is synthetic, so the data is clean and there is no missing value as well. That's why we didn't have to do more preprocessing.

Reduction: Data reduction is transformation technique which create ordered or simplified form of meaningful data that derive from multitudinous amount of data.

For the sake of our work we have dropped some columns.

3.4 Methodology

In this part we briefly discuss four well established and popular methods that used for our dataset.

Decision Tree **:** A decision tree is similar to a stream. It represents structure like a tree. In this tree we test some data and that test may have more than one result. This test is usually done on the attributes. It contains a root hub, branches and leaf hubs. Inward hub (non - leaf hub) presents test on characteristics, branches represent out- come of the last, leaf node holds a class label.

When we have multiple candidate first split at that time, there are multiple methods that one could use. Two well-known multiple methods are Information gain and Entropy. Entropy means disorder in a system. In a particular node all values are positive or all values are negative, that is represent all examples are the same class at that time entropy

is 0 or entropy is low. On the other hand, if the half values are positive and the half value is negative at that time entropy is highest.

When we choose the most useful attribute at that time one of the most useful criteria is information gain. Gain is measure how we reduce uncertainty (values lies between 0 and 1).

Let the arrangement of the examples S (preparing information) contains components p and n. p and n is from class P (Fraud=1) and N (Not Fraud=0) respectively.

The measure of data, expected to choose if a discretionary model in S has a place with P or N is characterized as far as entropy, I (p, n)

$$I(p,n) = -Pr(P)\log_2 Pr(P) - Pr(N)\log_2 Pr(N) \quad (1)$$

Where $Pr(P) = p/(p+n)$ and $Pr(n) = n/(p+n)$

I= Information Gain, Pr= Probability.

Information gain: If Si contains pi cases of P and in cases of N, the entropy, or the expected information needed to classify objects in all sub trees Si is

$$E(A) = \sum_{i=1}^{v} pr(Si)\, I(pi, ni) \quad (2)$$

Where $pr(Si) = \frac{si}{si} = \frac{pi+ni}{p+n}$

E= Entropy.

$$Gain(A) = I(p,n) - E(A) \quad (3)$$

Where A=Attribute.

Artificial Neural Network: ANN full format is Artificial Neural Network. This is computational models. ANN is stimulated by biological neural networks. It is used in generally unknown functions. ANN (Artificial Neural Networks) OR NN (Neural Networks) for solving a variety of problems in different fields of science and engineering it provides an exciting alternative method. It is broadly connected in classification and clustering. The neural network is not just an algorithm. It is also known as a framework. ANN is used to work together for a different machine learning algorithms and process complicated data inputs.

Working steps of artificial neural networks:

Start

- Read Dataset
- Encipher, the dependent variable
- Split the dataset into two parts (training and test)
- Tensor flow data structure or holding features, labels, etc.
- Implement model
- Train the model
- Reduce MSE
- Make prediction on the test set
- End

Support Vector Machine: SVM means Support Vector Machine. It is a supervised learning method. SVM use associated learning algorithms that analyze data used for classification and regression and other or outlier detection. So SVM is a supervised machine learning methods that looks at data and sorts it into one of the two categories. Support Vector Machine is one of the most effective classifiers which have sort of linear. It has a very good mathematical intuition behind the support vector machine, and we are able to handle certain cases where there is non-linearity by using non-linear basis functions or in particular we will see these are called kernel functions.

We will see that support vector machine have a clever way to prevent over fitting. And we can work with a relatively larger number of features without requiring too much computation

Working steps of SVM algorithm

- Prepare and format dataset
- Normalized Dataset
- Select activating functions
- Optimize parameters and using search algorithms after the cross validation
- Train the SVM network
- Test SVM network
- Evaluate model performance

Adaboost: There are lots of boosting algorithms. But the most popular is Adaboost, where the weak classifiers are decision trees. In this work, we use the AdaBoost. M1 with DT and BPN as weak classifiers.

In more detail the AdaBoost. M1 is working as follows: For a training set

$TS_n = \left[\left(x_{1,}y_1\right) \ldots \ldots (x_n, y_n)\right]$, with labels $yi \in Y = [1, \ldots \ldots, 1]$,

a weight $wr(i) = \frac{1}{n}$

Is initially assigned to every observation. These weights are recomputed, afterward, according to weak classifier achievements.

