Telecom Voice Traffic Termination Fraud Detection Using Ensemble Learning: The Case of Ethio Telecom

A Thesis Presented

By

Alemeshet Getahun Kassa

To

The Faculty of Informatics

Of

St. Mary's University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

In

Computer Science

July 27, 2020

ACCEPTANCE

Telecom Voice Traffic Termination Fraud Detection Using Ensemble Learning: The Case of Ethio Telecom

By

Alemeshet Getahun Kassa

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Dr. Asirat Mulatu

_____

Internal Examiner

Dr. Temtim Assefa

_____

External Examiner

Dr. Getahun Semeon

_____

Dean, Faculty of Informatics

July 27, 2020

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Alemeshet Getahun Kassa
Student

_____
Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Dr. Getahun Semeon
Advisor

_____
Signature

Addis Ababa

Ethiopia

June 2020

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Almighty GOD and His Mother St. Virgin Marry for all happened in my life and for giving me my lovely daughter Hildana Alemeshet while I was studying this postgraduate program.

Secondly, the success of this thesis is credited to the extensive support and assistance from my advisor Dr. Getahun Semeon. I would like to express my grateful gratitude and sincere appreciation to him for his keen insight, guidance, unreserved advising, constructive comments, encouragement and kindness to me throughout this study. Thank you so much!

Thirdly, I would like also to express my gratitude to all my instructors whose contribution helped me to succeed on this study.

Fourthly, I would like to thank officials and employees of Ethio Telecom for their supportive, genuine and elaborated explanation even much better than the questions asked and providing me appropriate professional comments and for their cooperation to give me valuable resources which are relevant for my study.

Finally, I would like to thank all my families and friends for the constant assistance and encouragement during the time of postgraduate program.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv
LIST OF ACRONYMS vii
LIST OF TABLES x
ABSTRACT xi
CHAPTER ONE 1
INTRODUCTION 1
1.1. Background 1
1.2. Statement of the Problem 4
1.3. Research Questions 5
1.4. Objective of the Study 5
1.4.1 General Objective of the Study 5
1.4.2 Specific Objectives of the Study 5
1.5. Methodology 6
1.6. Scope and Limitation of the study 9
1.7. Significance of the study 10
1.8. Thesis Organization 10
CHAPTER TWO 12
REVIEW OF LITERATURE 12
2.1 Introduction 12
2.2 Fraud in Telecommunication Industry 12
2.2.1 What is Telecom Fraud? 12
2.2.2 Telecoms Fraud Types 12
2.2.2.1 Interconnect bypass fraud 13
2.2.2.2 Subscription Fraud 13
2.2.2.3 Premium rate Fraud 14
2.2.2.4 Roaming Fraud 14
2.2.2.5 PBX Fraud 14
2.2.2.6 Prepaid Fraud 15
2.2.2.7 Cloning fraud 15
2.2.3 Difficulties in Detecting Fraud 16

2.3 Review of Data Mining and Knowledge Discovery (DMKD) 16
2.3.1 Basic Concepts 16
2.3.2 Data Mining Tasks 18
2.3.3 Classification learning algorithms (individual classifiers) 20
2.3.4 Ensemble Classification Methods 22
2.3.4.1 Bagging (Bootstrap aggregation) ensemble method 22
2.3.4.2 Boosting (AdaBoost) ensemble method 22
2.3.4.3 Stacking (Stacked generalization / Blending) method 23
2.3.4.4 Voting ensemble method 23
2.3.5 Data Mining Process Models 24
2.3.5.1 The KDD Process 24
2.3.5.2 CRISP-DM Process Model 26
2.3.5.3 The SEMMA Process Model 27
2.3.5.4 Hybrid Process Model 29
2.4 Review of Related Research Works 33

CHAPTER THREE 38
RESEARCH METHODOLOGY 38
3.1 Introduction 38
3.2 Research methods 38
3.3 Research design 38
3.3.1 Understanding of the problem domain 39
3.3.2 Understanding of the Data 41
3.3.2.1 Data source and Data collection 41
3.3.2.2 Description of Raw Data Quality 42
3.3.3 Preparation of the Data 46
3.3.3.1 Data Selection 46
3.3.3.2 Attribute Selection 47
3.3.3.3 Data Cleaning 49
3.3.3.4 Data Construction and Transformation 50
3.3.3.5 Data Formatting 50
3.3.4 Data mining for building predictive model 51

3.3.5 Evaluation of the discovered knowledge 52
3.3.6 Use of the discovered knowledge 53

CHAPTER FOUR 54

EXPERIMENTATION AND RESULT ANALYSIS 54
Introduction 55
4.1 Model building experiment using all attributes with ensemble methods 55
4.1.1 Model Building Experiment Using AdaBoost (Boosting) Method 55
4.1.2 Model Building Experiment Using Bagging Method 57
4.1.3 Model Building Experiment Using Stacking Method 61
4.1.4 Model Building Experiment Using Voting Method 61
4.1.5 Model building experiment using all attributes without Ensemble 63
4.3 Evaluation and comparison of the algorithm 65
4.4 Discussion of the results with domain experts 68
4.5 Specific Rule Extraction 69
4.6 Using Discovered Knowledge 72
4.7 Validity of user acceptance testing 72

CHAPTER FIVE 74

CONCLUSIONS AND RECOMMENDATIONS 74
5.1 Conclusions 74
5.2 Recommendations 75

REFERENCES 76

ANNEXES 81

List of acronyms

AdaBoost	Adaptive booster
EMs	Ensemble Methods
ETC	Ethiopian Telecommunication Corporation
CDR	Call Detail Record
CFCA	Communications Fraud Control Survey
WEKA	Waikato Environment for Knowledge Analysis
PART	Partial Decision Tree
MLP	multi-layer perceptron
CSV	Comma Separated Values
ANNs	Artificial Neural Networks
CRISP-DM	Cross-Industry Standard Process -Data Mining
SEMMA	Sample, Explore, Modify, Model, Assess
KDD	Knowledge Discovery in database
FMS	Fraud Management Systems
TCG	Test Call Generator
PRF	Premium Rate Fraud
PBX	Private Branch Exchange
SIM	Subscriber Identity Module
VoIP	Voice over Internet Protocol
ARFF	Attribute Relation File Format
GSM	Global System for Mobile Communication

LIST OF FIGURES

Fig 1. 1 The Hybrid Process Models (description of the six steps follows) 7
Fig 2. 1 Data mining: confluence of multiple disciplines [31] 17
Fig 2. 2 Data mining tasks and models [30].....18
Fig 2. 3 A Scenario that shows how decision tree is constructed 20
Fig 2. 4 The five stages of KDD (source: [42].....25
Fig 2. 5 CRISP-DM process model (source: [43].....26
Fig 2. 6 the Schematic of SEMMA 28
Fig 2. 7 The Hybrid Models (description of the six steps follows) (source: [45].....30
Fig 4. 1 Snapshot showing when first time data is loaded to WEKA tool 55
Fig 4. 2 Screen shoot of prototype developed for the rule generated 72

LIST OF TABLES

Table 2. 1 Summary of Ensemble methods in machine learning (compiled by the author) ..23
Table 2. 2 Summary of KDD, CRISP-DM, SEMMA and hybrid processes model [45].....31
Table 3. 1 List of attributes in the initial dataset with derived attributes 42
Table 3. 2 The Final List of Attributes used in this study.....48
Table 4. 1 Detailed performance measures for experiment One 56
Table 4. 2 Detailed performance measures for experiment Two.....58
Table 4. 3 Detailed performance measures for experiment Three.....60
Table 4. 4 Detailed performance measures for experiment four 62
Table 4. 5 Detailed performance measures for experiment four 64
Table 4. 6 Performance comparison of the selected models with ensemble methods) 65
Table 4. 7 Performance comparison of the selected models without ensemble methods 66
Table 4. 8 Confusion Matrix result for Boosted J48 decision tree algorithm with10-fold cross validation testing option 67
Table 4. 9 Evaluation on ensemble and single algorithm of prediction error.....68
Table 4. 10 Experts response summary on the proposed prediction model 73

ABSTRACT

One of the major developments in machine learning is the ensemble method, which finds highly accurate classifier by combining many moderately accurate component classifiers. In this thesis, ensemble classification methods were proposed. This proposed model provides the important information which can be used for decision making. A comparison study was also made for finding the suitable classifier on an ensemble technique used in the proposed model

We selected around 126736 records from two months' collection of call detail record data. After eliminating irrelevant and unnecessary data, a total of 50516 datasets were used for the purpose of conducting this study. The researcher also selected 10 attributes for this study based on their relevant for this research. Data preprocessing was done to clean the datasets. After data preprocessing, the collected data has been prepared in a format suitable for the DM tasks. The study was conducted using Waikato environment for knowledge analysis (WEKA) version 3.8.3 machine learning software and four ensemble based machine learning paradigms for classification techniques was used, namely boosting, bagging, stacking and voting classifiers, based on 2 basic learners (decision tree and neural network) algorithms. The training models are built using cross validation and tested for reliability by default values of percentage split (66%).

The performances of the model in this study were evaluated using the standard metrics of prediction accuracy, error rate analysis, FP rate, TP rate, recall, precision, F-measure and ROC curve which are calculated using the predictive classification table, known as Confusion matrix. Comparison of the performance of each algorithm made to select the algorithm with best performance. The results of the study show that ensemble J48 decision tree algorithm with 10-fold cross validation registered better performance of 96.73%. The boosting classifier provides highest prediction accuracy than the other classifiers. In this study, we found that the proposed ensemble methods provide significant improvement of prediction accuracy compared to individual classifiers.

Keywords*: Ensemble methods, Data mining, Boosting, Bagging, Stacking, Voting*

CHAPTER ONE

INTRODUCTION

1.1. Background

The introduction of telecommunication services in Ethiopia dates to 1894. Ethiopian Telecommunications Corporation is the oldest public telecommunications operator in Africa. In this year, the technological scheme contributed to the integration of the Ethiopian society when the extensive open wire line system was laid out linking the capital with all the important administrative cities of the country. The network has begun to expand starting from then [1]. Starting from November 2011, ETC changed the name to Ethio Telecom and is a government owned sole telecommunications service provider, Telecom Company. It provides voice, Internet and data services to the public through fixed line, mobile network and satellite communication throughout the country. From different services Ethio Telecom provides mobile communication service, which is one of the biggest and main services with more than 60 million of customers, using mobile networks. Globally, the development of telecommunication industry is one of the important indicators of social and economic development of a given country. In addition to this, the development of communication sector plays a vital role in the overall development of all sectors related to social, political and economic affairs. This sector is very dynamic in its nature of innovation and dissemination [2]. Technology advancements resulted in Telecom industry expanded; number of subscribers increased, subscribers demand increased, which increasingly motivated fraudsters, to expend and diversify fraud method and techniques [3].

Telecommunication fraud occurs whenever a person committing the fraud uses deception to receive telephony services free of charge or at a reduced rate. It is a worldwide problem with substantial annual revenue losses for many companies. However, it is difficult to provide precise estimates since some fraud may never be detected, and the operators are not willing to reveal figures on fraud losses. Sometimes they may not have the evidence and the technique to stop the fraud, but they have only the information from different sources. The situation can significantly be worse for mobile operators in Africa for, as a result of fraud, they become liable for large hard currency payments to foreign network operators. Thus,

telecommunication fraud is a significant problem which needs to be addressed, detected and prevented in the strongest possible manner [3]. According to [3] some popular examples of fraud in the telecommunication industry includes subscription fraud, identity theft, voice over the internet protocol (VoIP) fraud, cellular cloning, billing and payment fraud on telecom accounts, prepaid and postpaid frauds and PBX (Private Brach Exchange) fraud etc. Telecom service provider's operations and revenues are highly impacted due to telecom frauds. Large amount of revenue is lost due to telecom voice traffic termination fraud. Among the revenue sources of Ethio Telecom, international traffic takes the lion share of it. As [4] indicated 40% of Ethiopian Telecommunications Corporation, currently Ethio-telecom revenue is from international traffic [4]. Telecom voice traffic termination fraud is one of the fraud types that attack the revenue from international traffic. Telecom voice traffic termination fraud is affecting not only Ethio Telecom but also telecom operators in Africa. It is a system by which fraudsters re-route international calls by using local SIM cards. It is also one of the reasons for telecom operators for losing millions of dollars every year [5].
Industries like telecommunication produce and accumulate huge amount of data. These data comprise call detail data, network data and customer data. Because of its hugeness the data is difficult to analyze manually. Therefore, we need a mechanism to handle this data. As a result, the development of knowledge-based expert systems and automated systems performed important functions such as identifying telecom frauds [6]. Data Mining is also called Knowledge Discovery in Databases (KDD) which is defined as the science of extracting useful information from large datasets or databases. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information, i.e. information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [7].

The telecommunication industry was an early adopter of data mining technology and therefore many applications exist. Four major applications include: marketing customer profiling, fraud detection, churn management and network fault isolation [6]. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into

useful information, i.e. information that can be used to increase revenue, cut costs, or both. Data mining software is one of several analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [7]. Telecommunication companies utilize data mining technique to improve their marketing efforts, identify fraud and better manage their telecommunication network connection [3].

Data mining methods may be distinguished by either supervised or unsupervised learning methods. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. It has been observed that when certain classifiers are ensembled, the performance of the individual classifiers become increased. DM uses many machine learning (ML) methods, including ensemble learning methods. Ensemble methods helps improve machine learning results by combining multiple models. Using ensemble methods allows to produce better predictions compared to a single model [8]. The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. An ensemble learning is a machine learning process to get better prediction performance by strategically combining the predictions from multiple learning algorithms compared to a single model [9].

Recently, ensemble learning has been one of the active research fields in machine learning. Thus, it has been utilized in a very broad range of areas such as marketing, banking, insurance, health, telecommunication, and manufacturing. Ensemble classifiers pool the predictions of multiple base models. Much empirical and theoretical evidence has shown that model combination increases predictive accuracy [10, 11]. Some challenges of telecommunication operators are protecting them from attacks, mitigate fraud problems, minimizing fraud attempt, hardening the telecom ecosystem security to fraudsters, implementing better fraud detection and protection techniques [12]. In this study, the researcher proposed an effective solution to identify voice traffic termination fraud detection in telecommunications using ensemble based machine learning techniques that will have its own contribution for the company.

1.2. Statement of the Problem

Nowadays, even if telecom service providers and regulatory bodies were deploying different fraud management system (FMS) to detect this particular fraud, they still remain a main challenge for service providers. The changing behavior of the frauds from time to time is among the reason for the frauds to remain as a main challenge. Detection techniques for this fraud needs to be continuously assessed and improved so that telecom companies and regulatory bodies can go in line with the changing behavior of the frauds for efficient and effective fraud mitigation [13, 14].

In 2018, an estimated revenue of $4.2 Billion has been lost by telecom companies due to frauds [15]. Ethio Telecom, sole telecom service provider in Ethiopia, is among the telecom companies impacted by voice traffic termination fraud. Hence, voice traffic termination fraud mostly targets international call interconnection or termination fees, not only to Ethio Telecom but the country is also impacted because it is losing revenue in foreign currency. This fraud activity has been increasing dramatically each year due to the new modern technologies and the global superhighways of communication, resulting in the decreasing of the revenue and quality of service in telecommunication providers especially in Africa and Asia [16]. According to [17], telecom fraud is a burning problem, in addition to the increase and widespread of telecom frauds, it is a serious threat for national security beyond economic loss.

Telecom companies should continuously monitor their networks for such mobile fraud and resolve root causes at the very early stages. Currently, not only Ethio Telecom but also other telecom operators in Africa are suffering from such kind of fraud for different reasons due to the rising profile of the continent. This voice traffic termination fraud is practiced in order to avoid the high cost of international call termination tariff [5]. Currently, Ethio Telecom uses two mechanisms to detect the different types of frauds. These techniques were test call generation and fraud management system. Both techniques are used to detect fraud in real time and on daily basis. However, both systems have their own drawbacks. They didn't use intelligent machine learning tools. Hence, to improve prediction performance intelligent machine learning tools and data mining plays tremendous advantage. Beyond this, previous studies were conducted by using single algorithm pattern mining [46-51]. But, using single algorithm cannot improve the performance of the model to identify voice traffic termination

fraud prediction; or single algorithm cannot provide robust model. One of the major developments in machine learning in the past decade is the ensemble method, which finds highly accurate classifier by combining many moderately accurate component classifiers. Therefore, efficient fraud detection systems and analysis systems can save telecom operators a lot of money. In this study, we investigate how the technique proposed in ensemble method, could be adapted to solve a voice traffic termination fraud. Selecting a better ensemble based classification method in machine learning for building a model, which performs best in handling the prediction and identifying with a better performance is the aim of this study. The main purpose of this research is therefore, to build a predictive model that can accurately identify or predict voice traffic termination fraud by using ensemble learning classifiers.

1.3. Research Questions

In order to accomplish the purpose of the research, the following guiding questions or lines of inquiries were listed as follows: -

- *What are the appropriate data preprocessing techniques for building a fraud predictive model?*
- *What are the major attributes to consider in applying ensemble-based data mining technique for voice traffic termination fraud prediction?*
- *How we evaluate the accuracy of fraud predictive model?*

1.4. Objective of the Study

This study consists of general and some specific objectives to be achieved at the end.

1.4.1 General Objective of the Study

The general objective of this study is to investigate the potential applicability of the data mining and machine learning for voice traffic termination fraud prediction using ensemble based classification methods.

1.4.2 Specific Objectives of the Study

To accomplish the above general objective, the study focuses on the following specific objectives:

- To collect and prepare the data for analysis by cleaning and transformation of the data.

- To identify appropriate ensemble based data mining algorithms that are more appropriate to the problem domain.
- To build the fraud predictive model using ensemble.
- To evaluate the model so that classifier that shows better performance in terms of accuracy and speed can be selected.
- Deduce conclusion and recommendations for use by the study organization.

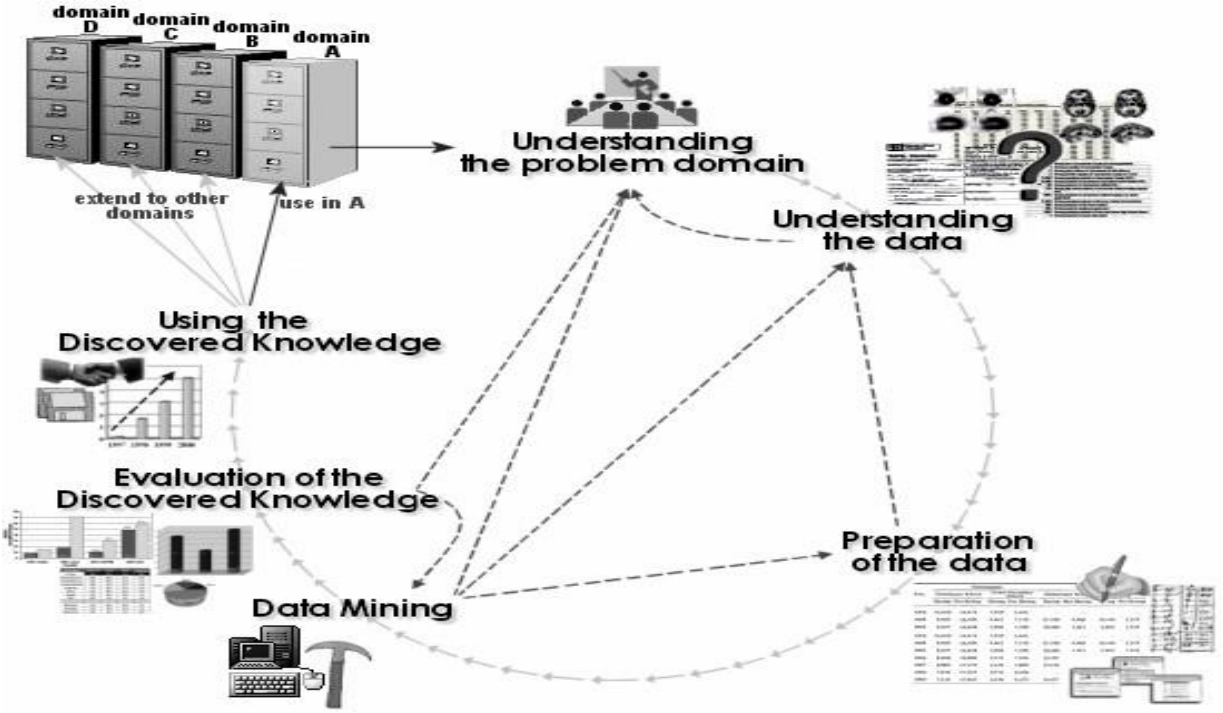
1.5. Methodology

In order to achieve the general and specific objectives of the study the following methods was used.

Research design:

This study follows experimental research. In order to apply this experimental research, we use the six-step process of Hybrid model. This model developed, by adopting CRISP-DM model to the needs of academic research community. Unlike the CRISP-DM process model, which is fully industrial, the hybrid of process model is both academic and industrial. The main extensions of the hybrid of process model include providing a more general, research-oriented description of the step, an introduction of a DM step instead of the modeling step, and an integration of several explicit feedback mechanisms. The overall design issues using Hybrid data mining methodology are represented diagrammatically as shown in figure 1 below.

Fig 1. 1 The Hybrid Process Models (description of the six steps follows)



All the below six steps in this process pursued.

Understanding of the problem domain:

In ordered to identify, define, understand and formulate the problem domain the researcher used different discussion points with domain experts. Those discussion points reflect telecom voice traffic termination fraud, so that the researcher closely works with the domain experts of Ethio Telecom and then determine attribute feature selection and understand some complex business process. In collaborating with the domain experts, the CDR data is selected as the main source of data collection. As a result of insight gained knowledge of the domain of telecom business and data mining problem was defined.

Understanding of the data:

This step is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy and plausibility of attribute values. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

In this research, in order to understand the data, the researcher briefly described the CDR data. The description includes listing out initial attributes, their respective values and the

researcher in collaboration with domain experts of Ethio Telecom evaluation was made on how the CDR is important for this study. The call data record (CDR) was used for mining the hidden patterns. Since Ethio Telecom is the only service provider in the country several transactions have been stored in the database every second. After discussing with the domain experts two-month call detail record along with selected key attributes were used. Data mining is the most relevant tool in such a situation whereas the size of the data set, which is stored in the CDR tables, is very large. It is impossible to find out the pattern of each number manually, so data mining tool has been used to analyze the data.

Preparation of the data:

In this phase the appropriate datasets perfect for the research was chosen. It was then cleaned and formatted in order for it to be fit for use. Data cleansing done to the data included removing duplicate record, correcting noisy data, removing irrelevant variable attributes and records i.e. attributes and records which are out of interest of the data mining problem

In this research, the researcher decides, together with domain experts, the CDR data is used as input for applying the ensemble DM techniques. The cleaned data is further processed by feature selection consulting the domain experts and the Weka attribute selection preprocessing techniques to reduce dimensionality and by derivation of new attributes. The result of these processes generates datasets for training and testing of the classification algorithms selected in this study.

Data mining for predictive modeling:

After the data has been prepared for analysis it is used to build classification models using ensemble learning classifiers. This has enabled to design a better strategy for voice traffic termination fraud detection. This study uses on an ensemble based machine learning paradigm like bagging, boosting, stacking and voting classifiers, based on 2 basic learners, i.e., Decision tree (DT) and Neural network (ANN) were applied in this research.

WEKA was adopted for mining of the data. The WEKA tool was chosen because [46]:

- Its functionality (support for many algorithms)
- Familiarity of the researcher with the software and ease of use (has a graphical user interface) and runs on almost any platform.
- It contains a compressive collection of data preprocessing and modeling techniques
- It is freely available under the General Public License (GNU)

Evaluation of the discovered knowledge:

In this research different ensemble based classification models were developed and evaluated using training and testing dataset. The experimental output of the classification models was analyzed and evaluated the performances accuracy using confusion matrix. After performing the confusion matrix, the results are evaluated by measuring its accuracy and Error rate. Furthermore, the effectiveness and efficiency of the model is also computed in terms of recall and precision. Here in collaboration domain experts of Ethio telecom, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge is evaluated. The performance of the algorithms were also evaluated time parameter: the time taken to build the model.

Use of the discovered knowledge:

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step knowledge discovered is incorporated into performance system and take this action based on the discovered knowledge.

In this research the discovered knowledge is used by integrating the user interface which is designed by VB.NET programming language with a Weka system in order to show the prediction of telecom voice traffic termination fraud based on the extracted rules.

1.6. Scope and Limitation of the study

The scope of this research is to develop detective and predictive model for telecom voice traffic termination fraud in the Ethio telecom service provision. In Ethio Telecom, the current approach is heavily reliant on investigations by expert engineer, this research sets out to apply data analysis techniques that can automate and standardize the detection and description of telecom voice traffic termination fraud, making their work faster and efficient. Today, there are many telecom frauds among those bypass frauds is very complex and transversal to the operator structure.

Hence, this study is limited to voice traffic termination fraud on mobile communication service in Ethio-telecom. Due to time, resource and knowledge, the depth of the research did not include all types of telecommunication frauds. Therefore, further research can be conducted to include the other types of telecommunication fraud. The other limitation of this study is also only two-month data has been used. This is because of Ethio-telecom does not keep data beyond this months.

1.7. Significance of the study

This study is significant to telecom companies because it introduces with the application of the ensemble based data mining and knowledge discovery methods and processes in telecommunications network operations especially to Ethio Telecom that enables it to extract new, potentially useful and novel knowledge from the data repository that will be used to give acceptable decisions in the detection of fraud in the form of proactive prevention technique that should be expected to save the time and increase revenue to the government in terms of hard currency and can achieve its development plan. Moreover, the output of this study helps to improving the efficiency of fraud detection mechanism of the current threat posed on Ethio Telecom service provision.

Accordingly, this study will play an important role to control and prevent the current threat on the mobile communication of Ethio Telecom. The outcome of this research shall also be used as a benchmark for domain expert as well as a source of methodological approach for studies dealing on the application of ensemble data mining on fraud management as well as other similar areas. The results of this research would support the routine and strategic decision making processes of the telecom operators in solving everyday problems on voice traffic termination fraud detection for further decision-making. The other contributions are improvements of a rule extraction techniques, resulting in increased comprehensibility and more accurate result by ensemble machine learning. Generally, this study is important since its application can reduce the revenue losses due to voice traffic termination fraud.

1.8. Thesis Organization

This thesis is organized into five chapters. The first chapter briefly discuses background of the problem area and states the problem, the general and specific objectives of the study, the research methodology, the scope and limitation of the study, and significances of results of the study.

The second chapter presents literature review on data mining, data mining overview, knowledge discovery process models, data mining tasks, classification algorithms and critical literature review on the telecommunication fraud that includes the common types of telecom fraud like premium rate fraud, internal fraud, prepaid fraud, cloning fraud, roaming fraud and PBX. Furthermore, this chapter also reviews local and international related works.

The third chapter presents the different DM steps that are undertaken by the methodology used in this study. This includes problem identification, data understanding, data preparation, model building and evaluating the results and using discovered knowledge are performed by WEKA software.

The fourth chapter deals with experimentations and result analysis in this chapter building of model with training dataset and validating the result with testing datasets and analysis of the result of the experimentation was the major concern. This chapter also shows how to use the final experimental results in real world application by doing prototype user interface.

The last chapter is devoted to concluding remarks and recommendations forwarded based on the research findings of the present study.

CHAPTER TWO

REVIEW OF LITERATURE

2.1 Introduction

In this chapter, the most relevant local and global literature on telecommunication fraud have been reviewed. More specific issues include: the detection and prevention techniques of telecom fraud, methods and algorithms, the DM process and the different tasks of DM.

2.2 Fraud in Telecommunication Industry

2.2.1 What is Telecom Fraud?

Many definitions in the literature exist, where the intention of the subscriber plays a central role. Johnson defines fraud as any transmission of voice data across a telecommunications network, where the intent of the sender is to avoid or reduce legitimate call charges [18].

Fraud is commonly described as the intent of misleading others in order of obtaining personal benefits [19]. In the telecommunications area, fraud is characterized by the abusive usage of and carrier services without the intention of paying.

In legislation, the term fraud is used broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences [3]. Telecommunication fraud occurs whenever a person committing the fraud uses deception to receive telephony services free of charge or at a reduced rate. Fraudsters see themselves as entrepreneurs, admittedly utilizing illegal methods, but motivated and directed by essentially the same issues of cost, marketing, pricing, network design and operations as any legitimate network operator [3].

Fraud covers a wide range of illicit practices and illegal acts that is intentional deception or misrepresentation. It is defined as any illegal act characterized by deceit, concealment, or violation of trust. Frauds are usually committed, by individuals and/or organizations, to secure personal or business advantage through unlawful act to obtain money, property, services or to avoid payment or loss of services [20].

2.2.2 Telecoms Fraud Types

Some of the major reasons why fraudsters commit frauds are: -

- Weaknesses in systems some of operational are easily prune to fraudsters to commit frauds.
- International call termination fee is very expensive compared to local call.
- The fast growth in technology and the fraudsters have varieties of ways in committing frauds

Different types of fraud exist based on the nature of the fraud committed.

According to [21], there are different types of fraud. However, the most common types of fraud are: - Interconnect bypass fraud, subscription fraud premium rate fraud (PRF) internal fraud, prepaid fraud, postpaid fraud, cloning fraud, roaming fraud, private branch exchange (PBX) fraud.

2.2.2.1 Interconnect bypass fraud

The international call comes through the internet and the gateway forward the call to one of the idle mobile numbers. Finally, it takes that number when the call reaches to the called number. Interconnect bypass fraud is device that maps the call from voice over internet protocol (VoIP) to a SIM card of the same mobile operator of the destination mobile. So that international call terminating as home call to subscriber country and usually cheap compared to the cost of terminating the international call. This is to just bypass international traffic. Interconnect bypass fraud also have a negative impact on call quality and create network congestion [22]. According to [22] the purpose of bypassing is making money by illegally terminating traffic into operator's network, without paying the interconnection fee, using VoIP technology. That is usually international traffic. They have contract with a wholesale operator for a determined number of minutes to be terminated via his SIM cards. Traffic is being received via IP and is routed through the fraudster's, SIM gateways, it reaches the destination as a national call. The fraudster will pay the network for a national call but will charge the wholesale operator for every minute he terminated; the network operator loses the interconnection fee.

2.2.2.2 Subscription Fraud

The subscription fraud is the most common since with a stolen or manufactured identity, there is no need for a fraudster to undertake a digital network ‘s encryption or authentication systems which is simply signing up for a service using fake or stolen identification, with no commitment of paying the bill [21].

Subscription fraud is also the most common type of fraud encountered on the GSM network. A person subscribes for a service by using false identification. Then the fraudster may use the service either for personal use or for profit making. According to GSM Association and the Communications Fraud Control Association, subscription fraud is the starting point for many other telecoms fraud and as such is recognized as the most damaging of non-technical fraud types [24].

2.2.2.3 Premium rate Fraud

Premium rate fraud mostly used in combination with other types of fraud, the main reasons are to make free calls to high cost numbers like competition or hot lines, and also to make money from falsely generating calls to a number owned and operated by the fraudster. The more calls generated, the higher the profit to be made [25]. According to [25] the premium rate services fraud currently considered the most financially damaging fraud type in combination with Roaming. This fraud type is international and so boundaries have no relevance in this case. In well-organized attacks the calls are made during weekends, holidays, etc. In order to take advantage of using a bigger time, until the first Roaming High usage report (HUR) is received.

2.2.2.4 Roaming Fraud

Roaming Fraud is the ability to use telecom products or services, such as voice or data services, outside the home network with no intention to pay for it. In these cases, fraudsters use the longer timeframes required for the home network. Roaming fraud can start as an internal or subscription fraud in the home network, when obtained SIM cards are sent to a foreign network [26]. The fraudsters use different fraud methods. Subscription fraud is one of the fraudsters' preferred methods for digital roaming fraud. Roaming fraud can happen when a subscriber that used the services of the visiting network refuses to pay for them either by claiming ignorance, insufficient knowledge of the additional costs, or by claiming that the service was never requested [21].

2.2.2.5 PBX Fraud

A PBX (Private Branch Exchange) is a switch station for telephone systems. It consists mainly of several branches of telephone systems and it switches connections to and from them, thereby linking phone lines. Most medium-sized and larger companies use a PBX for connecting all their internal phones to an external line. This way, they can rent only one line

and have many people using it, with each one having a phone at the desk with different number. There are different types of PBX fraud, all the call pays the companies for including not made for the business of the company. Some of these frauds are carried out by workers of the company and others hacking into the company network [8].

2.2.2.6 Prepaid Fraud

Prepaid fraud is an attempt made by the fraudsters to use free telecommunication services by using lost or stolen credit cards. Internal engineers can access and adjust the billing-activation Systems, stealing cards, and PIN numbers and recharge codes at production and support sites, scan for data from valid phones and duplicate the information onto stolen devices. The prepaid phones appear valid but steal minutes from an honest customer [21]. Credit information is stored in the SIM; it is not very difficult to modify the same illegally. In order to minimize the impact of prepaid fraud, the vendor community is developing techniques such as real time billing and rating systems, combining pre-paid and postpaid systems into one, generation of logs for changes made to Intelligent Network (IN) based systems; activation at point of sale systems; migration to IN based systems [21].

2.2.2.7 Cloning fraud

Cloning fraud is created by reprogram techniques that mean copying the identity of valid mobile telephone to another mobile telephone. During cloning it requires access to electronic serial number and mobile identification number. A cloned mobile phone is a form of reprogram to access to a mobile network as a valid cell phone. The legal phone user then gets billed for the cloned phone‘s calls. Cloning mobile phones is achieved by cloning the SIM card contained within, not necessarily any of the phone's internal data. There are various methods used to obtain the Mobile Identification Number (MIN) and its Electronic Serial Number (ESN); the most common are to crack the cellular company, or listen in on the cellular network.

Whenever a cellular phone is on, it periodically transmits two unique identification numbers: its Mobile Identification Number (MIN) and its Electronic Serial Number (ESN). These two numbers together specify the customer‘s account. These numbers are transmitted unencrypted, and they can be received, decoded and stored using special equipment. Cloning occurs when a customer‘s MIN and ESN are programmed into a cellular telephone not belonging to the customer. The attraction of free and untraceable communication makes

cloned phones very popular in major metropolitan areas [27]. The arrival of new technologies has provided fraudsters new techniques to commit fraud. Voice traffic termination fraud is one of such fraud that has emerged recently with the use of VOIP technologies and the major causes of loss of revenue in the telecommunication industry. Due to this particular fraud, millions of dollars are lost from telecom operators every year and Ethio telecom is not an exception. Therefore, this research focused on telecom voice traffic termination fraud detection which is the most worrying type of fraud in today's telecom business is used in international calls.

2.2.3 Difficulties in Detecting Fraud

Detecting fraud is a challenging task and is a continuously evolving discipline. Whenever it becomes known that one detection method is in place, the fraudsters will change their tactics and try others. For example, on an industrial scale, telecommunication fraud is mainly perpetrated by organized criminal gangs, professional hackers and service providers own employees [18]. The availability of numerous hacking tools on the internet makes telecommunication fraud a widespread crime that can be committed by anybody using various methods/means depending on one's individual goal. The main motivation to commit telecommunication fraud is to make money (revenue fraud), for example, by selling fraudulently obtained telephone services at cheap rates. Other motivations are non-revenue fraud, for example, by avoiding or reducing payment of services used, demonstrating ability to outmaneuver the service provider's system security [18].

2.3 Review of Data Mining and Knowledge Discovery (DMKD)

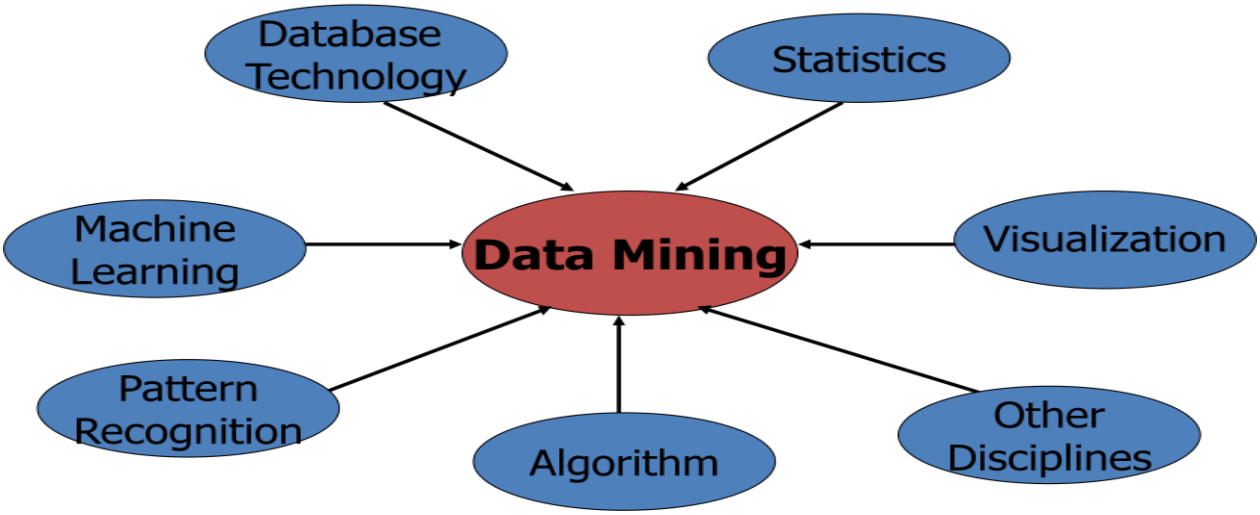
2.3.1 Basic Concepts

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, potentially useful patterns and relationships in a massive data set. There are a number of applications for machine leaning (ML), the most significant of machine leaning is data mining technology [28]. Complex and huge amount of data is generated and captured in every aspect of the industry such as marketing, insurance, finance, health, telecommunication, etc. However, it is difficult for a human being to manually collect and process these data. This vast amount of data should be baked up on computerization technology to create a valuable and interesting knowledge, which brought data mining

technology into existence.

As discussed in [29] and [30], data mining is a multidisciplinary field of information technology that adopts the techniques and terminologies from various disciplines such as Artificial Intelligence, Statistics, Database Systems, Data Warehouse, Information Retrieval, Machine Learning, Applications, Pattern Recognition, Visualization, Algorithms, and High Performance Computing. Data mining can be also considered as an exploratory data analysis [31]. Generally, Data mining uses advanced data analysis tools to find out previously unknown (hidden), valid patterns and relationships among data in large data sets. As can be seen from Figure 2.1 data mining is the core field for different disciplines such as database, machine learning, and pattern recognition etc.

Fig 2. 1 Data mining: confluence of multiple disciplines [31]



Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories as well as the organized process of identifying crucial knowledge from large and complex data sets. Data Mining (DM) is the critical task in the KDD process involving the selection of algorithms that fits the data exploration task, development of model and discovery of previously unknown patterns. The model can be used to visualize the data generally and to analyze, predict, group, associate, formulate rules… etc. specifically [32].

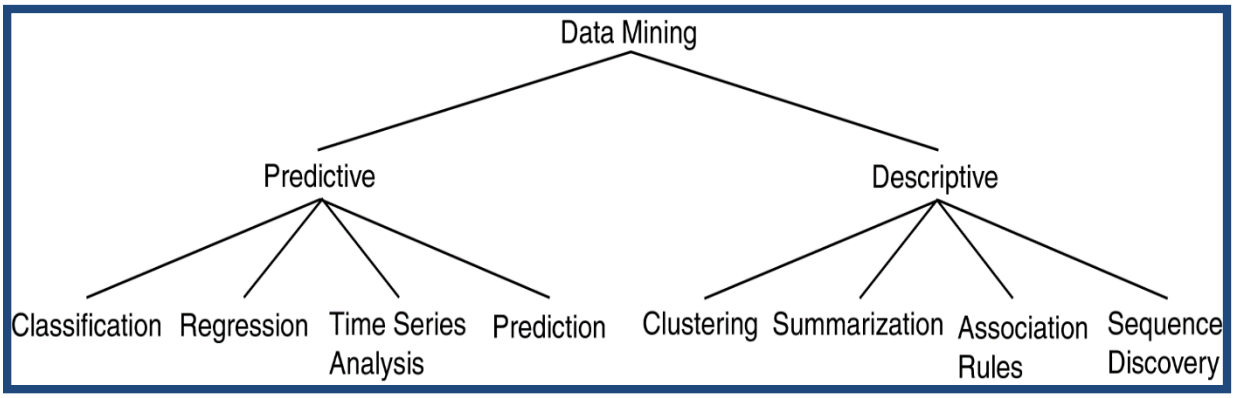
The move towards offering wide range of contemporary services such as cellular phone, smart phone, Internet access, email, text messages, images, computer and web data transmissions,

and other data traffic made the telecommunications industry to handle huge and complex data. This has made the application of data mining mandatory in the area for the accomplishment of many tasks such as detect fraudulent activities, efficient resource utilization, understand business dynamics, understand customers, and improve service quality which is in line with this study [30].

2.3.2 Data Mining Tasks

Functionalities of data mining specify the kinds of patterns found in data mining tasks that can be generally classified into two categories, descriptive and predictive. Although the boundaries between prediction and description are not sharp in such a way that some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa, the distinction is useful to understand the overall discovery goal. The relative importance of prediction and description is specific to each data mining application as the goals is achieved using a variety of particular data mining methods [33].

Fig 2. 2 Data mining tasks and models [30]



Each model encompasses different techniques so as to achieve its task objective. According to [34], different data mining techniques have been developed and used in data mining projects with the aim of achieving the data mining objectives. Some of these techniques are association, classification, and clustering. Some of the data mining models and techniques are examined as follows.

i. Predictive methods/model

Predictive mining tasks perform induction on the current data in order to make prediction which is often referred to as supervised data mining. Moreover, prediction involves using some variables or fields in the database to predict the values of unknown or future values of

other variables through the application of classification, regression, biases/anomalies detection… etc. Basic prediction methods are described in the coming paragraphs as discussed by different scholars.

Classification is an important data mining technique with broad applications. It is used to classify each item in a set of data into one of predefined set of classes or groups. It plays an important role in document classification. In this research, we have analyzed four ensemble classifiers namely bagging, boosting, stacking and voting classifiers, based on 2 basic learners, i.e., Decision tree (DT) and ANN were applied in this research.

Statistical regression is a supervised learning technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item. This model predicts classes of new variables to which they belong. Predict one or more continuous numeric variables, such as profit or loss, based on other attributes in the dataset.

Time series analysis usually involves predicting numbering value for instance about market price of future. Time series are sequences of events. For example, the final sales income is an event that occurs each day of the week and each week in the month. Time series analysis can be used to identify the sales trends of organizations like Ethio Telecom [35] also indicate the major tasks of time series in data mining are indexing, clustering, classification, anomaly detection, and segmentation.

ii. Descriptive Methods/model

Descriptive data mining tasks characterize properties of the data in a target data set. It focuses on finding human-interpretable patterns describing the underlying relationships in the data. Descriptive methods reveal patterns in data through the application of clustering, association rules, sequential patterns …etc. Moreover, descriptive data mining includes the unsupervised and visualization aspects of data mining. Descriptive task encompasses methods such as clustering, summarizations, association rules, and sequence analysis. Descriptive analysis is the task of providing a representation of the knowledge discovered without necessarily modeling a septic outcome. From a machine learning perspective, we might compare these algorithms to unsupervised learning. Basic descriptive methods are described in the coming paragraphs as discussed by different scholars.

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the

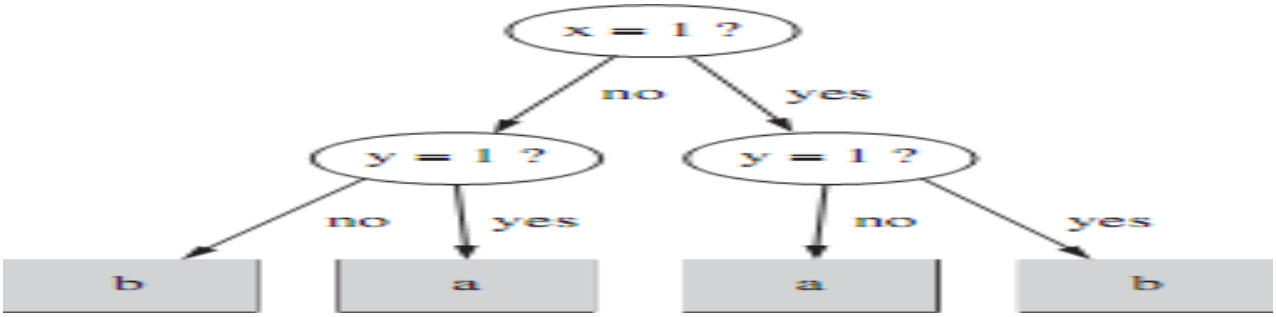
classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. Clustering (which is also called Unsupervised learning) groups similar data together into clusters. It is used to find appropriate groupings of elements for a set of data. Unlike classification, clustering is a kind of undirected knowledge discovery or unsupervised learning; that is, there is no target field, and the relationship among the data is identified by bottom-up approach [34].