Iteratively, for r=1,…..,k, a weak classifier Cr(x) is trained on TSn in order to minimize the following error

$$Er = \sum_{i=14}^{n} wr(i)\, I(Cr(xi) \neq yi) \quad (4)$$

Where I is the indicator function, equal to one when its argument is true, zero otherwise. After r iterations the weights are initially updated as follows:

$$Wr + 1(i) = wr(i) exp\left(arI(Cr(xi) \neq yi)\right) \quad (5)$$

Where,

$ar = .5ln\left(\frac{1-er}{er}\right)$ And $er = \frac{Er}{\sum_{i=1}^{n} wi(i)}$

After the initial update the weights are re-normalized. The final-boosted classifier is

$Cfinal(x) = argmax \sum_{r=1}^{k} arI(Cr(xi) = j)$, j2Y

The major strength of Adaboost:

- Adaboost is capable to convert weak classifier into strong classifier
- In Adaboost, weak learners are decision trees with single split.
- It works by putting more weight on difficult to classify instances.
- It can handle both classification and regression problem.

The following figure shows basic methodology

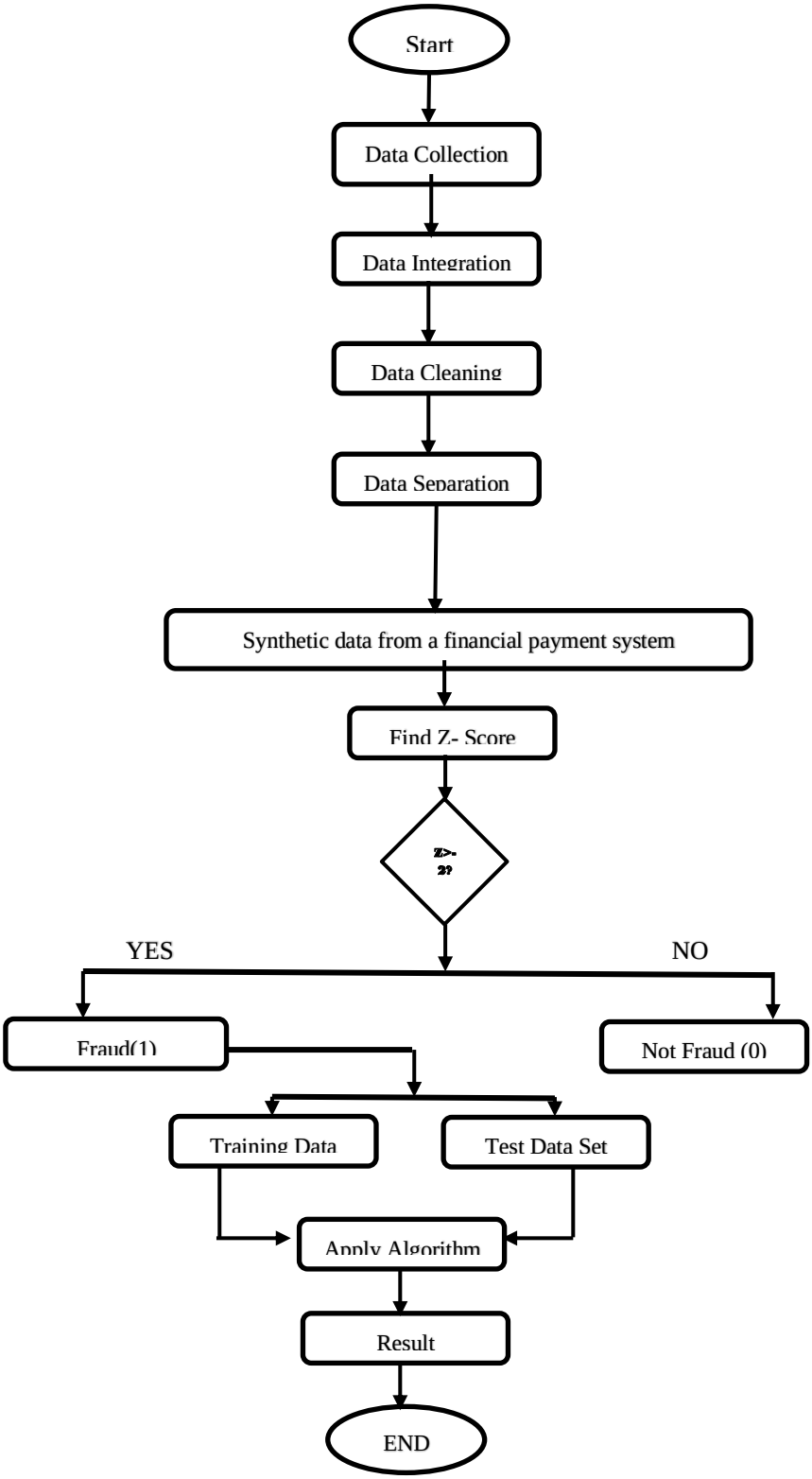

Fig 3.4.1 Methodology

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Result

Confusion Matrix:

Confusion matrix represents a table which is applied on a test data set to analyze the performance of a classification model or classifier. There are two parts of confusion matrix one is predicted part and another is actual part. Actual values represent which value is true from the previous stage and predicted value represents after experiment or observation we have to say something as like as value true or false.

Key matrix:

$$Accuracy = TP+TN/(TP+TN+FP+FN)$$

$$Recall = TP/(TP+FN)$$

$$Precision = TP/(TP+FP)$$

Where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

The classification accuracy measured the proportion of cases correctly classified sensitive measured the fraction of positive cases that were classified as positive, whereas the function of negative cases that were classified as negative are measured in specificity. The algorithm's predictive performance gets better when the values of these statistics gets higher. Confusion matrix of Decision Tree is shown below:

TABLE 4.1.1: DECISION TREE PERFORMANCE BASED ON CONFUSION MATRIX

	Precision	Recall	F1-score	Support
NO	.99	.99	.99	2818
YES	.60	.55	.57	53

From trainee data 99% precision and 99% recall in No.

Where 60% precision and 55% recall in Yes.

That means for non-fraud data in precision 99% of instances that the classifier predicted as fraud that are actually fraud and in recall 99% of fraud instances that the classifier predicted correctly as fraud.

In another hand for fraud data in precision 60% of instances that the classifier predicted as fraud that are actually fraud and in recall 55% of fraud instances that the classifier predicted correctly as fraud.

The accuracy of SVM is better than Decision Tree:

TABLE 4.1.2: SVM PERFORMANCE BASED ON CONFUSION MATRIX

	Precision	Recall	F1-score	Support
NO	.99	1	.99	2812
YES	.91	.52	.66	56

From trainee data 99% precision and 100% recall in No.

Where 91% precision and 52% recall in Yes.

For non-fraud data in precision 99% of instances that the classifier predicted as fraud that are actually fraud and in recall 100% of fraud instances that the classifier predicted correctly as fraud And for fraud data in precision 91% of instances that the classifier predicted as fraud that are actually fraud and in recall 52% of fraud instances that the classifier predicted correctly as fraud.

The accuracy for ANN

TABLE 4.1.3: ANN PERFORMANCE BASED ON CONFUSION MATRIX

	Precision	Recall	F1-score	Support
NO	.99	.99	.99	2815
YES	.73	.70	.71	57

From trainee data 99% precision and 99% recall in No.

Where 73% precision and 70% recall in Yes.

Which is better than both SVM and Decision Tree.

For non-fraud data in precision 99% of instances that the classifier predicted as fraud that are actually fraud and in recall 99% of fraud instances that the classifier predicted correctly as fraud And for fraud data in precision 73% of instances that the classifier predicted as fraud that are actually fraud and in recall 70% of fraud instances that the classifier predicted correctly as fraud.

The accuracy for Adaboost

TABLE 4.1.4: ADABOOST PERFORMANCE BASED ON CONFUSION MATRIX

	Precision	Recall	F1-score	Support
NO	1	1	1	2824
YES	.94	.71	.81	57

From trainee data 100% precision and 100% recall in No.

Where 94% precision and 71% recall in Yes.

For non-fraud data in precision 100% of instances that the classifier predicted as fraud that are actually fraud and in recall 100% of fraud instances that the classifier predicted correctly as fraud And for fraud data in precision 94% of instances that the classifier

predicted as fraud that are actually fraud and in recall 71% of fraud instances that the classifier predicted correctly as fraud.