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction for example calling number and peak status. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Telecom can also use association technique between calling date and time with zone to identify its customer's calling habits [34]. Sequence analysis is used to determine sequential patterns in data. The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time.

2.3.3 Classification learning algorithms (individual classifiers)

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items [35].

Fig 2. 3 A Scenario that shows how decision tree is constructed



A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), C4.5, J48 and Random forest. Decision tree models are constructed in a top-down recursive divide-and-conquer manner. J48 decision tree algorithms have adopted this approach. The training set is recursively partitioned into smaller subsets as the tree is being built [36].

J48 algorithm has various advantages. Some of these advantages are: -

- Gains of balanced flexibility and accuracy
- Capability of limiting number of possible decision points

Artificial neural networks

Artificial neural networks are computing models for information processing and are particularly useful for identifying the fundamental relationship among a set of variables or patterns in the data. They grew out of research in artificial intelligence; specifically, attempts to mimic the learning of the biological neural networks especially those in human brain which may contain more than 1011 highly interconnected neurons. They do share two very important characteristics with biological neural networks - parallel processing of information and learning, and generalizing from experience [32]. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an expert in the category of information it has been given to analyze.

A neural network consists of several neurons that are arranged in layers and are linked to every neuron in the previous layer by connections with different strengths or weights associated to them. The learning adapts weights at each point of iteration and provides the system of a method to learn by example. Multi-layer perceptron (MLP) is currently the most popular type of neural network. It is a simplified model of the human mind elaboration process and works by simulating and elevated number of simple elaboration units that resemble abstract versions of neurons. Artificial neural network has the capability to learn from the environment and enhance their performance through learning which is achieved by an iterative process of adjusting the weights and bias level.

2.3.4 Ensemble Classification Methods

Ensemble algorithms are a powerful class of machine learning algorithm that combine the predictions from multiple models [38]. The idea of ensemble methodology is to build a predictive model by integrating multiple models. It's a well-known that ensemble methods can be used for improving performance. An ensemble classifier is also a machine learning method or paradigm which uses or combines multiple classifiers to improve robustness as well as to achieve an improved classification performance from any of the consistent classifiers [38]. Furthermore, this technique is more resilient to noise compared to the use of a single classifier. This method uses a divide and conquer approach' where a complex problem is decomposed into multiple sub-problems that are easier to understand and solve international call fraud prediction problems. Ensemble methods are also meta-algorithms that combine several machine leaning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

2.3.4.1 Bagging (Bootstrap aggregation) ensemble method

Bagging stands for bootstrap aggregation, is one of the simplest but most successful ensemble methods for improving unstable classification problems [39]. For example, weak classifiers such as decision tree algorithms can be unstable, especially when the position of a training point changes slightly and can lead to a very different tree. The method is usually applied to decision tree algorithms, but it can be used with other classification algorithms such as MLP, IBK and PART algorithm. The most well-known independent method is bagging (bootstrap aggregating). After obtaining the base learners, Bagging combines them by majority voting and the most voted class is predicted. In this method aims to increase accuracy by creating an improved composite classifier and by amalgamating the various output of learned classifiers into a single prediction.

2.3.4.2 Boosting (AdaBoost) ensemble method

AdaBoost is an ensemble machine learning algorithm for classification problems that add new machine learning models in a series where subsequent models attempt to fix the prediction errors made by prior models. Boosting is also an ensemble for boosting the performance of a set of weak classifiers into a strong classifier. This technique can be viewed as a model averaging method and it was originally designed for classification, but it can also be applied to regression. Boosting provides sequential learning of the predictors. The first one learns

from the whole data set, while the following learns from training sets based on the performance of the previous one. The miss classified examples are marked and their weights increased so they will have a higher probability of appearing in the training set of the next predictor. AdaBoost technique is used in this research work and it's often provides a better result than individual classifier [40]. In this method boosting method aims to support decision maker's preferences, thus increasing performance and comprehensibility.

2.3.4.3 Stacking (Stacked generalization / Blending) method

Stacking combines several classifiers using the stacking method [40]. It can do classification or regression. We can choose different meta-classifiers and the number of stacking folds. We can choose different classifiers, different level-0 classifiers and a different meta-classifier. In order to create multiple level-0 models, we can specify a meta-classifier as the level-0 model. Stacking works a similar to boosting. We also apply several models to our original training data. Stacking is a process where the output of one level of classifiers is used as input for the next level. That is, the predictions of some classifiers are the features for other classifiers. For this, we would need to retrain one of the models with the outputs of the first classifier as input. In this study stacking aims to achieve the highest generalization accuracy.

2.3.4.4 Voting ensemble method

Voting is the simplest ensemble algorithm, and is very effective. It can be used for classification and regression problems. Voting works by creating two or sub-models. Each sub-model makes predictions which are combined in some way, such as by taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be. It's also possible that the voting ensemble results in a better overall score than the best of the base estimators, as it aggregates the predictions of multiples models and tries to cover for potential weakness of the individual models

Table 2. 1 Summary of Ensemble methods in machine learning (compiled by the author)

Common types of Ensemble Methods	
Bagging	• Reduces variance and increases accuracy • Robust and against outliers or noisy data • Often used with Decision Trees

Boosting	• Also reduces variance and increases accuracy • Not robust against outliers or noisy data • Flexible-can be used with any loss function
Stacking	• Used to ensemble a diverse group of strong learners • Involves training a second-level machine leaning algorithm called a "meta-learner" to learn the optimal combination of the base learners
Voting	• Useful technique, which comes especially handy when a single model shows some kind of bias • It's also possible that the voting ensemble results in a better overall score than the best of the base estimators, as it aggregates the predictions of multiples models and tries to cover for potential weakness of the individual models

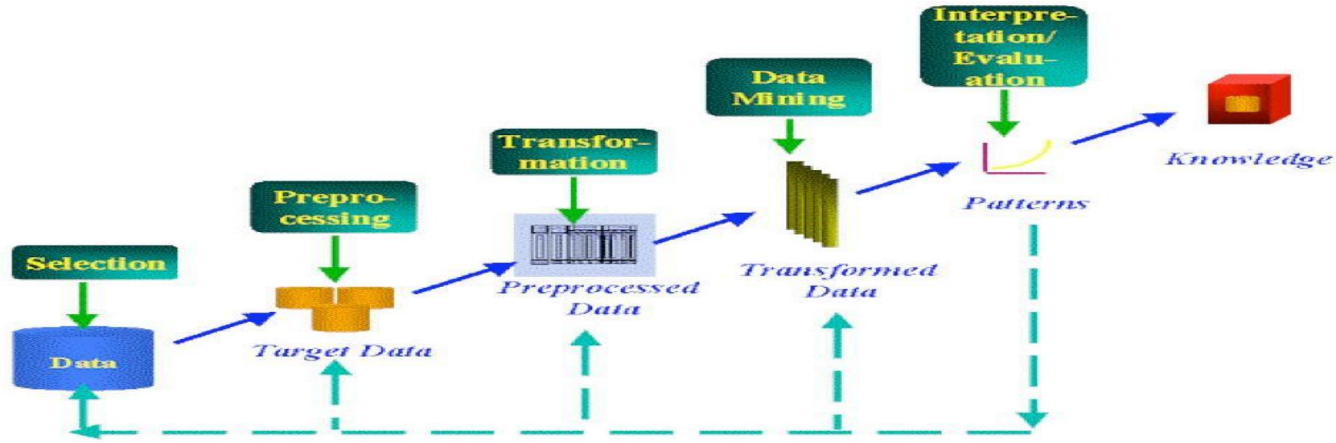
2.3.5 Data Mining Process Models

As there are many types of data mining process models, there is not much difference among the various steps since most of them are interrelated. One could follow a model that fits the specific data mining project under study. In the article by [41], one of the 10 challenging problems to be solved in data mining research is data mining process-related problems. Important topics exist in improving data-mining tools and processes through automation, as suggested by several researchers. Specific issues include how to automate the composition of data mining operations and building a methodology into data mining systems to help users avoid many data mining mistakes.

2.3.5.1 The KDD Process

The basic steps of data mining for knowledge discovery (KDD) are: defining business problem, creating a target dataset, data cleaning and pre-processing, data reduction and projection, choosing the functions of data mining, choosing the data mining algorithms, data mining, interpretation, and using the discovered knowledge. A short description of these steps follows in the coming paragraphs [42]. The KDD process is shown diagrammatically in Figure 2.4 below.

Fig 2. 4 The five stages of KDD (source: [42]



Selection - This stage is concerned with creating a target data set or focusing on a subset of variables or data samples, on which discovery is to be performed by understanding the data and the business area. Because, Algorithms alone will not solve the problem without having clear statement of the objective and understanding.

Pre-processing – this phase is concerned in removing noise or outliers if any, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes.

Transformation - The transformation of data using dimension reduction or transformation methods is done at this stage. Usually there are cases where there are large numbers of attributes in the database for a particular case. With the reduction of dimension there will be an increase in the efficiency of the data-mining step with respect to the accuracy and time utilization.

Data Mining - this phase is the major stage in data KDD because it is all about searching for patterns of interest in a particular representational form or a collection of such representations. These representations include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. Therefore, selecting the right algorithm for the right area is very important.

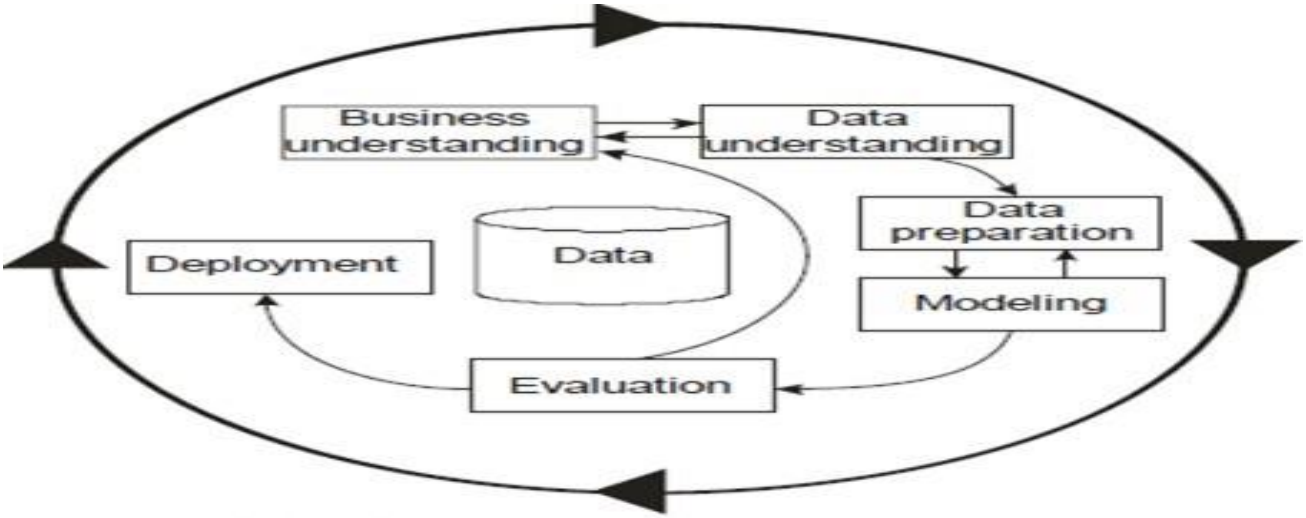
Evaluation – In this stage the mined data is presented to the end user in a human viewable format. This involves data visualization, where the user interprets and understands the discovered knowledge obtained by the algorithms.

Using the Discovered Knowledge - Incorporating this knowledge into a performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving for conflicts with previously acquired knowledge are tasks in this phase.

2.3.5.2 CRISP-DM Process Model

As described in [43], CRISP (CRoss-Industry Standard Process for Data mining) process mode involves six consecutive phases as shown in Figure 2.5 There are so many lessons to be learned during the process and from the deployed solution that can trigger new, often more-focused business questions. Moreover, the subsequent data mining processes will benefit from the experiences of previous ones.

Fig 2. 5 CRISP-DM process model (source: [43]



Each phase will be outlined as follows.

Business Understanding: The business understanding phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding: This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data and detection of interesting data subsets. Data understanding is further broken down into collection of initial data, description of data, exploration of data and verification of data quality.

Data Preparation: Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. The purpose of data preprocessing is to clean selected data for better quality. The data selected may have different formats as the selection could be from different sources

Modeling: Data modeling is where the data mining software is used to generate results for various situations. A cluster analysis and visual exploration of the data are usually applied first. Depending upon the type of data, various models might then be applied.

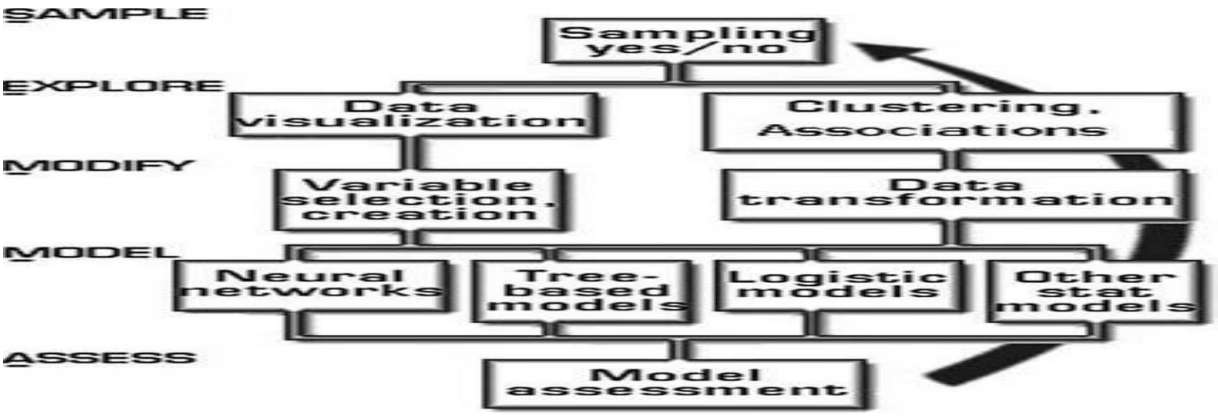
Evaluation: This is the fifth stage of CRISP-DM, after one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key sub steps in this step include evaluation of the results, process review and determination of the next step.

Deployment: The results of the data mining study need to be reported back to project sponsors. The data mining study has uncovered new knowledge, which needs to be tied to the original data mining project goals. Management will then be in a position to apply this new understanding of their business environment.

2.3.5.3 The SEMMA Process Model

The SEMMA process was developed by the SAS (Statistical Analysis System) Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process. A pictorial representation of SEMMA is given in figure 2.6 [44].

Fig 2. 6 the Schematic of SEMMA



Steps in SEMMA Process

Sample: This is where a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly is extracted. A sampling strategy, which applies a reliable, statistically representative sample of the full detail data, is advocated for optimal cost and computational performance. It is also advised to create partitioned data sets for better accuracy assessment.

Training – used for model fitting.

Validation – used for assessment and to prevent over fitting.

Test – used to obtain an honest assessment of how well a model generalizes.

Sample: This is the first and optional stage of SEMMA process which focuses on sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.

Explore: This is the second stage of SEMMA process which focuses on exploration of data. This can help in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

Modify: This is the third stage of SEMMA process which focuses on modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.

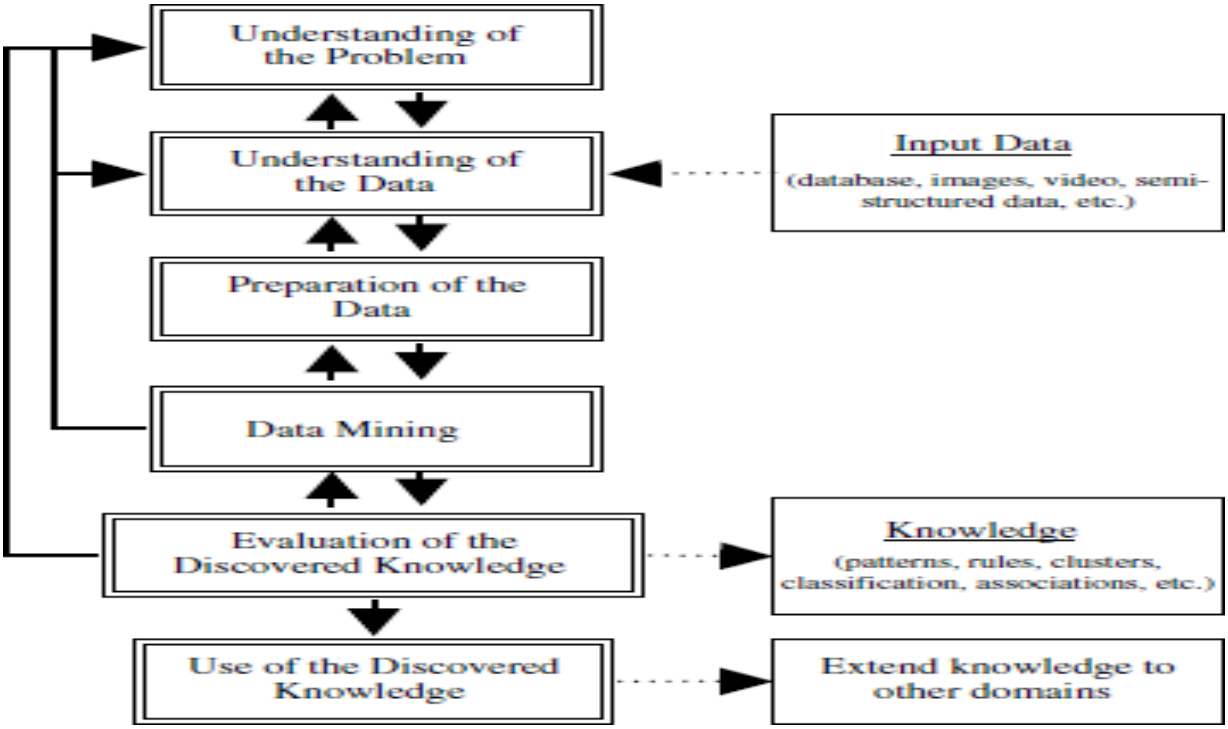
Model: This is the fourth stage of SEMMA process which focuses on modeling of data. The software for this automatically searches for combination of data. There are different modeling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.

Access: This is the fifth and final stage for SEMMA process focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

2.3.5.4 Hybrid Process Model

According to [45] KDD, CRSP-DM, SEMMA and Hybrid model are commonly used for data mining process. In this thesis hybrid data mining methodology was used, because the hybrid data mining methodology has five iterative steps, but CRISP has only three feedback mechanism. In addition, SEMMA and KDD have no feedback mechanism. Moreover, hybrid data mining methodology is more general research-oriented description of the steps and introducing 'using discovered knowledge' step instead of the 'deployment' in CRISP. As a result, to encourage knowledge discovery process for this research domain hybrid methodology used to detect and predict voice traffic termination, mainly voice bypass fraud. In addition, CRISP-DM and SEMMA mostly company oriented especially SEMMA that is used by SAS enterprise miner and integrate with their software. However, CRISP-DM is more complete as compare to SEMMA. Furthermore, all these process models guide and helps the people and experts to know that how they can apply data mining into practical scenarios. However, hybrid methodology is the extension of CRISP-DM so, hybrid data mining methodology is more complete and better than others and this methodology is used for this research domain. On the other hand, the development of academic and industrial models has led to the development of hybrid data mining methodology. Hybrid methodology is a six-step KDP model developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research i.e. hybrid data mining methodology is recommended for academic research. In this study based on the six step hybrid data mining methodology the following steps with detailed task are performed.

Fig 2. 7 The Hybrid Models (description of the six steps follows) (source: [45]



Understanding of the problem domain: This helps to work closely with domain experts to define the problem and determine the project goals by selecting key people and learning about current solutions to the problem. It also involves learning domain-specific terminology. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

Understanding of the data: This is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

Preparation of the data: This phase uses for preparing necessary data for subsequent operations. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values. The cleaned data are fed to next operations like reducing dimensionality, Discretization and data granularization. The end results are data that meet the specific input requirements for the DM tools selected in first step.

Data mining: At this point the researcher will be used different data mining algorithms to extract knowledge from preprocessed data.

Evaluation of the discovered knowledge: In this step the researcher in collaborating with domain experts of Ethio-telecom, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge.

Use of the discovered knowledge: This final steps consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

Comparative analysis of KDD, CRISP-DM, SEMMA & Hybrid Methodology

Table 2. 2 Summary of KDD, CRISP-DM, SEMMA and hybrid processes model [45]

Data mining process model	KDD	CRISP DM	SEMMA	Hybrid

No. of steps	9	6	5	6
	Developing and Understanding of the Application	Business Understanding	---------	Understanding of the problem domain
	Creating a Target Data Set	Data Understanding	Sample	Understanding of the data
	Data Cleaning and Pre-processing		Explore	
	Data Transformation	Data Preparation	Modify	Preparation of the Data
	Choosing the suitable Data Mining Task	Modeling	Model	Data mining
	Choosing the suitable Data Mining Algorithm			
	Employing data mining Algorithm			
	Interpreting Mined Patterns	Evaluation	Assessment	Evaluation
	Using Discovered Knowledge	Deployment	----------	Use of the discovered knowledge

2.4 Review of Related Research Works

In this section, different works which are related with fraud detection in telecommunication sectors are presented. The papers presented are knowledge discovery from mobile network and different works in telecom sector using data mining techniques are selected and presented based on the relevance and similarities of the papers with this thesis work. So, the related works are presented in the following two sections in two categories. The first section focuses on global works related to fraud detection and the second is about the local DM researches.

Related Works on Fraud Detection (Global Context)

Abidogun [3] also conducted a research on data mining, fraud detection and mobile telecommunications. One of the strategies for fraud detection is to check for signs of questionable changes in user behavior. Although the intentions of the mobile phone users cannot be observed, their intentions are reflected in the call data which define usage patterns. Over a period of time, an individual phone generates a large pattern of use. While call data are recorded for subscribers for billing purposes, we are making no prior assumptions about the data indicative of fraudulent call patterns, i.e. the calls made for billing purpose are unlabeled. Further analysis is thus, required to be able to isolate fraudulent usage. An unsupervised learning algorithm can analyses and cluster call patterns for each subscriber in order to facilitate the fraud detection process. This research investigates the unsupervised learning potentials of two neural networks for the profiling of calls made by users over a period in a mobile telecommunication network. The study provides a comparative analysis and application of Self-Organizing Maps (SOM) and Long Short-Term Memory (LSTM) recurrent neural networks algorithms to user call data records in order to conduct a descriptive data mining on users call patterns.

The method of research proceeds with the normalization of call data records which contained a 6-month call data set of 500 masked subscribers from a real mobile telecommunication network. The researcher extracted from the data set, mobile originating calls (MOC). These are calls that were initiated by the subscribers. Within the 6 months' period, a total of 227,318 calls originated from the 500 subscribers. The SOM and LSTM RNN models are then applied to unsupervised discrimination of the normalized call data set. Results are reported which estimates the performances of the two learning models.

Abdi Karim and Rosalina Salahuddin [12] proposed the artificial neural network and support vector machine to detect GSM gateway bypass in SIM box fraud. The suitable features of data obtained from the extraction process of call detail records (CDRs) are used for classification in the development of ANNs and SVM models. The performance of ANN compared with SVM to find which model gives the best performance from the experiments, it's found that SVM model gives higher accuracy compared to ANN by giving the classification accuracy of 99.06% compared with ANN model, 98.71% accuracy.

Ojuka [7] also conducted a research in the area of fraud detection in telecommunication. According to the researcher, detecting fraud is a challenging task and is a continuously evolving discipline. Whenever it becomes known that one detection method is in place, the fraudsters will change their tactics and try others. For example, on an industrial scale, telecommunication fraud is mainly perpetrated by organized criminal gangs, professional hackers and service providers own employees.

Burge-Shawe and Fawcett-Provost [46] presented adaptive fraud detection. Fraudster adapt to new prevention and detection measures, so fraud detection needs to be adaptive and evolve over time. However, legitimate account users may gradually change their behavior over a longer period of time, and it is important to avoid spurious alarms. Models can be updated at fixed time points or continuously over time.

Local DM Research Works

Jember [47] conducted a research on data mining in supporting fraud detection on mobile communication service in the case of Ethio - mobile. According to the researcher, the application of data mining methods and tools to large quantities of data generated by the call detail record (CDR) of telecommunication switch machine will be the art of the day to address the serious problems of telecommunication operators. The CDR consists of a vast volume of data set about each call made and it is a major resource of data for research works to find out hidden patterns of calls made by customers in addition to the typical use for bill processing activities.

The data used by the researcher was a three-month data from October and November of 2003 and March of 2004. From the total of 9153 customers, 900 customers who made an international call of more than one minute per call were selected using stratified sampling technique in order to get representative data from all the postpaid mobile subscribers. The

methodology used had three basic steps, data collection, data preparation, model building and testing. Matlab tool and neural network data mining technology were employed to build the models from which an accuracy of 89% was achieved.

The research selected the attributes minimum number of calls, maximum number of calls, average number of calls, standard deviation of number of calls, total number of calls minimum duration, average duration, standard deviation of duration and total duration. The research scope was limited to exploring the possibility of application of the DM technology to supply fraud detection in mobile communication network using artificial neural network of the postpaid mobile phone. The researcher recommended giving enough time to get representative samples of the data and enough time to building an appropriate model to get the best performance.

Gebremeskel [48] also conducted a research on data mining application in supporting fraud detection: on Ethio-mobile services. An attempt had been made to assess the possible application of data mining technology to support mobile fraud detection on Ethio-Mobile Services. The data source of this study was taken from call detail record (CDR). The researcher used a two-month data from the month of February to March of 2005. The methodology by Gebremeskel included data collection, data preparation, and model building and testing. SPSS software tool and artificial neural network algorithm were used as data mining techniques. The researcher took 29,463 records as a sample and selected 9 attributes for that process. From these 29,463 call records, 9186 were identified as fraudulent calls. In general, the numbers of fraudulent call became 31.18 percent of the total sample size (total record). Additionally, the study was targeted at prepaid mobile service. The number of post-paid mobile subscribers was 53614 and the prepaid ones were 381417. The researcher further recommended that additional research could be done by including other attributes of the call detail record to build models with better performance and better accuracy. Feven [49] tried to study on fraud detection in pre-paid mobile phone and the scope of her study was limited to prepaid mobile customer and here paper followed CRISP-DM process model. Here classification techniques have been used to conduct the experiments are, multilayer perceptron, J48, k nearest neighbor and self-organizing map (SOM). As a result of the experiments, it is found that BFTree algorithm from decision tree model gives higher accuracy compared to the others by giving the classification accuracy of 99.9%. The data she used is

restricted on one months and from CDR data only. Finally, the researchers use the rules generated from BFTree algorithms to construct telecom fraud prediction model. The researcher also further recommended that additional research could be done in post-paid mobile services and by including other attributes of the call detail record to build models. Kahsu [51] still also tried to develop predictive modeling to differentiate fraudulent from legitimate subscribers using data mining techniques. Three classification techniques, Random forest (RF), Artificial neural network (ANN), Support vector machine (SVM), and user profiling data set were proposed. Results of the work show that Random forest performed better among the three algorithms with accuracy of 95.99% and a lesser false positive rate. The researcher also further recommended that additional research could be done by including other attributes and methods may improves performance and accuracy of the technique, for instance including the ratio of International mobile subscriber identity (IMSI) to International mobile equipment identity (IMEI). Yeshinegus [52] tried to study on fraud detection in telecommunication. The scope of his study was limited to prepaid mobile customer and he followed CRISP-DM process model. His classification methods of data mining are applied using J48, PART and multilayer perceptron algorithms and experimentation result showed that the model from the PART algorithm exhibited 100% accuracy level followed by J48 algorithm with 99.98%.

Birhanu [50] conducted a research on fraud detection in telecommunication networks using self-organizing map (SOM): the case of Ethiopian Telecommunication Corporation. He used unsupervised feed-forward neural network model, which is SOM, it helps to analyze and visualize high dimensional data. SOM also enables clustering data without knowing the class membership of the input data, unlike neural network models based on supervised learning. Then the clustering capability of SOM is used to group similar call pattern behavior analysis. He used extended map model to identify suspicious calls and the result has shown that these calls are identified as fraudulent or suspicious call patterns.

From the above related works mentioned in both sections, it can be seen that type of algorithms used to detect fraudulent activities by analyzing the patterns of each calls, dataset, and attributes/features used in the process have significant impact on classification performance. In addition that, the scope of the existing publicly available training datasets limited to prepaid mobile subscriber's data or doesn't consider the postpaid mobile

subscriber's data, some algorithms were also not tested in detecting and predicting telecom fraud. Though the aforementioned researches have a substantial contribution for showing the directions in conducting this study. Considering these factors, selected ensemble machine learning algorithms have been tested in detecting and predicting telecom voice traffic termination fraud using labeled training dataset.

The output of this research gives new patterns, new information and new insight which helps to improve the quality of service provided. The improvement leads to save the time and revenue maximization. Generally, this research is more interested in generating rules that best predict and detect the voice traffic termination fraud and to come to an understanding of the most important factors (variables) affecting the mobile telecom to be fraudulent.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter the study describes data mining goals, sources of data and techniques that have been used in preprocessing and model building phases are discussed in detail.

3.2 Research methods

Methodology means the steps or procedures that the researcher follows to achieve the objectives stated. It is a road map that shows the direction of how the research is going to be done to reach the end. In this research, the researcher aimed to use the six-step process of Hybrid model to achieve the stated objectives. This is because the model has been widely applied for data mining studies in academic research [45]. Accordingly, this study has the following methodologies in order to develop classification model that predict voice traffic termination fraud behavior in call detail record historical data.

3.3 Research design

This research is an experimental research. This is because experiments that will occur in order to extract results from real world implementations and it is important to restate that all the experiments and results should be reproducible [53]. In this chapter the methods, techniques and tools used to conduct the research were discussed in detail based on the steps of hybrid data mining process model selected to guide the entire process for this research.

As it is discussed in chapter two data mining process models were developed for purely academic purposes and for industrial purposes. The six-step process of hybrid model was developed for both academic and industrial. The process model adopted to undertake this research is the hybrid data mining process model due to the reason that this model describes each of the knowledge discovery process steps in better way and it is flexible since it has a feedback mechanism in more steps than the other process model [45].

The hybrid data mining process model has six steps that are: Understanding of the problem domain, understanding of the data, preparation of the data, data mining for building predictive model, evaluation of the discovered knowledge, and discovered knowledge usage. In this

thesis the following specific tasks were performed under each of the six step hybrid data mining process model.

3.3.1 Understanding of the problem domain

In this section, we tried to understand the domain to set the data mining goal and the objective of this research as per the real situation of Ethio Telecom working area and as per the domain perspective. This step helps to work closely with domain experts to define the problem and determine the project goals by selecting key people and learning about current solutions to the problem. It also helps to learn domain specific terminology.

In this research, in order to identify, define, understand and formulate the problem domain different discussion points reflect the voice traffic termination fraud were used, so as to closely works with the domain experts of Ethio telecom, then determine attribute feature selection and understanding business processes. Peer reviewed journals, document analysis, direct observation, interview and additional authenticated sources also examined to get a better insight. In consultation with the domain experts the customer data record (CDR) data was selected based on the selected key attributes as the main source of data. The call detail record database serves as a suitable source of information where useful knowledge about callers can be extracted to identify the call whether is legitimate or not among the dataset. Domain experts are consulted in order to understand the data and to know more about the problem. Based on the insight and knowledge gained about the domain of telecom business, data mining problem was defined.

Experts from Information technology/Information system security/technical audit department are communicated to provide their technical advice on the problem to be addressed in this study. They are doing the analysis using statistical tools. The output from the analysis is used to enhance the quality of service and to take appropriate actions to deliver the service to the customers as well as sent a report for higher concerned officials for further decisions. It has own drawback to analyze huge amount of data using simple statistical tools. This data is stored in different systems like customer data record, customer billing system, customer relation management system and other more systems and databases. Ethio telecom major aim is to become world class Telecom Company by providing best quality of service. So as stated in the above paragraph the data is large amount so it's impossible to do analysis manually and identify the call patterns of a legitimate and a fraud customer. So, to overcome this problem,

the concept ensemble machine learning techniques and approaches are used to identify the call patterns of customers. Furthermore, it hoped that the company revenue could be protected, and decision making can be simplified. The business goal to achieve is that to help the operator or service provider during decision making by the providing the required tool and technique.
In order to have detail understanding and knowledge about the problem domain, the researcher discussed with domain experts based on the discussion points as shown below concerning telecom voice traffic termination fraud.

- How does the current fraud management system and test call generation detect fraud numbers from legitimate calling activities?
- Does the current system detect all fraud calls efficiently?
- How do you explain voice traffic termination fraud in telecom sector?
- Why people commit telecommunication fraud?
- Which key attributes are voice traffic termination fraud indicators from the fields of CDR data?

According to the domain experts of Ethio Telecom and internal documents, the Ethiopian government has decided to transform the telecommunication infrastructure and services to world class standard. Thus, Ethio Telecom is born from this ambition in order to bring about a paradigm shift in the development of the telecom sector to support the steady growth of our country.
The vision the company is to be a world-class telecommunications service provider. So that the following points are the mission of the company

- Connect every Ethiopian through information communication technology.
- Provide telecommunication services and products that enhance the development of our nation.
- Build reputable brand known for its customers' consideration.
- Build its managerial capability that enables Ethio Telecom to operate an international standard.

In line with its ambitious mission, Ethio Telecom has the following ambitious goals:

- being a customer centric company
- offering the best quality of services

- meeting world-class standards
- building a financially sound company

3.3.2 Understanding of the Data

This step was used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values. Finally, verification of the usefulness or quality of the data with respect to the DM goals was conducted. The sections below discus the nature and structure of the collected data.

3.3.2.1 Data source and Data collection

The source of data for this study has been collected from Ethio-telecom business database records (CDR) by using cooperation letter from St. Mary's University. Then the letter was directed to concerned sections from chief officer of the company. The CDR data for this study was so bulky. Storage were a big challenge before processing the data. So sample is taken based on call duration (talk time) and amount paid by the billing mobile number in cents. The database was manipulated by using MS-Excel filtering techniques. The first and the major source of data that the researcher identified was the call detail records. When a call is to be found on a telecommunication network, descriptive information about the call is saved as a call detail record. Call detail record include enough information to describe the important characteristics of each call. This means that the CDR contains each call information related to mobile fraud call, such as billing number, time of call initiation, duration of call in second, mobile number receiving the call, recharge ID number, error call type (error CDR indicator), the amount charged or to be charged for the call duration, subscriber ID number used for the billing are some of the details of the call. In the application of the data mining technology and in developing a model that can support fraud detection, the goal of this study is to developing model so as discover the presence of incoming international calls that are terminated by using local mobile numbers.

The illegitimate calling activity may not be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call detail record, which in turn can be observed. Therefore, it is reasonable to use call detail record to apply data mining technology and to formulate model using training and testing dataset and evaluate the accuracy of the model using the WEKA software. From the company business database of two months'

data were taken which is more than 126736 of records. From these records, 50515 instances with 6 basic attributes and 4 derived attributes were taken for this study. The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered to achieve the objectives of the study. Sample data was collected from the whole dataset on which feature selection and preprocessing is conducted. List of attributes in the initial dataset is found in detail under description of raw data quality (3.2.2.2).

3.3.2.2 Description of Raw Data Quality

Description of the data is very important in data mining process in order to have clearly understand the data. As indicated before the researcher collected the data from Ethio Telecom postpaid mobile CDR database records. All the CDR were not relevant or important for this study. Therefore, the researcher tried to filter the relevant data from the large database based on the call duration (talk time) and amount paid by the billing mobile number in cents. Among the 37 attributes or features 29 of them have blanks / no data/zero value due to security reasons and other attributes are not important for this study. The database was thoroughly studied and as a result of the study, 6 attributes were found to be important or relevant for this study. All the 6 basic attributes were extracted in Excel format directly from the target database. In this section, from the source described above, the attributes with their data types and descriptions are shown in the table 3.1 below. The table below shows that the initial basic attributes and derived attributes from mobile postpaid subscribers CDR database.

Table 3. 1 List of attributes in the initial dataset with derived attributes

No	Attributes Name	Data Type	Description	Attributes remark
1	EVENT_INST_ID	Number	Unique ID given for a single call	original & not selected because of not relevant for this study
2	RE_ID	Number	CDR type ID for voice, SMS and GPRS)	original & it is constant & not relevant for this study
3	BILLING_NB	Number	Bill paying number (Mobile number to be charged)	Original & the same as calling number & not selected for the sake of privacy reason

4	CDR_TYPE	Number	Call type Id (for outgoing and incoming call)	original & it is relevant for this study & selected based on reviewing literatures in the field & domain experts
5	CALLING_NBR	Number	The party that is initiating the event	Original & the same as billing number & not selected for the sake of privacy reason
6	CALLED_NBR	Number	The party that is receiving the event	Original & not selected for the sake of privacy reason
7	CALLING_IMSI	Number	Calling IMSI	original & not selected because it is blank
8	THIRD_NBR	Number	Third Party Number	original & not selected because most of its value is blank & not relevant for this study
9	DATE START TIME	Date	The start time of the event	original & not selected because it is not important for this study
10	DATE END TIME	Date	The end of the event	original & not selected because it is not important for this study
11	DURATION	Number	Total call time/duration of call in second	original & it is relevant for this study & selected based on the literature recommendation in the field & domain experts
12	CALL_FEE /Charge	Number	Amount paid in cents	original & it is relevant for this study & selected based on the literature recommendation in the field & domain experts

13	CALLED_COUNTRY	Number	Called country	original & not selected due to most of its value is blank
14	CALLING_CARRIER	Number	Calling carrier	original & not selected due to not relevant for this study
15	CALLED_CARRIER	Number	Called carrier	original & not selected due to not relevant for this study
16	CELL_A	Number	Calling distinct (Mobile BTS cell sector A number (BTS-ID) where the call is originated	original & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
17	CELL_B	Number	Called district	original & not selected due to most of its value is blank
18	STATE_DATE	Number	Billing date	original & it is not relevant for this study
19	CALLING_SUB_ID	Number	Calling subscriber ID	original & it is not relevant for this study
20	BILLING_CYCLE_ID	Number	Billing cycle ID	original & it is not relevant for this study
21	CHARGE1	Number	the amount paid by the billing mobile number in cents	original & not selected due to similar with Call FEE or charge
22	CHARGE2	Number	the amount paid by the billing mobile number in cents	original & not selected due to similar with Call FEE or charge
23	PRICE_ID1	Number	Rate ID	original & it is not relevant for this study
24	ACCT_ITEM_ID1	Number	Account item ID	original & it is not relevant for this study
25	TRAFFIC_UP	Number	Uploaded data	original & not selected due to the researcher is focused on voice CDR data
26	TRAFFIC_DOWN	Number	Download data	original & not selected due to the researcher is focused on voice CDR data

27	BILLING_OFFERING_ID	Number	Billing offering ID	original & it is not relevant for this study
28	ERROR_CDR_TYPE	Number	Error CDR Indicator	original & it is not important for this study
29	CALL_FORWARD_INDICATOR	Number	Call Forward Indicator	original & not selected due to not relevant for this study
30	HOT_LINE_INDICATOR	Number	Hot Line Indicator (voice mail)	original & it is not relevant for this study
31	PRD_ID	Number	PRD ID	original & it is not relevant for this study
32	INPUT_DATE	Number	INPUT DATE	original & it is not relevant for this study
33	ETL_FILE_NAME	String	File name from CDR data department	original & it is not relevant for this study
34	TASK_ID	Number	TASK ID	original & it is not relevant for this study
35	CALLING_TRUNK_ID	Number	CALLING TRUNK ID	original & not selected because most of the value is blanks when the call is international
36	CALLED_TRUNK_ID	Number	CALLED TRUNK ID	original & not selected because most of the value is blanks when the call is international
37	RECORD_SEQ	Number	Record sequence number give when the call is made	original & it is not relevant for this study
38	CALLED NUMBER_COUNT	Number	Number of outgoing calls originated from the given subscriber	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
39	NUMBER OF UNIQUE OUTGOING CALL	Number	Number of unique subscribers called	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts

40	INCOMING NUMBER COUNT	Number	Number of Incoming calls terminated in the subscriber	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
41	TOTAL NUMBER OF CALLOUT	Number	Number of distinct cells Accessed by the subscriber	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
42	CALL RATIO	Number	The ratio of incoming to outgoing calls	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
43	IS FRAUD	Char	To categorize as fraudulent & legitimate/fraudulent activities	Derived and relevant for this study

3.3.3 Preparation of the Data

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks likely to be performed multiple times, and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools and other activities are performed to preparing the data for mining. Moreover, in this step before mining the patterns, the researcher balancing the negative and positive class by using WEKA class balancer. During this pre-processing stage, several processes took place and each of these steps are described briefly below:

3.3.3.1 Data Selection

The data selection for this study is based on recommendation of domain experts of the company as well as the researchers. The collected data for analysis in this study was Ethio Telecom mobile subscriber's usage pattern of over the period of two months. The whole target dataset may not be taken for the data mining task. Irrelevant or unnecessary data are

eliminated from the data mining data base before starting the actual data mining function. The dataset consists of ten columns and 50515 rows. From this, 25266, represents those of fraudulent records and the other half, 25249, represents the legitimate or non-fraudulent records. Further the data instances are spilt into two set. The splitting is made by rule thumb which says more than two third of data can sufficiently represent the whole data for training. Also, this idea is supported by Weka data mining tool which sets default splitting to 66% for training and the remaining for testing purpose. The rest of dataset is used for training purpose. Since this data set contains irrelevant and unnecessary data, all are not used for training. Those irrelevant records for this study are short message service (SMS), short numbers, fixed telephone and data service etc. are eliminating from the record. So, after eliminating irrelevant and unnecessary data only a total of 50515 datasets are used for the purpose of conducting this study.

The above elaborated table (Table 3.1) of the CDR database consists of 43 attributes. The first was to remove from the database those fields or attributes, which were irrelevant to the task at hand. As shown in table 3.1, the following are the initial sets of attributes, which are further preprocessed to select the final attributes used in the study.

Accordingly, EVENT_INST_ID,RE_ID, BILLING_NB, CDR_TYPE, CALLING_NBR, CALLED_NBR,CALLING_IMSI,THIRD_NBR,START_TIME,END_TIME,DURATION, CALL_FEE/CHARGE,CALLED_COUNTRY,CALLING_CARRIER,CALLED_CARRIER,CELL_A,CELL_B,STATE_DATE,CALLING_SUB_ID,BILLING_CYCLE_ID,CHARGE1,CHARGE2,PRICE_ID1,ACCT_ITEM_ID1,TRAFFIC_UP,TRAFFIC_DOWN,BILLING_OFFERING_ID,ERROR_CDR_TYPE,CALL_FORWARD_INDICATOR,HOT_LINE_INDICATOR,PRD_ID,INPUT_DATE,ETL_FILE_NAME,TASK_ID,CALLING_TRUNK_ID,CALLE_TRUNK_ID and RECORD_SEQ were the initial datasets of attributes that are selected.

3.3.3.2 Attribute Selection

In data mining technology not, all attributes are relevant. Selection of relevant attribute is necessary for data mining, among all original attributes. Many irrelevant attributes may be present in data to be mined. So, they need to be removed. Also, many mining algorithms do not perform well with all features or attributes. Therefore, features selection techniques need to be applied before any kind of mining algorithm is applied. The main objectives of feature

selection are to avoid over fitting and improve model performance and provide faster and more cost-effective models [55]. As a result, all attributes are not necessary for the experiment. So that very limited numbers of attributes that are most important for the study at hand are selected because of this reason. In order to select the best attributes from this initial collected dataset, the researcher evaluates the information content of the attributes with the helping of the domain expert and selected based on the literature recommendation in the domain. Hence, the researcher together with the domain expert, removes those attributes, which have less important for the research and the remain attributes as shown in the table 3.2 are the final list of attributes that have been used in this study.