Here below all of algorithms accuracy and precision, recall and F1 score are represent in bar chart:

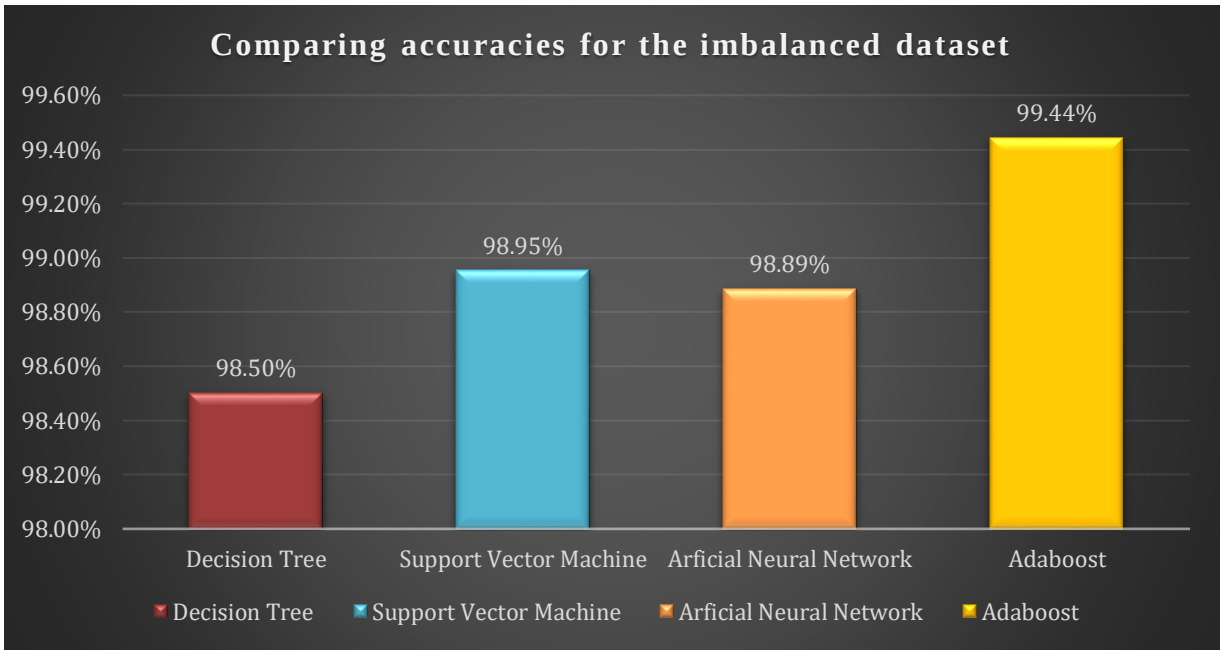

Fig4.1.5. Accuracy for DT, SVM, ANN & Adaboost Algorithms

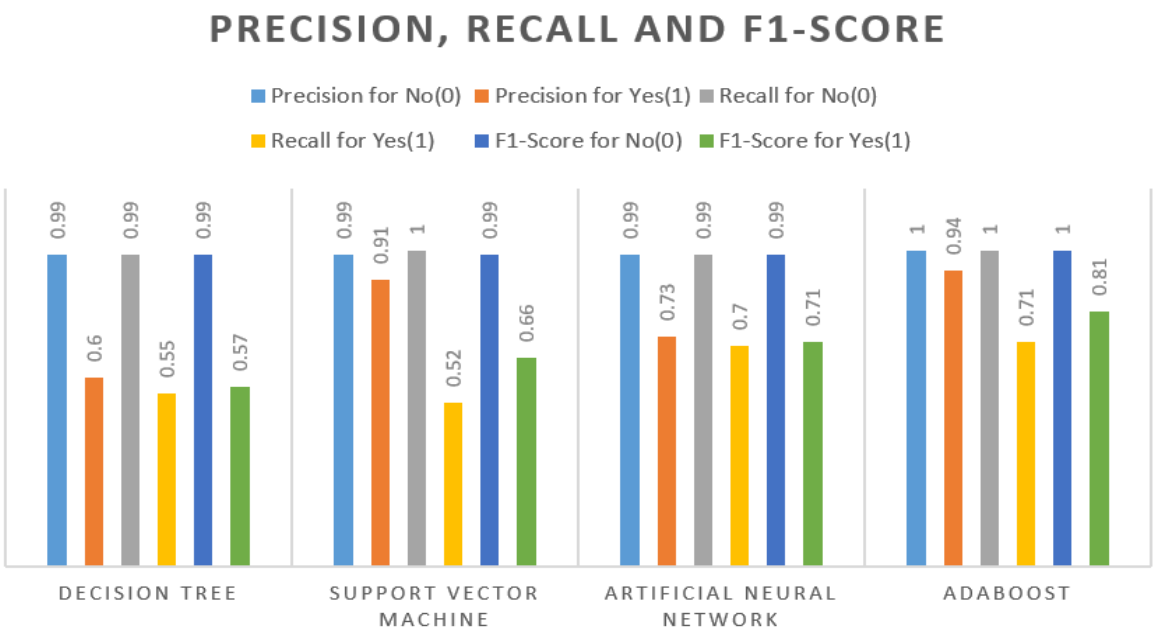

Fig4.1.6. Precision, recall & F1 score

4.2 Descriptive Analysis

For our research we applied four methods of classification algorithm. These are decision tree, support vector machine and artificial neural network and adaboost. Among those four methods we try to find out the best accuracy and executing time. Our total data is 594643. This data type is synthetic. It's a huge amount of data. This large amount of data run very tough and it takes more time to executing. Firstly we use 9571 data for test from 594643. Here the amount of fraud data is 162 and not fraud 9409. And also use 6699 data for training set (70% of taken data). Here fraud is 109 and not fraud data is 6590. And the rest of 30% data is used as a test set.

CHAPTER 5

CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Conclusions

Fraud in financial payment system we understand cash-in, cash-out, mobile research, national and international transfer, bill payments etc. Fraud in financial payment system is increasing day by day. If we do not take action now, then it will have a huge impact. Because of these billions of dollars go to the hands of fraudsters. This research gives us the idea which type of transaction is fraud transaction in financial payment system

5.2 Implication for Future Study

For any kind of research in data mining the first challenge is collecting data. Ours is not different. We have to face many problems for collecting data. For example, privacy issues or organizations rules etc. So, we have to go for synthetic data. But in future we will do research on the real-world data. It will help to predict the real fraud and its behavior in the financial service.

REFERENCES

[1] Sun, Yanmin, Mohamed S. Kamel, Andrew KC Wong, and Yang Wang. "Cost-sensitive boosting for classification of imbalanced data." *Pattern Recognition* 40, no. 12 (2007): 3358-3378.

[2]Ahmed, Sajid, FarshidRayhan, Asif Mahbub, Md Rafsan Jani, SwakkharShatabda, and Dewan Md Farid. "LIUBoost: Locality Informed Under-Boosting for Imbalanced Data Classification." In *Emerging Technologies in Data Mining and Information Security*, pp. 133-144. Springer, Singapore, 2019.