Based on domain experts, real working situations, and literature recommendation in the field the selected attributes for this thesis work are listed in the following table.

Table 3. 2 The Final List of Attributes used in this study

No	Attributes Name	Data Type	Description	Attributes remark
1	CDR-TYPE	Number	Call type Id (for outgoing and incoming call)	original & it is relevant for this study & selected based on reviewing literatures in the field & domain experts
2	DURATION	Nominal	Total call time /duration of call in second	original & it is relevant for this study & selected based on the literature recommendation in the field & domain experts
3	Call-FEE /CHARGE	Nominal	Amount paid in cents	original & it is relevant for this study & selected based on the literature recommendation in the field & domain experts
4	CELL_A	Number	Calling distinct (Mobile BTS cell sector A number (BTS-ID) where the call is originated	original & it is relevant for this study & selected based on the literature recommendation in the field & domain experts

5	CALLED NUMBER COUNT	Number	Number of outgoing calls originated from the given subscriber	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
6	NUMBER OF UNIQUE OUTGOING CALL	Number	Number of unique subscribers called	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
7	TOTAL NUMBER OF CALLOUT	Number	Number of distinct cells Accessed by the subscriber	Derived & it is relevant for this study & selected based on the reviewing literatures in the field & domain experts
8	INCOMING NUMBER COUNT	Number	Number of Incoming calls terminated in the subscriber	Derived and relevant for this study & selected based on literature recommendation in the field
9	CALL RATIO	Number	The ratio of incoming to outgoing calls	Derived and relevant for this study & selected based on literature recommendation in the field
10	IS FRAUD	Char	To categorize as Fraudulent and Legitimate or non-fraudulent activities	Derived and very relevant for this study

3.3.3.3 Data Cleaning

This phase was used for making sure that the data is free from different errors. To do this, different operations were performed including removing or reducing noise by applying smoothing techniques, correcting missing values by replacing with the most commonly occurring value for that attribute [54]. Consequently, the data cleaning has become a must in order to improve the quality of data to improve the performance of the accuracy and efficiency of the data mining techniques. The researcher was cleaned the data, by removing the records

that have incomplete or invalid data and under each column. Removing of such records was done and their removal does not affect the entire dataset. The researcher makes use of the MS-Excel application and manually tracing, and fixing is other technique of the researcher for cleaning the data.

Generally, in this thesis detecting and correcting errors in the data were done. In order to provide access to accurate and consistent data, outlier detection, data reduction and data transformation were performed.

3.3.3.4 Data Construction and Transformation

The other important step in preprocessing is deriving other fields from the existing ones. Adding fields that represent the relationships in the data are likely to be important in increasing the chance of the knowledge discovery process yield useful result [54]. In consultation with the domain experts at Ethio Telecom IT and Network security department, Leghar branch the following fields or attributes are considered essential in determining the fraudulent and non-fraudulent /legitimate activities and were derived from the existing fields. The derived attributes include: Total number of outgoing calls (this refers to the number of outgoing calls originated from the given subscriber), Number of unique called numbers (refers to the number of unique subscribers called), Total number of incoming calls (refers to the number of Incoming calls terminated in the subscriber), Call ratio (refers to the ratio of incoming to outgoing calls), Call type (to identify whether the call is fraudulent or legitimate/ non-fraudulent activities), number of callout (refers to number of distinct cells accessed by the subscriber).

3.3.3.5 Data Formatting

At this step the researcher changed the data into a format which was suitable for the data mining tool algorithms. The preprocessing of the data performed in WEKA 3.8.0 and MS-Excel. The final data set was also in MS-Excel format. However, the selected tool WEKA does not accept the data in comma delimited (CSV) text file. It rather accepts data in ARFF (Attribute Relation File Format). Comma delimited applied for a list of records where the items are separated by commas, whereas ARFF is extension of a file format that the WEKA software can read. The dataset, which is in ARFF file format, was ready to be used.

3.3.4 Data mining for building predictive model

In hybrid data mining process model, the fourth step is model building. After the data has been prepared for analysis it is used to build classification models using ensemble learning methods. Ensemble models often place high in data mining competitions and thus demonstrate the advantage of ensemble learning over any single data mining algorithm. Ensemble classifiers in ML play a key role in prediction problems. The use of Ensemble Methods (EMs) for classification is among the recent areas of research in ML. Many recent researches specify that EMs lead to a major improvement in classification performance by choosing suitable class. For this work, several ensemble ML techniques are explored and evaluated on real international incoming telephone datasets [8]. In this thesis, new ensemble classification methods were proposed. Recently, advances in knowledge extraction techniques have made it possible to transform various kinds of raw data into high level knowledge. However, the classification results of these techniques were affected by the limitations associated with individual techniques. Hence, hybrid or ensemble approach is widely recognized by the data mining research community in order to improve the performance of a single classifier.

Ensemble methods have been suggested to overcome the defects of using a single supervised learning method [8].

In statistics and machine learning, ensemble classifiers are used to improve predictive results obtained by using its constituent algorithms. An ensemble combines several 'weak' learners and aggregates their results into one 'strong' leaner. An ensemble is a supervised algorithm itself. This is because it can be trained on labeled data and later used to predict labels for previously untested data. Ensembles thus work in a two-step process. Firstly, a set of classifier methods are learnt. The results of these methods are combined to obtain higher accuracy. Ensemble algorithms has many advantages over using a singular classification method. Most important advantage is that the results of the classification are less dependent on the peculiarities of a single algorithm. As a result, in this study, we investigate the potential applicability of data mining technique using ensemble learning method to solve voice traffic termination fraud detection problems. So basically, ensemble ensures a more generic inspection of data at hand. Another advantage is that ensemble different methods makes the entire more expressive and unbiased than a single model [55].

Therefore, this study uses on an ensemble based machine learning paradigm like bagging, boosting, stacking and voting classifiers, based on 2 basic learners, i.e., Decision tree (DT) and Artificial neural network (ANN) were applied in this research work due to the fact that they are supported /cited by most of literatures that showed promising results [8,9]. Depending on the nature of the data and the purpose of the study, the mentioned techniques and algorithms are also selected. Also, the model building process performed on Explorer environment on WEKA 3.8.3. WEKA's ability to operate on csv format made is easy to convert the selected subscriber's data set to suit this important requirement. This has enabled to design a better strategy for voice traffic termination fraud detection. WEKA was adopted for mining of the data. The WEKA tool was chosen because [46]:

- Its functionality (support for many algorithms)
- Familiarity of the researcher with the software and ease of use (has a graphical user interface) and runs on almost any platform.
- It contains a compressive collection of data preprocessing and modeling techniques
- It is freely available under the General Public License (GNU)

3.3.5 Evaluation of the discovered knowledge

This is the fifth step of hybrid data mining methodology. Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yardstick to examine the performance of any model. The evaluation is important for understanding the quality of the model or technique, for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set or model or techniques.

In this research different ensemble based classification models were developed and evaluated using training and testing dataset. In this study, the researcher used mean absolute error (MAE), root mean squared error (RMSE), confusion matrix, ROC curve analysis, F-measure, accuracy, TP rate, FP rate and time taken to build the model to evaluate the performance of the discovered knowledge. Mean absolute error and Root mean squared error are used in this study to evaluate the magnitude average error of the models. Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. Root

mean absolute error (RMSE) is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation.

Furthermore, the effectiveness and efficiency of the model was also computed in terms of recall (sensitivity) and precision (specificity). Here in collaboration with domain experts of Ethio telecom, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge is evaluated.

3.3.6 Use of the discovered knowledge

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step knowledge discovered is incorporated into performance system and take this action based on the discovered knowledge. In this research the discovered knowledge was used by integrating the user interface which is designed by VB.NET programming language with a Weka system in order to show the prediction of voice traffic termination fraud based on the extracted rules.

CHAPTER FOUR

EXPERIMENTATION AND RESULT ANALYSIS

Introduction

As indicated at the outset, the objective of this research is to build a predictive model to detect telecom voice traffic termination fraud using ensemble learning classifiers based on the data extracted from the Ethio Telecom live CDR database. Thus, the experimentation is conducted based on the selected process model which is hybrid indicated in Section 1.5.1 and 3.3 respectively.

The models are built with four ensemble-based machine learning paradigms. These methods are boosting, bagging, stacking and voting classifiers, based on 2 basic learners (decision tree and neural network) classifiers using WEKA 3.8.3 machine learning software. These classifier algorithms are used because it is recommended and commonly used in different literature reviews then do the comparison and selection of the best classifier based on their performance.

The classifiers were extracted from the dataset which were required for training and testing the models created by the classifiers. For creating predictive model, a total size of 50515 records were used for training and testing. The experimentation is conducted based on two classification test options, 10-fold cross validation and percentage split. The default value of percentage split, which is 66% was for training and 34% was used for testing. In both model evaluation techniques, all attributes are considered. On the top of all these, Knowledge from domain experts was crucial to understand the application domain where this study was conducted.

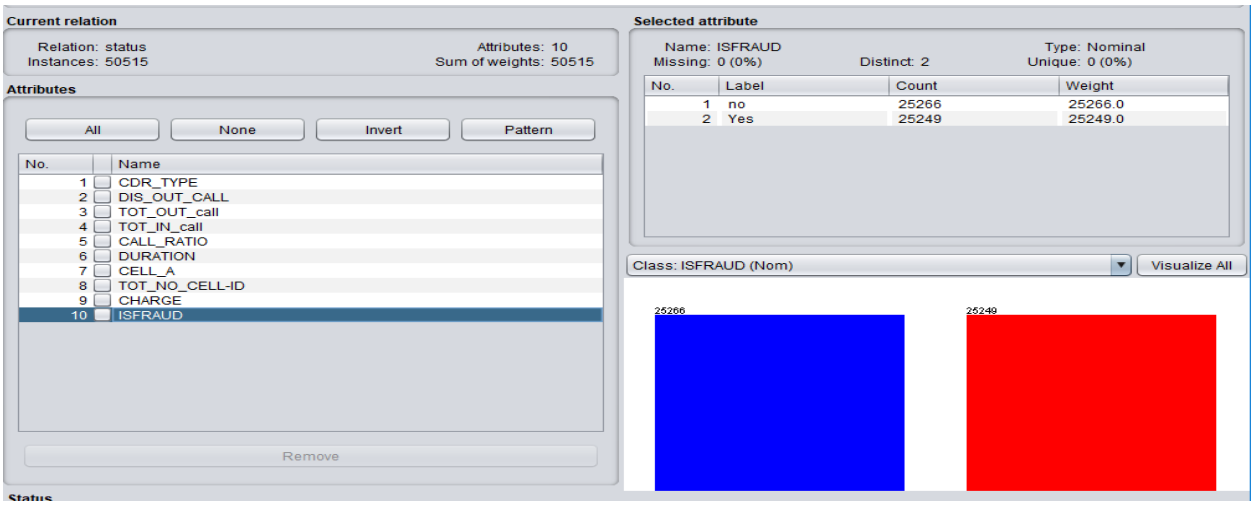
The performances of the models in this study are evaluated using the standard metrics of accuracy, FP rate, TP rate, recall, precision and F-measure which are calculated using the predictive classification table, known as Confusion matrix. ROC curve analysis was also used to compare the performances of classifies. These performance evaluation methods are well known measures for evaluation of ensemble data mining models for classification. In this research, the analysis and interpretation of the results was made by the researcher and domain experts.

4.1 Model building experiment using all attributes with ensemble methods

4.1.1 Model Building Experiment Using AdaBoost (Boosting) Method

AdaBoost is an ensemble machine learning algorithm for classification problems, that add new machine learning models in a series where a subsequent model attempt to fix the prediction errors made by prior models. Boosting is also an ensemble for boosting the performance of a set of weak classifiers into strong classifiers [36]. It can be used with classification algorithms such as J48, MLP and PART algorithm. In this study boosting method aims to support decision maker's preferences, thus increasing performance and compressibility. The model was built based on 10-fold cross validation and default values of 66% percentage split. The following attributes were selected for the experiments indicated below. The following snapshot, Fig 4.1, shows when the prepared data is loaded to WEKA for the first time.

Fig 4. 1 Snapshot showing when first time data is loaded to WEKA tool



Experiment One:

The first experiment was designed to evaluate the performance of boosting ensemble method. In this experiment the 10-fold cross validation and the default percentage split test option in WEKA were used. Table 4.1 shows the resulting performance measurement of boosting ensemble method with 10-fold cross validation and default percentage split of the model.

Table 4. 1 Detailed performance measures for experiment One

Method	Algorithm	Test Option	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
AdaBoost	**J48**	**10-fold**	**96.73%**	**0.967**	**0.033**	**0.967**	**0.967**	**0.967**	**0.988**
		66%	96.12%	0.961	0.039	0.961	0.961	0.961	0.987
	PART	10-fold	96.43%	0.964	0.036	0.964	0.964	0.964	0.987
		66%	95.99%	0.960	0.040	0.960	0.960	0.960	0.985
	MLP	10-fold	81.69%	0.817	0.183	0.819	0.817	0.817	0.888
		66%	81.75%	0.817	0.183	0.817	0.817	0.817	0.888

KEY: Accuracy: Registered performance of model Average, (TPR) True Positive Rate. (FPR) False Positives Rate, (PR) precision rate, (RR) Recall rate, (ROC) Area: Region of Convergence curve.

As shown from the resulting performance measurements in table 4.1, the boosting of J48 learning algorithm (using 10-fold cross validation test option) scored the highest accuracy of 96.73%. Which means that out of the total training datasets (50515), 467527 (96.73%) records were correctly classified instances, while 1652 (3.2703%) of the records are incorrectly classified instances. In second case using default percentage split (66%) test option, out of the 17175 testing records 16509 (96.1223%) of them were correctly classified instances. Only 666 (3.8777%) records are incorrectly classified instances. Considering the average, No and Yes class better result on true positive rate, false positive rate, precision, recall, F-Measure and ROC area of this model are 0.967, 0.033, 0.967, 0.967, 0.967 and 0.988 respectively. From the results it can be seen that boosted J48 decision tree with 10-fold cross validation has relatively better performance than the percentage split case. The snapshot running information for this experiment is shown in Annex-1.

As shown from table 4.1 boosted PART algorithm with 10 fold cross validation achieved an accuracy of 96.4327%. This result shows that out of the total training datasets 50515

(96.4327%) records were correctly classified instances, while 1802 (3.5673%) of the records were incorrectly classified instances. In second case using 66% split test option in WEKA: 16486 (95.99%) records were correctly classified instances, while 689 (4.0116%) of the records were incorrectly classified instances. From the results it can be seen that boosted PART algorithm with 10-fold cross validation has relatively better performance than the percentage split case. The snapshot running information for this experiment is shown in Annex-1. The snapshot running information for this experiment is shown in Annex-2.

As shown from the performance result of boosted MLP algorithm with 10-fold cross validation scored an accuracy of 81.69%. Which means that out of the total 50515 records used for training 41264 (81.69%) records were correctly classified instances and 9251 (18.3134%) of the records were incorrectly classified instances. In second case using 66% split: 14040 (81.75%) records were correctly classified instances, while 3135 (18.2533%) of the records were incorrectly classified instances. This shows that the experiment conducted with the 10-fold cross validation algorithm is better performance than the experiment conducted with default percentage split 66%. The snapshot running information for this experiment is shown in Annex-3.

4.1.2 Model Building Experiment Using Bagging Method

The second ensemble learning classifier applied in this research was bagging (bootstrap aggregating). Bagging stands for bootstrap aggregation, is one of the simplest but most successful ensemble methods for improving classification problems [35]. The method is usually applied to decision tree algorithms, but it can be used with other classification algorithms such as PART, MLP and IBK algorithm. In this study this method aims to increase classification performance by creating an improved composite classifier and by amalgating the various output of learned classifiers into a single prediction. The model is built based on the default values of 66% percentage split and 10-fold cross validation test option.

Experiment Two:

This second experiment was designed to evaluate the performance of bagging ensemble method. In this experiment also the 10-fold cross validation and percentage split model evaluation techniques were used. The outcome of this experiment is presented in table 4.2 below.

Table 4. 2 Detailed performance measures for experiment Two

Method	Algorithm	Test Option	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
Bagging	**J48**	**10-fold**	**95.87%**	**0.959**	**0.041**	**0.959**	**0.959**	**0.959**	**0.985**
		66%	95.08%	0.951	0.049	0.951	0.951	0.951	0.981
	PART	10-fold	94.86%	0.949	0.051	0.949	0.949	0.949	0.983
		66%	94.29%	0.943	0.057	0.944	0.943	0.943	0.980
	MLP	10-fold	81.94%	0.819	0.181	0.821	0.819	0.819	0.904
		66%	81.74%	0.819	0.181	0.820	0.819	0.819	0.904

As shown from the performance result, the bagging of J48 algorithm with 10-fold cross validation scored an accuracy of 95.8705%. This result shows that out of the total 50515 records used for training 48429 (95.8705%) records were correctly classified instances and 2086 (4.1295%) of the records were incorrectly classified instances. In second case using percentage split of 66%: 16330 (95.08%) records were correctly classified instances, while 845 (4.9199%) of the records were incorrectly classified instances. Generally, when we compare from the two test option conducted, the bagging J48 decision tree algorithms with 10-fold cross validation and bagging J48 decision tree algorithms with default percentage split (66%), the model developed with 10-fold cross validation algorithm gives a better classification performance of identifying newly occurring voice traffic termination fraud than default percentage split (66%). The snapshot running information for this experiment is shown in Annex-4.

As shown from table 4.2 bagging PART algorithm with 10 fold cross validation achieved an accuracy of 94.861%. This result shows that out of the total 50515 datasets used for training 47919 (94.861%) records were correctly classified instances, while 2596 (5.1391%) of the records were incorrectly classified instances. In second case using 66% split test option in

WEKA: 16195 (94.294%) records were correctly classified instances and 980 (5.706%) of the records were incorrectly classified instances. From the results it can be seen that bagging PART algorithm with 10-fold cross validation has relatively better performance than the percentage split case. The snapshot running information for this experiment is shown in Annex-5.

As shown from the performance result of bagging MLP algorithm with 10-fold cross validation scored an accuracy of 81.94%. Which means that out of the total 50515 records used for training 41393 (81.94%) records were correctly classified instances, while 9122 (18.058) of the records were incorrectly classified instances. In second case using 66% split: 14069 (81.9156%) records were correctly classified instances, while 3106 (18.0844%) of the records were incorrectly classified instances. This shows that the experiment conducted with the percentage split (66%) is better performance than the experiment conducted with 10-fold cross validation test option in WEKA. The model developed with the default percentage split of 66% gives a better classification accuracy of identifying newly occurring voice traffic termination fraud. The snapshot running information for this experiment is shown in Annex-6.

4.1.3 Model Building Experiment Using Stacking Method

The third ensemble learning classifier applied in this research was stacking. Stacking or stacked generalization combines several classifiers using the stacking method [36]. It is a simple extension to voting that can be used for classification or regression. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features. It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms. It can be used with the classification algorithms such as J48 and PART, J48 and MLP, PART and MLP and J48, PARTand MLP algorithms. In this study stacking aims to achieve the highest generalization accuracy. The model is built based on the default values of 66% percentage split and 10-fold cross validation.

Experiment Three:

This third experiment was designed to evaluate the performance of stacking ensemble method. In this experiment also the 10-fold cross validation and percentage split model evaluation techniques were used. The outcome of this experiment is presented in table 4.3 below.

Table 4. 3 Detailed performance measures for experiment Three

Method	Algorithm	Test Option	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
Stacking	J48 & PART	10-fold	92.11%	0.921	0.079	0.922	0.921	0.921	0.931
		66%	91.63%	0.916	0.084	0.916	0.916	0.916	0.916
	J48 % MLP	10-fold	81.59%	0.816	0.184	0.816	0.816	0.816	0.821
		66%	80.30%	0.803	0.198	0.819	0.803	0.800	0.803
	PART & MLP	10-fold	81.59%	0.816	0.184	0.816	0.816	0.816	0.844
		66% split	81.46%	0.815	0.186	0.816	0.815	0.814	0.873
	J48, PART & MLP	**10-fold**	**92.16%**	**0.922**	**0.078**	**0.922**	**0.922**	**0.922**	**0.954**
		66%	91.67%	0.917	0.083	0.917	0.917	0.917	0.935

As shown from the performance result, the stacking of J48 and PART algorithm with 10-fold cross validation scored an accuracy of 92.11%. This result shows that out of the total training datasets (50515), 46527 (92.11%) records were correctly classified instances and 3988 (7.89%) of the records were incorrectly classified instances. Using percentage split of 66%: 15738 (91.63%) records were correctly classified instances, while 1473 (8.3668%) of the records were incorrectly classified instances. This shows that the experiment conducted with 10-fold cross validation test option in WEKA is better performance than the experiment conducted with the percentage split (66%) test option. The snapshot running information for this experiment is shown in Annex-7.

As shown from the performance result of stacking J48 and MLP algorithm with 10-fold cross validation scored an accuracy of 81.58%. Which means that out of the total 50515 records used for training 41215 (81.58%) records were correctly classified instances, while 9300 (18.41%) of the records were incorrectly classified instances. In second case using 66% split: 14069 (80.30%) records were correctly classified instances, while 3383 (19.69%) of the

records were incorrectly classified instances. This shows that the experiment conducted with the 10-fold cross validation test option in WEKA is better performance than the experiment conducted with percentage split of 66%.

As shown from table 4.3, the stacking of PART and MLP algorithm with 10-fold cross validation scored an accuracy of 81.59%. This result shows that out of the total training datasets (50515), 41215 (81.59%) records were correctly classified instances, while 9300 (18.4104%) of the records were incorrectly classified instances. Using 66% split: 13991 (81.46%) records were correctly classified instances, while 3184 (18.53%) of the records were incorrectly classified instances. This also shows that the experiment conducted with the 10-fold cross validation test option in WEKA is better performance than the experiment conducted with percentage split of 66%.

As shown from the performance result, the stacking of J48 and PART, J48 and MLP, PART and MLP and J48, PART and MLP algorithms with 10-fold cross validation scored an accuracy of 92.16%. This result shows that out of the total training datasets (50515), 46555 (92.46%) records were correctly classified instances and 3960 (7.84%) of the records were incorrectly classified instances. Using percentage split of 66%: 45745 (91.67%) records were correctly classified instances, while 1430 (8.33%) of the records were incorrectly classified instances. The snapshot running information for this experiment is shown in Annex-8.

4.1.4 Model Building Experiment Using Voting Method

The fourth ensemble learning classifier applied in this research was voting. Voting is the simplest ensemble algorithm, and is very effective. It can be used for classification and regression problems. Voting works by creating two or sub-models. Each sub-model makes predictions which are combined in some way, such as by taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be [56]. It's also possible that the voting ensemble results in a better overall score than the best of the base estimators, as it aggregates the predictions of multiples models and tries to cover for potential weakness of the individual models.

Experiment Four:

The fourth experiment was designed to evaluate the performance of voting ensemble method. In this experiment the 10-fold cross validation and the default percentage split test option in

WEKA were used. Table 4.4 shows the resulting performance measurement of voting ensemble method with 10-fold cross validation and default percentage split of the model.

Table 4. 4 Detailed performance measures for experiment four

Method	Algorithm	Test Option	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
Voting	J48 & PART	**10-fold**	**95.23%**	**0.948**	**0.054**	**0.953**	**0.952**	**0.952**	**0.981**
		66%	94.62%	0.946	0.054	0.947	0.946	0.946	0.978
	J48 & MLP	10-fold	94.39%	0.944	0.056	0.945	0.944	0.944	0.967
		66%	94.05%	0.941	0.060	0.941	0.941	0.941	0.961
	PART & MLP	10-fold	92.23%	0.922	0.078	0.923	0.922	0.922	0.963
		66% split	91.62%	0.916	0.084	0.917	0.916	0.916	0.959
	J48, PART & MLP	10-fold	94.36%	0.944	0.056	0.944	0.944	0.944	0.975
		66%	93.74%	0.937	0.063	0.938	0.937	0.917	0.971

From the table it can be see that, the voting of J48 and PART algorithm with 10-fold cross validation scored an accuracy of 95.23%.Which means that out of the total training datasets (50515), 48106 (95.23%) records were correctly classified instances, while 2409 (4.77%) of the records are incorrectly classified instances. Using default percentage split (66%) test option in WEKA out of the 17175 testing records 16251 (94.62%) of them were correctly classified instances. Only 924 (5.38%) records are incorrectly classified instances. From the results it can be seen that voting (J48 and PART) with 10-fold cross validation has relatively better performance than the percentage split case. The snapshot running information for this experiment is shown in Annex-9.

As shown from table 4.4, voting (J48 and MLP) algorithm with 10 fold cross validation achieved an accuracy of 94.39%. This result shows that out of the total 50515 records used for training 47686 (94.39%) records were correctly classified instances, while 2829 (5.94%)

of the records were incorrectly classified instances. In second case using 66% split test option in WEKA: 16154 (94.05%) records were correctly classified instances and 980 (5.706%) of the records were incorrectly classified instances. The snapshot running information for this experiment is shown in Annex-10.

As shown from the performance result, the voting of PART and MLP algorithm with 10-fold cross validation scored an accuracy of 92.23%. This result shows that out of the total (50515) records used for training, 46593 (92.23%) records were correctly classified instances and 3922 (7.76%) of the records were incorrectly classified instances. Using percentage split of 66%: 15736 (91.62%) records were correctly classified instances, while 1439 (8.38%) of the records were incorrectly classified instances. This shows that the experiment conducted with 10-fold cross validation test option in WEKA is better performance than the experiment conducted with the percentage split (66%) test option.

As shown from the performance result of voting (J48, PART and MLP) algorithm with 10-fold cross validation scored an accuracy of 94.36%. Which means that out of the total training datasets 47666 (94.36%) instances were correctly classified and 2849 (5.64%) instances were incorrectly classified from 50515 testing instances. In second case, using percentage split of 66%: 16098 (93.73%) instances were correctly classified and 1077 (6.27%) instances were incorrectly classified from 17175 of testing instances. From the results it can be seen that voting (J48, PART and MLP) with 10-fold cross validation test option has relatively better performance than with 66% percentage split.

4.1.5 Model building experiment using all attributes without Ensemble

Experiment Five:

This fifth experiment was designed to evaluate the performance of individual algorithms without ensemble method. In this experiment also the 10-fold cross validation and percentage split model evaluation techniques were used. The outcome of this experiment was presented in table 4.4 below.

Table 4. 5 Detailed performance measures for experiment four

Method	Algorithm	Test Option	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
Without ensemble	**J48**	**10-fold**	**94.72%**	**0.947**	**0.053**	**0.948**	**0.947**	**0.947**	**0.972**
		66%	94.15%	0.942	0.059	0.942	0.942	0.967	0.964
	PART	10-fold	92.32%	0.923	0.077	0.925	0.923	0.923	0.974
		66%	91.63%	0.916	0.084	0.916	0.916	0.916	0.970
	MLP	10-fold	81.71%	0.923	0.077	0.925	0.923	0.923	0.974
		66%	81.74%	0.817	0.083	0.817	0.817	0.817	0.902

In this experiment the 10-fold cross validation and percentage split model evaluation techniques were used. In 10-fold cross validation test option in WEKA: 47846 (94.7164%) instances were correctly classified, while 2669 (5.2866%) instances were in correctly classified from 50515 testing instance. In second case, using percentage split of 66%: 16171 (94.1573%) instances were correctly classified, while 1004 (4.8457%) instances were incorrectly classified from 17175 of testing instance. From the results it can be seen that decision tree classifier (J48) with 10 fold cross validation has relatively better performance than 66% percentage split. The snapshot running information for this experiment is shown in Annex-11.

As shown from table 4.5, PART rule induction classifier with 10-fold cross validation achieved an accuracy of 92.3191. This result shows that out of the total training datasets 406635 (92.3191%) instances were correctly classified and 3880 (7.6809%) instances were incorrectly classified from 50515 testing instance. In second case, using percentage split of 66%: 17738 (91.6332%) instances were correctly classified and 1437 (8.3668%) instances were incorrectly classified from 17175 of testing instance. From the results it can be seen that PART rule induction algorithm with 10 fold cross validation has relatively better performance

than 66% percentage split. The snapshot running information for this experiment is shown in Annex-12.

As shown from the performance result of Multi-layer perceptron (MLP) algorithm with 10-fold cross validation scored an accuracy of 81.7124%. Which means that out of the total training datasets 41277 (81.71%) instances were correctly classified and 9238 (18.29%) instances were incorrectly classified from 50515 testing instance. In second case, using percentage of 66%: 14040 (81.74%) instances were correctly classified and 3135 (18.25%) instances were incorrectly classified from 17175 of testing instance. From the results it can be seen that multi-layer perceptron algorithm with 66% percentage split has relatively better performance than 10-fold cross validation. The snapshot running information for this experiment is shown in Annex-13.

4.2 Evaluation and comparison of the algorithm

In any branch of science, it is almost a common requirement that performance of various model have to be compared with each other to understand the suitability of a model to a given problem. In this section also after performing the experiments comparing the model and selecting the best model is the task to be done. Basically the experiments were conducted on all attributes that selected by discussion with domain experts and purpose of the research. The models were compared using different performance measures like correctly classified instance, incorrectly classified instance, ROC area analysis. Table 4.6 and table 4.7 summarized the performance of each classifier on the basis of which the best performing classifier was selected.

Table 4. 6 Performance comparison of the selected models with ensemble methods)

Method	Algorithm	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.F-measure	Av.ROC area
AdaBoost	**J48**	**96.73%**	**0.967**	**0.033**	**0.967**	**0.967**	**0.967**	**0.988**
	PART	96.43%	0.964	0.036	0.964	0.964	0.964	0.987
	MLP	81.74%	0.817	0.183	0.817	0.817	0.817	0.888
Bagging	J48	95.87%	0.959	0.041	0.959	0.959	0.959	0.985
	PART	94.861%	0.949	0.051	0.949	0.949	0.949	0.983
	MLP	81.94%	0.819	0.181	0.821	0.819	0.819	0.904

Stacking	J48 & PART	92.11%	0.921	0.079	0.922	0.921	0.921	0.931
	J48 & MLP	92.32%	0.923	0.077	0.925	0.923	0.923	0.974
	PART & MLP	81.59%	0.816	0.184	0.816	0.816	0.816	0.844
	J48, PART & MLP	92.16%	0.922	0.078	0.922	0.922	0.922	0.954
Voting	J48 & PART	95.23%	0.948	0.054	0.953	0.952	0.952	0.981
	J48 & MLP	94.39%	0.944	0.056	0.945	0.944	0.944	0.967
	PART & MLP	92.23%	0.922	0.078	0.923	0.922	0.922	0.963
	J48, PART & MLP	93.74%	0.937	0.063	0.938	0.937	0.917	0.971

Table 4. 7 Performance comparison of the selected models without ensemble methods

Method	Algorithm	Accuracy	Av.TP rate	Av.FP rate	Av.Precision	Av.Recall	Av.measure	Av.ROC area
Without Ensemble	J48	94.72%	0.947	0.053	0.948	0.947	0.947	0.972
	PART	92.32%	0.923	0.077	0.925	0.923	0.923	0.974
	MLP	81.74%	0.817	0.083	0.817	0.817	0.817	0.902

In this study, comparison of classification techniques with ensemble method and single algorithm were performed. As described above one of the basic aims of data mining is to compare different models and to select the better classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. As a result, the best classification classifier which is appropriate for this problem domain has been selected. The above experimental result shows that in terms of average positive rate the model built based on boosted J48 decision tree algorithm shows 0.033 and which is very low compared to single algorithm and in performance registered of correctly classified instances in both test modes boosted J48 decision tree algorithm from ensemble methods is the highest one.

As the above table 4.6 shows the boosted J48 decision tree algorithm classified 48863 instances which is 96.73% correctly classified with average false positive rate of 0.033 with 10-fold cross validation. Furthermore, different experiments are done using 10-fold cross validation and 66% split. The experiment results shows that with 10-fold cross validation on single algorithm J48 decision tree and PART algorithm achieved better result. J48 decision

tree algorithm produces 94.72% prediction accuracy followed by PART algorithm with 92.32%. Boosting ensemble using 10-fold cross validation J48 decision tree and PART produced 96.73% and 96.43% prediction accuracy respectively.

In this study, ensemble method given better prediction accuracy rather than single algorithm with 10-fold cross validation and 66 percentage split testing method. In this study, ensemble methods are more efficient than individual algorithm. For this dataset, boosting ensemble works best than bagging, stacking and voting. But, bagging, stacking and voting are better than single algorithm. Hence boosted J48 decision tree algorithm with 10-fold cross validation is used as a model of this study because it has high performance than all the other algorithms study in this research. From the above results, the overall performance of the 10-fold cross validation experiment was better than the other options. The snap shot running information boosted J48 decision tree algorithm with 10-fold cross validation is shown in annex-1. The resulting confusion matrix for the best model (boosted J48 decision tree) is presented in table 4.8 below.

Table 4. 8 Confusion Matrix result for Boosted J48 decision tree algorithm with10-fold cross validation testing option

Confusion Matrix		
A	B	Classified as
24313	953	No= Not fraud
699	24550	Yes= Fraud

The entries in the confusion matrix have the following meaning:

- 24313 is the number of correct predictions that an instance is No (Non-fraud) TP
- 953 is the number of incorrect predictions that an instance is Yes (Fraud) FP
- 699 is the number of incorrect predictions that an instance is No (Non-fraud) FN
- 24550 is the number of correct predictions that an instance is Yes (Fraud) TP

24313 and 24550 values of the confusion matrix are the correctly predicted results of the classifier, whereas the values 953 and 699 are incorrectly predicted results. In 10-fold cross validation 953 instances are misclassified as yes (fraud) which were actually class of no (non-fraud) and 699 instances are misclassified as no (non-fraud) which were actually class of yes (fraud). As discussed with domain experts, the reason for the missclassifacion of the two classes was if fraud case occur, there is also a possibility that non-fraud to be occurred. Some error measures are more useful than others. In selecting the best algorithms and parameters

generated the best model for voice traffic termination fraud detection, the following performance metrics were also used.

Table 4. 9 Evaluation on ensemble and single algorithm of prediction error

Method	learning algorithm	Mean absolute error (MAE)	Root mean squared error (RMSE)
Single algorithm	J48	0.08	0.21
Boosting	**J48**	**0.03**	**0.17**

The above table 4.9 showed that boosting classifiers has led to considerable decrease of prediction error from (Mean absolute error = 0.08) to (Mean absolute error = 0.03) compared to single algorithm with all attributes. Also the prediction error in boosting decrease from (Root mean squared error = 0.21) to (Root mean squared error = 0.17) compared to single algorithm.

4.3 Discussion of the results with domain experts

This section includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts and checking the impact of the discovered knowledge. Usually real world databases contain incomplete noisy and inconsistent data and such unclean data may cause confusion for the data mining process [36]

The rules indicate the possible conditions in which the CDR record could be classified in each of the fraud and non-fraud of telecom subscribers. From the generated rules it is observed that most determinant factors are number of cells accessed by the subscriber followed by number of mobile BTS nodes accessed by subscriber and number of outgoing calls made by the subscriber. The domain experts argued that, the discovered rules are important and great impact on telecom voice traffic termination fraud detection.

In general, in this study we tried to revealed different related works concerning the issue of telecom voice traffic termination fraud. Under this study depending on the knowledge generated by the boosted J48 decision tree algorithm it has been found that call type id for outgoing and incoming call, number of outgoing calls originated from the given subscriber, number of unique subscribers called, number of distinct cells accessed by the subscriber of study become a new finding that the other researcher didn't consider it. The other findings are also improvements of a rule extraction techniques, resulting in increased comprehensibility and more accurate result by ensemble machine learning.

4.4 Specific Rule Extraction

The model developed with boosted J48 decision tree classifier was selected as the best model for this study and this model generated significant rules that are useful for prediction of voice traffic termination fraud. The rules generated from the models can help Ethio telecom or other telecom companies to revise or cross check the voice traffic termination fraud detection policies and also help managements for easy and quick decision making. From this model a set of rules are extracted by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node and rules that contain most instances of the dataset were extracted. The following are some of the rules derived from the models as presented below and some interpretation is given for some of the rules. After selecting that rules the researcher turned to the domain expert for discussion. To extract rules boosted J48 decision tree is more human-readable and easy with each path. If the condition is satisfied then the conclusion follows.

Rule 1: If TOT_NO_CELL-ID >=7 AND CELL_A >= 636011500614211 AND CALL RATIO = 0.75 AND TOT_IN_Call >1 AND TOT_OUT_Call <= 26, then Class=Yes (Fraud).

This rule show, if the total number of cells accessed by the subscriber (number of BTS nodes accessed by the subscriber) is greater than 7, the subscriber accesses BTSs whose IDs are higher than 636011500614211, the call ratio is 0.75, the total number of incoming call is no more than 1 and the total number of outgoing calls made by the subscriber is not more than 4 then there is a more probability of the subscriber to be predicted as a fraudulent one.

Rule 2: If CALL_RATIO >= 0.04 AND TOT_NO_CELL-ID >=73 AND TOT_OUT_call <= 26 AND TOT_NO_CELL-ID <= 1, then Yes (Fraud)

This rule show, if call ratio is more than 0.04 and total number of cells is greater than or equal to 73, total number of outgoing calls made by the subscriber for a random range of two months is not more than 26 and the total number of BTS nodes accessed by the subscriber is not more than 1, then there is no chance of the subscriber to be predicted as a fraudulent.

Rule 3: If CELL_A > 636012214322992 AND DURATION = high, then Yes (Fraud)

This rule show, if a subscriber accesses BTSs who's IDs are higher than 636012214322992 and duration of the call is high, then there is a more probability of the subscriber to be predicted as a fraudulent one.

Rule 4: If TOT_NO_CELL-ID > 4 AND DIS_OUT_CALL <= 1 AND TOT_OUT_call <= 3, then Yes (Fraud).

This rule show, if the total number of BTS nodes accessed by the subscriber is more than 4 and the total number of outgoing calls made by the subscriber is not more than 1, then there is a more likely chance of the subscriber to be predicted as a fraudulent one.

Rule**:** 5 If CELL_A > 636012214322992 AND DURATION = low, then No (Non fraud).

This rule show, if the IDsof the BTS nodes accessed by the subscriber during the experimentation period is greater than 636012214322992 and the duration of the call is low, then then the call instances encountered are of non-fraudulent type.

Rule 6: If TOT_NO_CELL-ID > 1012 AND CELL_A <= 636011200214567, then No (Non fraud)

This rule show, if the total number of cells accessed by the subscriber is greater than 1012 and the IDs of those cells falls below 636011200214567, then the call instances encountered are of non-fraudulent type.

Rule 7: If TOT_NO_CELL-ID > 447, then No (Non fraud).

This rule show, if number of distinct cells accessed by the subscriber is greater than 447 then the call instances are of non-fraudulent type.

Rule 8: If CELL_A > 636011100213355 AND TOT_NO_CELL-ID <= 22, then Yes (Fraud).

This rule show, if a subscriber accesses BTSs who's IDs are higher than 636011100213355 and the total number of BTS nodes accessed by the subscriber is not more than 22, then there is a more probability of the subscriber to be predicted as a fraudulent one.

Rule 9: If TOT_NO_CELL-ID > 22 then No (Non fraud).

This rule show, the total number of BTS nodes accessed by the subscriber is greater than 22, then there is no chance of the subscriber to be predicted as a fraudulent.

Rule 10: If CELL_A > 636011101716846 AND TOT_NO_CELL-ID <= 69 AND CELL_A <= 636011102412501 AND CELL_A<=636011102113973 AND CELL_A <= 636011101813746 AND TOT_NO_CELL-ID <= 38, then Yes (Fraud).

This rule show, if the IDs of the BTS nodes accessed by the subscriber during the experimentation period falls in between 636011101716846 and 636011102412501, and the total number of BTS nodes is not more than 69, then there is a more likely chance of the subscriber to be predicted as a fraudulent one.

Rule 11: If CELL_A > 636011400218181 AND CELL_A<= 636011400218183 AND TOT_NO_CELL-ID <= 36 then Yes (Fraud).

This rule show, if the IDs of the BTS nodes accessed by the subscriber lies between 636011400218181 and 636011400218183, and the total number of cells is no more than 36, then the subscriber can be predicted as a fraudulent one.

Rule 12: If TOT_NO_CELL-ID > 34 AND CELL_A <= 636011400230093 AND CELL_A <= 636011400218181 AND CELL_A <= 636011400213217 then Yes (Fraud).

This rule show, if the total number of cells accessed by the subscriber is more than 34 and the IDs of those cells lies below 636011400230093, then the subscriber can be predicted as a fraudulent one.

Rule 13: If TOT_NO_CELL-ID > 158 AND TOT_OUT_call <= 13 AND TOT_NO_CELL-ID <= 186 then Yes (Fraud).

This rule show, if the total number of outgoing calls made by the subscriber is not more than 13 and the total number of cells accessed by the same subscriber is less than or equal to 186, then the call instances encountered are of fraudulent type.

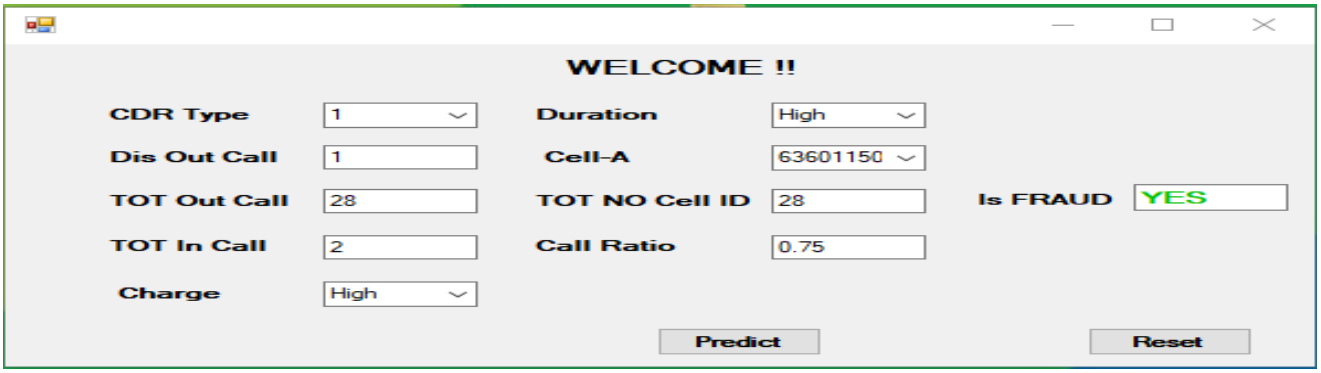
From the generated rules it is observed that most determinant factors are number of cells accessed by the subscriber followed by number of mobile BTS nodes accessed by subscriber and number of outgoing calls made by the subscriber. The domain experts argued that, the

discovered rules are important and great impact on telecom voice traffic termination fraud detection.

4.5 Using Discovered Knowledge

After evaluating the discovered knowledge, the last step is using this knowledge for the industrial purposes. In this step the knowledge discovered is incorporated into performance system and take this action based on the discovered knowledge. In this research the discovered knowledge is used by integrating the user interface which is designed by C# programing language with a Weka system in order to show the prediction of voice traffic termination fraud. As it had been described in section 4.4 above the rules are listed as IF-THEN rule. The designed user prototype accept user query and suggest fraud status. Figure 4.1 below show the user interface that enable user interaction with the user interface and sample of the code is given in the annexes part.

Fig 4. 2 Screen shoot of prototype developed for the rule generated



4.6 Validity of user acceptance testing

After we developed a prototype that can predict voice traffic termination fraud status, we prepared a questioner (attached in annex-15) to check the validity of the prototype. In this study, a total of 6 domain experts from technical audit or IT management audit specifically from Information technology and Network security department were participated to evaluate the systems acceptance. Each of the participants are asked to give feedback on the acceptability of the prediction and to rate it on a scale of 1 (Strongly disagree) to 5 (Strongly agree). Summary of the result was presented in table 4.10 below.