[3] Subudhi, Sharmila, and SuvasiniPanigrahi. "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud." In *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 528-531. IEEE, 2018.

[4] Brown, Iain, and Christophe Mues. "An experimental comparison of classification algorithms for imbalanced credit scoring data sets." *Expert Systems with Applications* 39, no. 3 (2012): 3446-3453.

[5] Sun, Yanmin, Mohamed S. Kamel, Andrew KC Wong, and Yang Wang. "Cost-sensitive boosting for classification of imbalanced data." *Pattern Recognition* 40, no. 12 (2007): 3358-3378.

[6] Dhankhad, Sahil, Emad Mohammed, and Behrouz Far. "Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study." In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 122-125. IEEE, 2018.

[7] Xuchun Li, Lei Wang, Eric Sung, "Adaboost with SVM-based component classifiers", Engineering Applications of Artificial Intelligence 21 (2008) 785–795.

[8] Thanathamathee, Putthiporn, and ChidchanokLursinsap. "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques." *Pattern Recognition Letters* 34, no. 12 (2013): 1339-1347.

[9] Brown, Iain, and Christophe Mues. "An experimental comparison of classification algorithms for imbalanced credit scoring data sets." *Expert Systems with Applications* 39, no. 3 (2012): 3446-3453.

[10] Thammasiri, Dech, DursunDelen, PhayungMeesad, and Nihat Kasap. "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition." *Expert* Systems with Applications 41, no. 2 (2014): 321-330.

[11] Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. "Learning from class-imbalanced data: Review of methods and applications." *Expert Systems with Applications* 73 (2017): 220-239.

[12] Hassan, Amira Kamil Ibrahim, and Ajith Abraham. "Modeling insurance fraud detection using imbalanced data classification." In *Advances in Nature and Biologically Inspired Computing*, pp. 117-127. Springer, Cham, 2016.

[13] Sawant, Abhijit A., and P. M. Chawan. "Comparison of Data Mining Techniques used for Financial Data Analysis." *International Journal of Emerging Technology and Advanced Engineering* (2013).

[14] Li, Jianping, Jingli Liu, Weixuan Xu, and Yong Shi. "Support vector machines approach to credit assessment." In *International Conference on Computational Science*, pp. 892-899. Springer, Berlin, Heidelberg, 2004.