Table 4. 10 Experts response summary on the proposed prediction model

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency: ▪ In terms of time (reducing time to accomplish the tasks) ▪ In terms of accuracy	95%	5%	–	–	–
	90%	10%	–	–	–
Effectiveness': ▪ In terms of the output result (quality of work output) ▪ In terms of performance ▪ In terms of error tolerance	85%	15%	–	–	–
	95%	10%	–	–	–
	70%	15%	5%	5%	5%
Easy to understand: ▪ In terms of the feature of user interface (easy to learn) ▪ In terms of platform	90%	10%	–	–	–
	80%	20%	–	–	–
Easy to remember: ▪ In terms of remembering the way to use the prototype ▪ In terms of users productivity	85%	15%	–	–	–
	90%	10%	–	–	–

According to this study, the performance result of the prediction model scored an accuracy of 96.73%. Based on discussion with domain experts, most of the domain experts satisfied with the prediction results. In order to make the prediction more accurate and error tolerable, we advise integration of the discovered classification rules with knowledge based system. In the overall user acceptance criteria, more than 85% of the domain experts agreed that this prediction model is efficient, produces a desirable result, user friendly, easy to understand, remember, increase user's productivity, quality of work output and reducing time to accomplish the tasks. The domain experts also suggested that there need to enhance the performance of the model to make the prediction near to 100%. Also even if the purpose of the study is academic purpose and the use of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the company can use it.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

In this thesis an effort has been made to examine the effect of ensemble learning, specifically, ensemble and single based classification techniques. The hybrid, iterative methodology was employed in this study which consists of six basic steps such as problem domain understanding, data understanding, data preparation, data mining, evaluation and use of the discovered knowledge.

The researcher selected around 126765 records from the huge CDR database. After eliminating irrelevant and unnecessary data only a total of 50515 datasets were used for the purpose of conducting this study. Ten attributes were selected from initial attributes or the fields. It has been preprocessed and prepared in a format suitable for the DM tasks. The study was conducted using WEKA 3.8.3 machine learning software and four ensemble-based machine learning paradigms for classification techniques was used, namely boosting, bagging, stacking and voting methods, based on 2 basic learners (decision tree and neural network) classifiers. These classifier algorithms were used because it is recommended and commonly used in different literature reviews then do the comparison and selection of the best classifier based on their performance.

The performances of the model in this study were evaluated using the standard metrics of prediction accuracy, FP rate, TP rate, recall, precision and F-measure and ROC curve which are calculated using the predictive classification table, known as Confusion matrix. Error rate analysis was also used to compare the performances of classifies. Percentage split and 10-fold cross validation model evaluation techniques was used for training and testing the model. All models that built performed well in predicting voice traffic termination fraud behavior. The most effective model to predict voice traffic termination fraud behavior is boosted j48 decision tree using 10-fold cross validation model evaluation techniques implemented on all attributes with classification accuracy of 96.73%. While considering the method which increases the efficiency of j48 method the most, we find that BOOSTING does this much efficiently than others. Other ensembles like BAGGING, STACKING and VOTING also increase the

efficiency of j48. Finally, on the result of experiment we discussed with domain experts and they confirm the rules generated are useful for voice traffic termination fraud prediction. The rules generated from the models can help Ethio Telecom or other telecom companies to revise or cross check the telecom fraud detection policies that they have. In this study, we found that the proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers.

5.2 Recommendations

Data mining tools and techniques are being used to solve different types of problems in various industries, particularly in telecommunication data to support decision making. In this study data mining classification and prediction techniques were applied on Ethio telecom company datasets and a good performance was achieved in this technique. The researcher recommends the following points based on the outcome of the research.

- The technique employed in this study were boosting, bagging, stacking and voting classifiers, based on 2 basic learners (decision tree and neural network) algorithm. Even though an encouraging result was obtained, using other types of techniques with changing different parameter might perform better therefore; we recommended other researchers to test with other types of techniques like: ensemble other classification algorithms such as support vector machine, Bayesian network or using other machine learning software.
- We recommend that other researchers to conduct similar researches by using quality of the audio of the calls. But there is a need to check if the data is available for research.
- This research is only used to detect voice traffic termination fraud by using ensemble classification methods, but other researchers can also use similar process to detect other fraud types and methods.

REFERENCES

[1] Ethio -Telecom Company profile magazine, 2013.

[2] E. Taye, Telecommunication in Ethiopia. Geneva: UNCTAD, 2010.

[3] O. Abidogun, Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks (Doctoral dissertation, University of the Western Cape), 2005.

[4] N. Asfaw, Challenges Facing International Telecom Business and the Way Forward, Ethiopian Telecommunication Corporation's Perspectives. Master's Thesis (Telecom MBA), College of Telecommunication and Information Technology, Management Department, 2006.

[5] N. Adu-Boafo, '' the big issue: Perspective on SIM Box Fraud in Ghana,'' Africa Telecom & IT, 4, 10-17, 2013.

[6] M. Gary, ''Data Mining in Telecommunications. In O. Maimon and L. Rokach, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,'' Kluwer Academic Publishers, 1189-120, 2005.

[7] N. Ojuka, ''Detection of subscription fraud in telecommunications using decision tree learning'', 2009.

[8] M. Govindarajan, "Intrusion detection using ensemble of classification methods," in I.J. Computer Network and Information Security, IEEE, pp. 45–53.8, 2014.

[9] O. Adnan, M. Abuassba, "Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines," Hindawi Computational Intelligence and Neuroscience Volume 2017, Article ID 3405463, 11 pages.

[10] S. Finlay, ''Multiple classifier architectures and their application to credit risk assessment,'' European Journal of Operational Research, 210, 368-378.10, 2011.

[11] G. Paleologo, A. Elisseeff, & G. Antonini, ‘’Subagging for credit scoring models, ’’European Journal of Operational Research, 201, 490-499.11, 2010.

[12] J. Shawe-Taylor, K. Howker, & P. Burge, ‘’Detection of fraud in mobile telecommunications,’’ Information Security Technical Report, 4(1), 16-28, 1999.

[13] L. Cortesao, F. Martins, A. Rosa, and P. Carvalho, “Fraud management systems in telecommunications: A practical approach,” in 12th Int. Conf. on Telecommun. (ICT), pp. 167–182, 2005.

[14] E. Tarmazakov and D. Silnov, “Modern approaches to prevent fraud in mobile communications networks,” in Conf. of Russian Young Researchers in Elect. And Electron. Eng. (EIConRus), IEEE, pp. 379–381, 2018.

[15] C. F. C. Association et al., “Global fraud loss survey,” Press Release, New Jersey, NJ (CFCA), vol. 10, p. 2013, 2017.

[16] H. Abdikarim and S. Roselina, ‘’classification of SIM Box fraud Detection using support vector machine and Artfical neural network. University of Technology Malaysia’’, 2014.

[17] G. Negarit, ‘’Telecom Fraud Offence Proclamation’’, 2012.

[18] E. Jonsson, E. Lundin, & H. Kvarnström, ‘’Combining fraud and intrusion detection-meeting new requirements’’ Paper presented at the NORDIC WORKSHOP ON SECURE IT SYSTEMS-NORDSEC, 2000.

[19] P. Estévez & C. Perez, ‘’ Subscription fraud prevention in telecommunications using fuzzy rules and neural networks’’ Expert Systems with Applications, 31(2), 337-344, 2006.

[20] A. Tariku, ‘’Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company,’’ AAU, Faculty of Informatics, Department of Information Science, 2011.

[21] S. Brown, ‘’Telecommunication fraud management: Wave road security’’, 2005.

[22] T. Cohen, & R. Southwood, ‘’an overview of VoIP regulation in Africa: policy responses and proposals, ’’Commonwealth Telecommunications Organization (CTO), 2004.

[23] I. Murynets, I, M. Zabarankin, P. Jover, and A. Panagia, “Analysis and detection of simbox fraud in mobility networks,” in INFOCOM, Proceedings IEEE, IEEE, 2014, pp. 1519–1526. 2014.

[24] M. Ghosh, ‘’Telecoms fraud, ’’Computer Fraud & Security, 14-17, 2010.

[25] J. Van Heerden, “Detecting fraud in cellular telephone networks (Doctoral dissertation, University of Stellenbosch)”, 2005.

[26] J. Brooks, M. Hussin, A. Soto, F. Jacob, V. Sinha, & T. Eisner, ‘’Fraud Classification Guide,’’ TM Forum GB954, 2012.

[27] T. Fawcett, & F. Provost, ‘’Adaptive fraud detection: Data mining and knowledge discovery, 1(3), 291-316’’, 1997.

[28] V. Vaithiyanathan, “comparison of different classification techniques using different datasets”, International Journal of Advances in Engineering & Technology, vol. 6, pp.764-768, May 2013.

[29] F. Gorunescu, “Data Mining: Concepts, Models and Techniques”. Springer, 2011.

[30] J. Han, ‘’Data mining techniques’’. In ACM SIGMOD Record, international conference on Management of data (Vol. 25, No. 2, p. 545). ACM, 1996.

[31] D. Olson, & D. Delen, ‘’Advanced data mining techniques’’. Springer, 2008.

[32] L. Rokach & O. Maimon, “Data mining with decision trees: Theory and Applications” World Scientific Publishing Co. Pte. Ltd, 2008.

[33] U. Fayyad, G. Piatetsky-Shapiro & P. Smyth, “From data mining to knowledge discovery in databases” AI magazine, 17(3), 37, 1996.

[34] J. Han, ‘’Data mining techniques’’. In ACM SIGMOD Record, international conference on Management of data (Vol. 25, No. 2, p. 545). ACM, 1996.

[35] R. Roiger & M. Geatz,” Data Mining: A Tutorial-Based Primer, Addison-

Wesley, ISBN 0-201-74128-8, Boston", 2003.

[36] J. Han & M. Kamber, ''Data Mining: Concepts and Techniques'', Second Edition. Elsevier Inc., 2006.

[37] M. Zaki & J. Meira, "Data Mining and Analysis: Fundamental Concepts and Algorithms". 2013.

[38] L. Rokach. "Ensemble-based classifiers". Springer Science and Business Media. [Online]. Available: 10.1007/s10462-009-9124-7, 2009.

[39] A. Tiwari and A. Prakash, "Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA", International Journal of Engineering and Technical Research (IJETR), vol. 2, pp. 207-209, Sept, 2014.

[40] Galar et.al. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transactions on systems, man, and cybernetics—part C: Applications and reviews, pp. 1-22, 2011.

[41] Q. Yang & X. Wu, "10 Challenging Problems in Data Mining Research". International Journal of Information Technology & Decision-Making Vol. 5, No. 4, pp. 597–604, World Scientific Publishing Company, 2006.

[42] U. Fayyad, G. Piatetsky, & P. Smyth, "Mining Scientific Data". Communications of the ACM, 39(11), pp. 51-57, 1996.

[43] P. Chapman, ''CRISP-DM 1.0: Step-by-step data mining guide,'' SPSS Inc., 2000.

[44] L. Kurgan & P. Musilek, "A survey of Knowledge Discovery and Data Mining process models" Knowledge Engineering Review, 21(1), 1-24, 2006.

[45] A. Azevedo, "KDD, SEMMA and CRISP-DM: a parallel overview", 2008.

[46] P. Burge & J. Shawe-Taylor, ''Detecting Cellular Fraud Using Adaptive Protypes'', 1997.

[47] G. Jember, ''Data mining application in supporting fraud detection on mobile

communication: the case of Ethio mobile'' Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia, 2005.

[48] G. Gebremeskel, ''Data mining application in supporting fraud detection: on Ethio-mobile services, '' Unpublished Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia, 2006.

[49] H. Birhanu, ''Fraud Detection in Telecommunication Networks Using Self-Organizing Map: The Case of Ethiopian Telecommunication Corporation (ETC),''Master's Thesis, College of Telecommunication and Information Technology, Department of Information Technology, 2006.

[50] F. Feven, ''Predictive Sim Box fraud detection model for Ethiotelecom, ''published Master, 2017.

[51] A. Kahsu, ''SIM Box fraud detection using data mining techniques,'' unpublished Master, 2018.

[52] G. Yeshinegus, ''Predictive Modeling for Fraud Detection in Telecommunications: The Case of Ethio-Telecom. AAU, (Master's thesis, Addis Ababa University), 2013.

[53] M .Ayash, "Research Methodologies in Computer Science and Information Systems", Alquds Open University College of Technology & Applied Science.

[54] H. Jantan, & Z. Othman, ''Human Talent Prediction in HRM using C4.5 Classification Algorithm''. International Journal on Computer Science and Engineering, 02(08), 2526–2534, 2010.

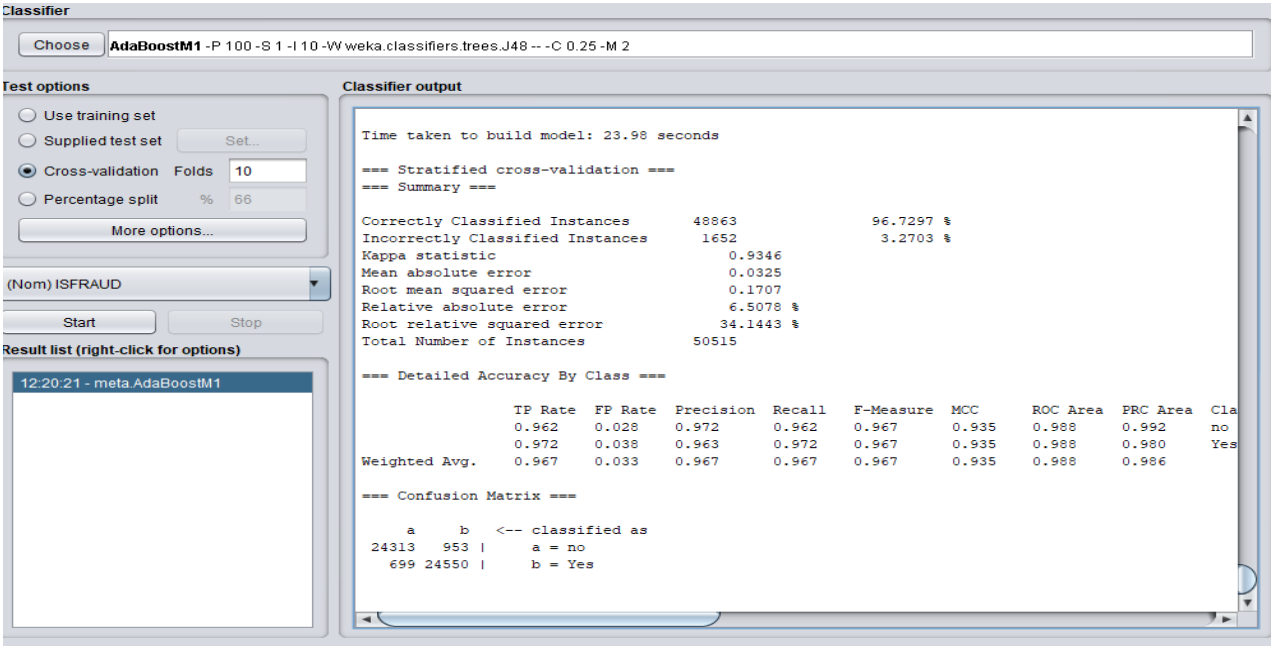
[55] Lusia and Suhartono, "Ensemble Method Based for Rainfall Forecasting in Indonesia "International Journal of Science and Research (IJSR), vol. 2, February 2013.

[56] L. Kabari & C. Onwuka, ''Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a classifier.'' International Journals of research in Computer Science and Software Engineering, ISSN: 2277-128X

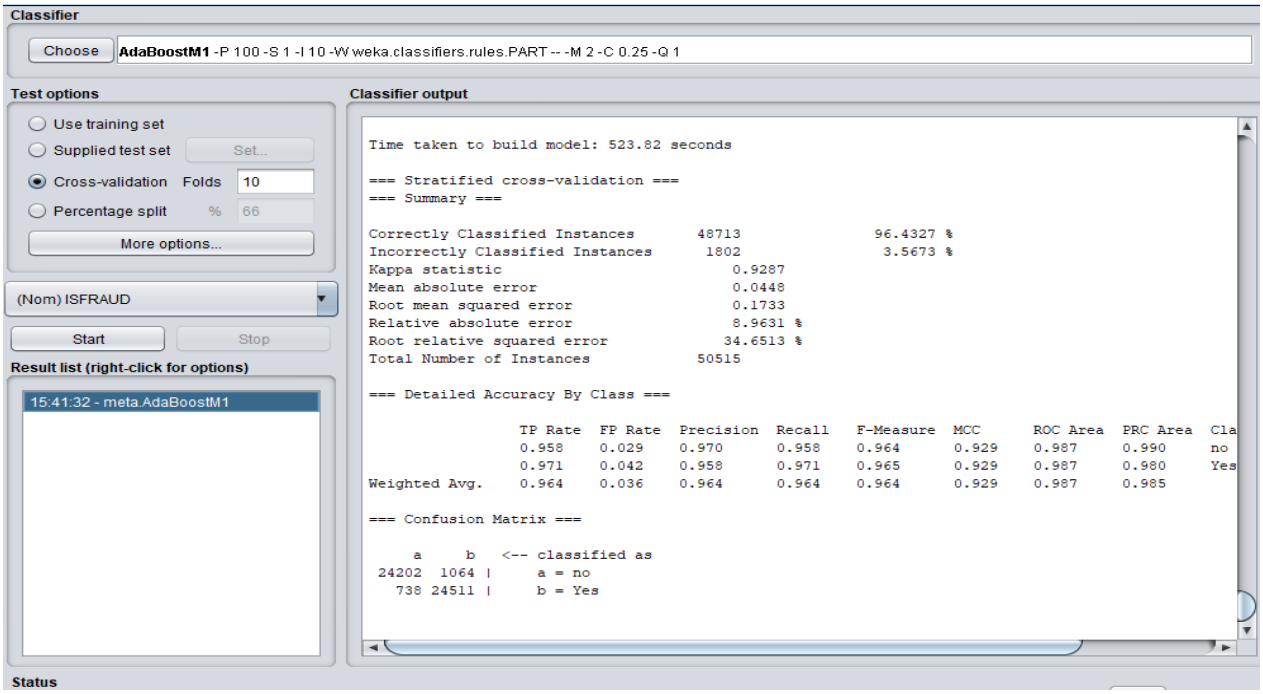
(Volume-9, Issue-3), March 2019.

ANNEXES

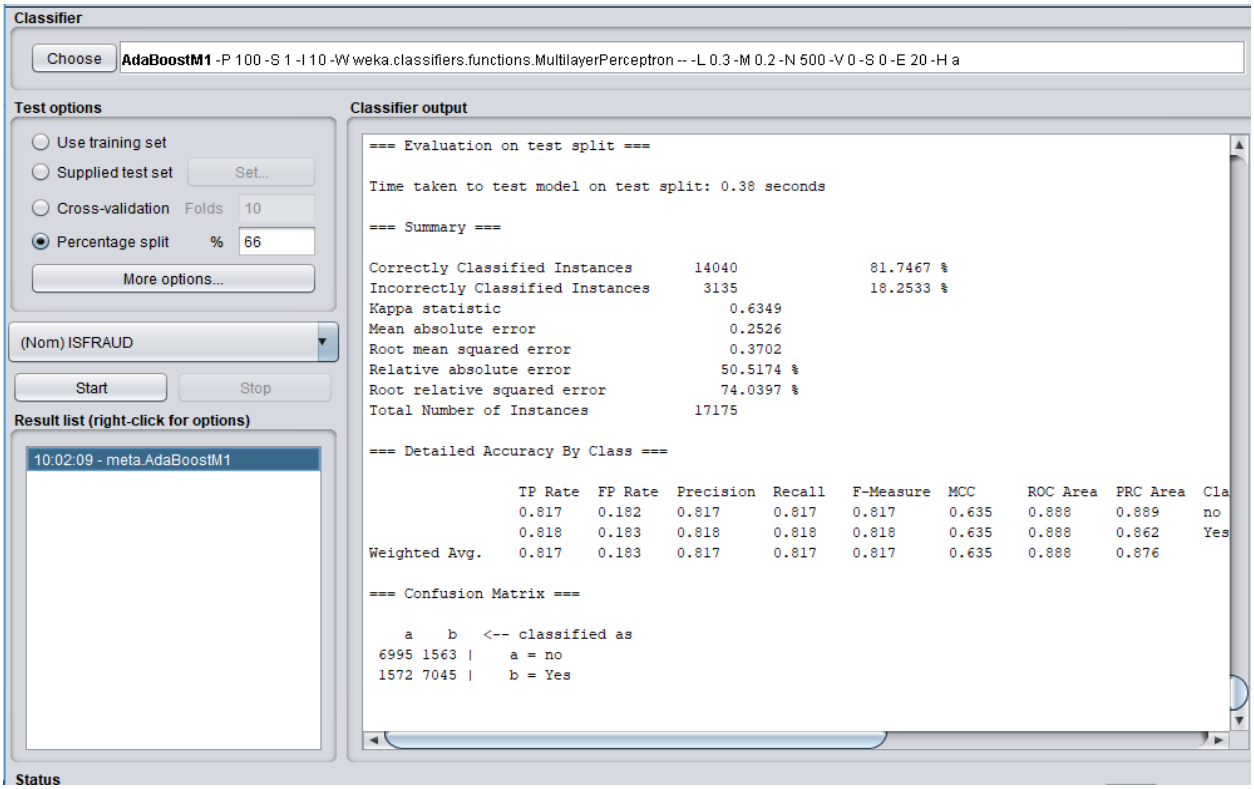
Annex-1: The snapshot running information of boosted J48 decision tree classifier with 10-fold cross validation technique



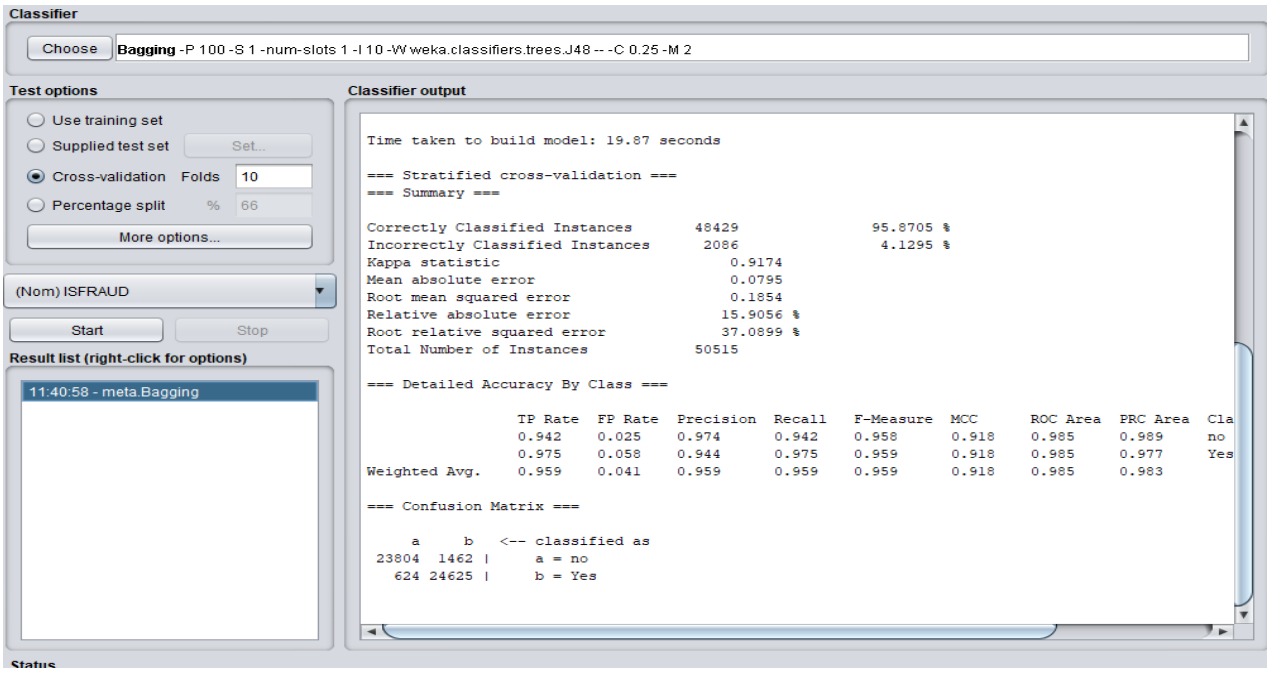
Annex-2: The snapshot running information of boosting ensemble method PART algorithm with 10-fold cross validation technique



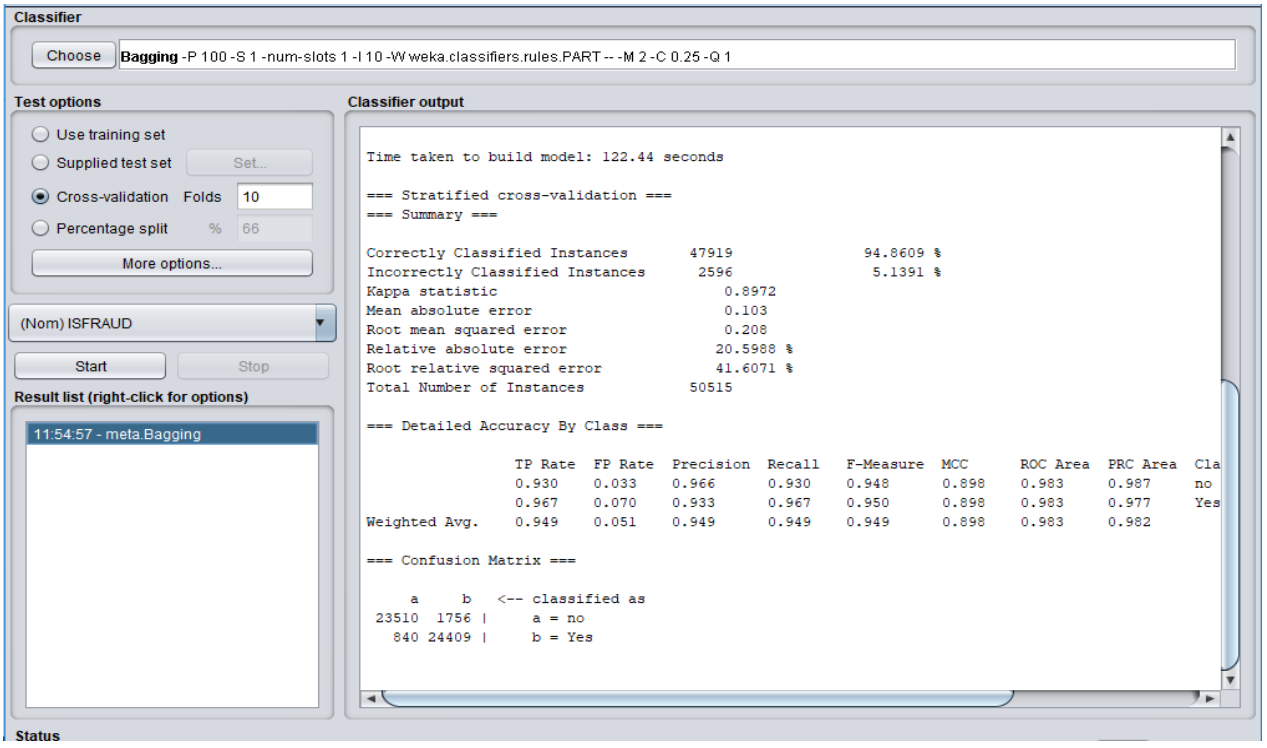
Annex-3: The snapshot running information of boosted MLP classifier with default percentage split of 66% validation technique



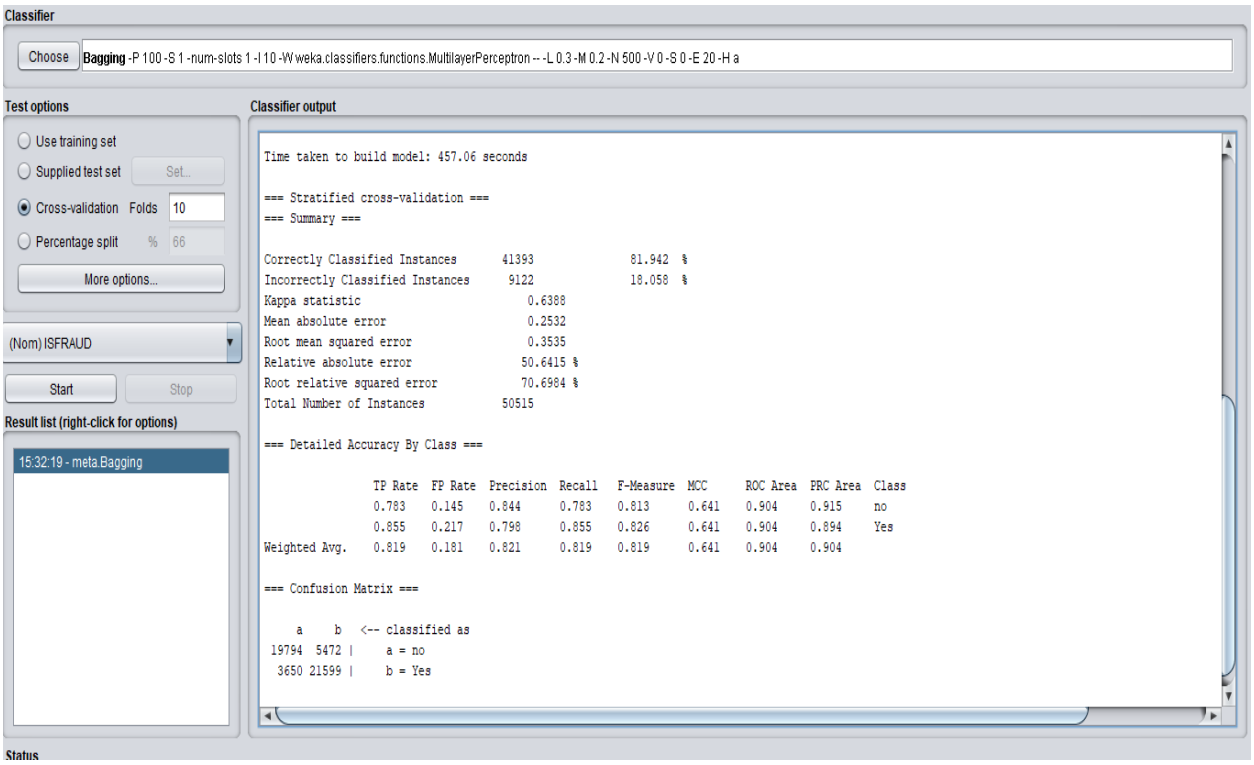
Annex-4: The snapshot running information of bagging J48 decision tree classifier with 10-fold validation technique



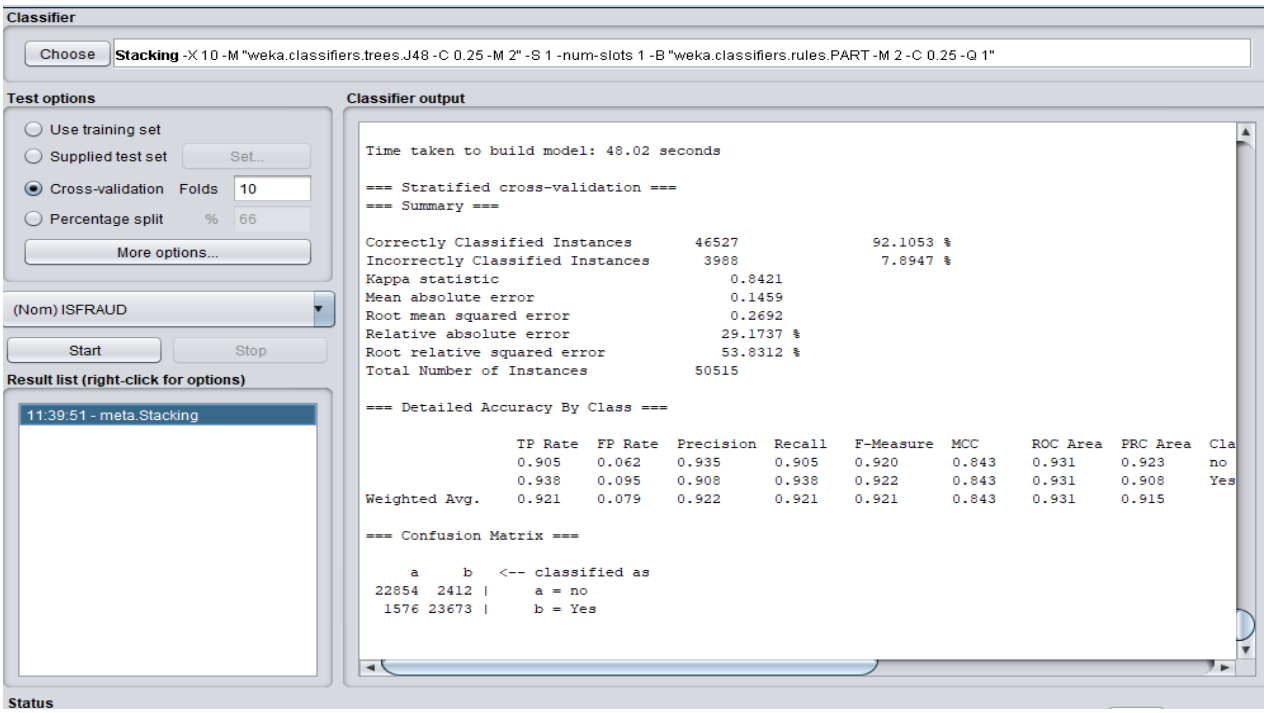
Annex-5: The snapshot running information of bagging PART classifier with 10-fold cross validation technique testing option



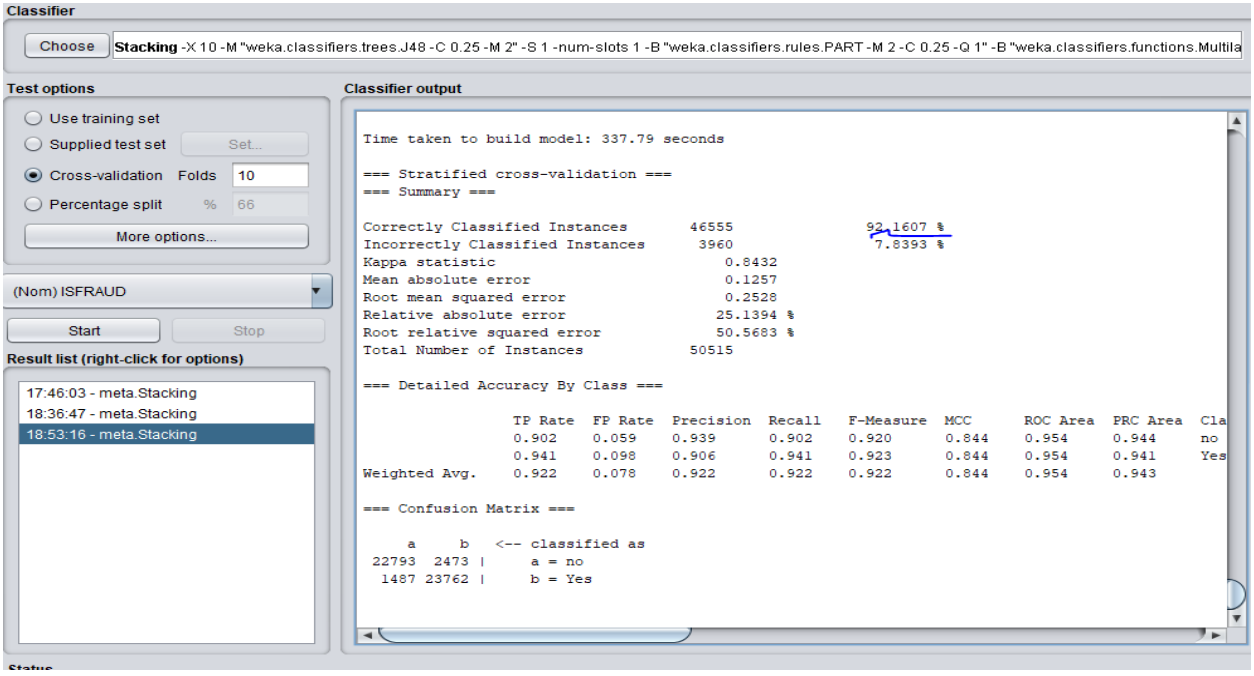
Annex-6: The snapshot running information of bagging ensemble method MLP algorithm with 10-fold cross validation technique



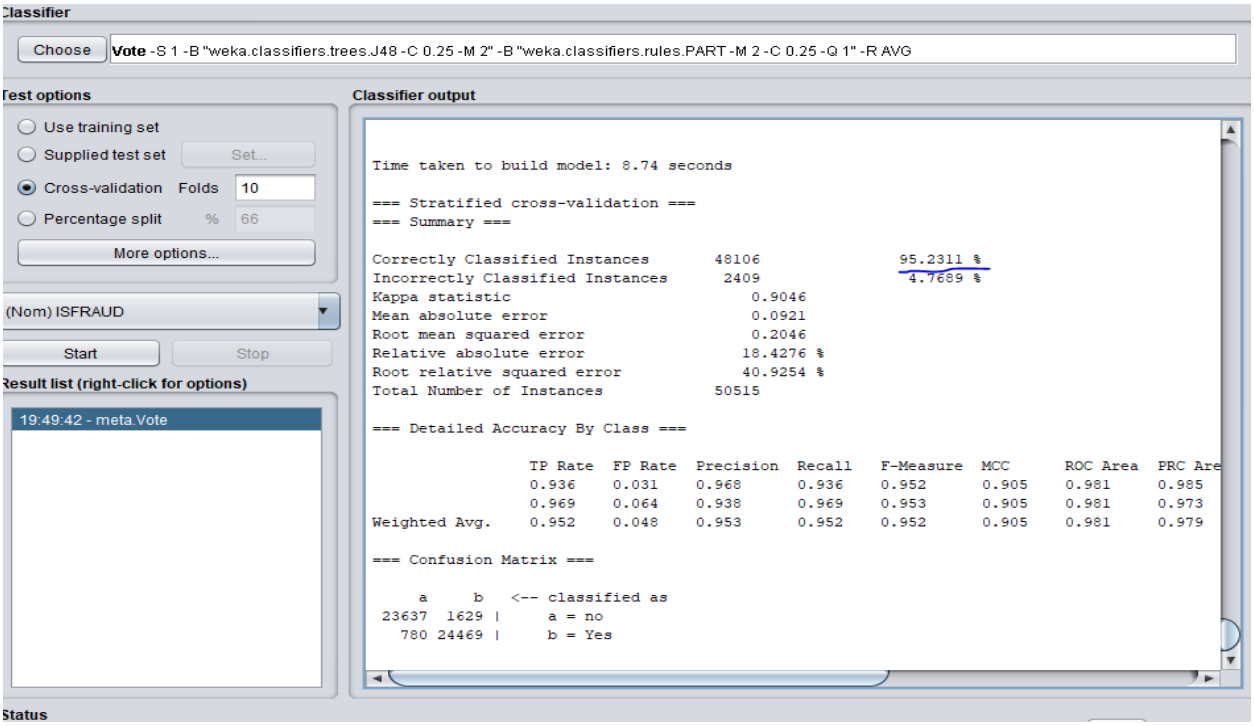
Annex-7: The snapshot running information of stacking J48 decision tree and PART algorithm with 10-fold cross validation technique



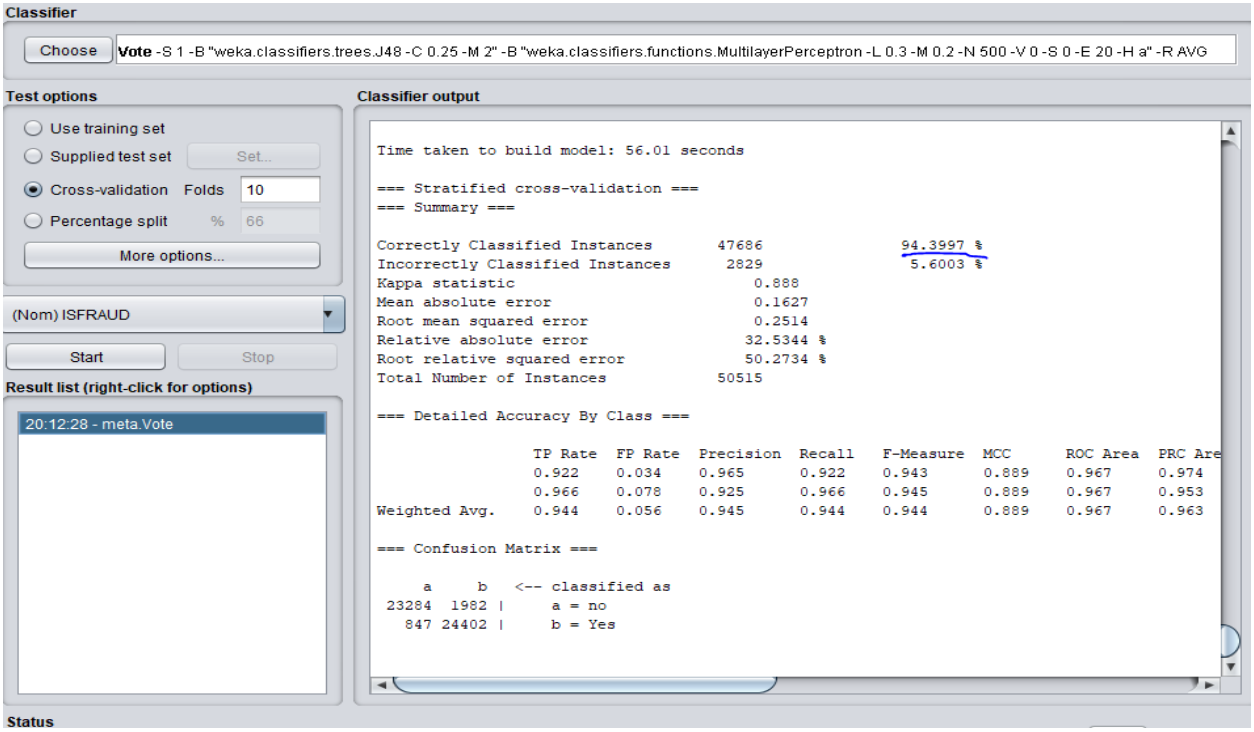
Annex-8: The snapshot running information of stacking J48, PARTand MLP algorithm with 10-fold cross validation technique



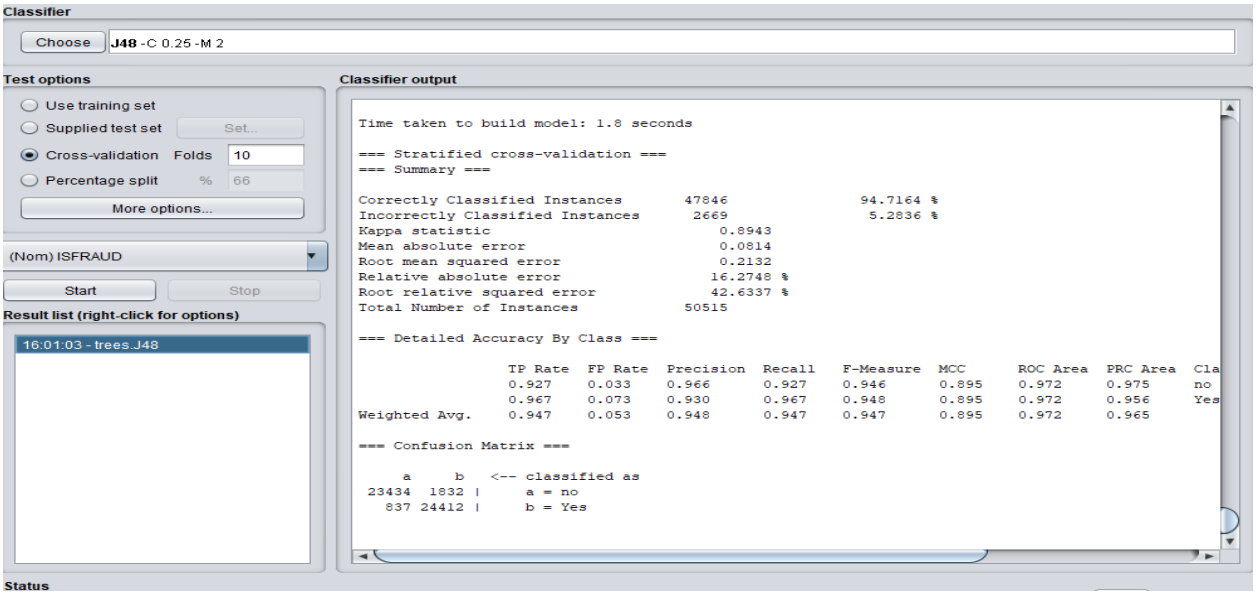
Annex-9: The snapshot running information of voting (J48 and PART) algorithm with 10-fold cross validation technique



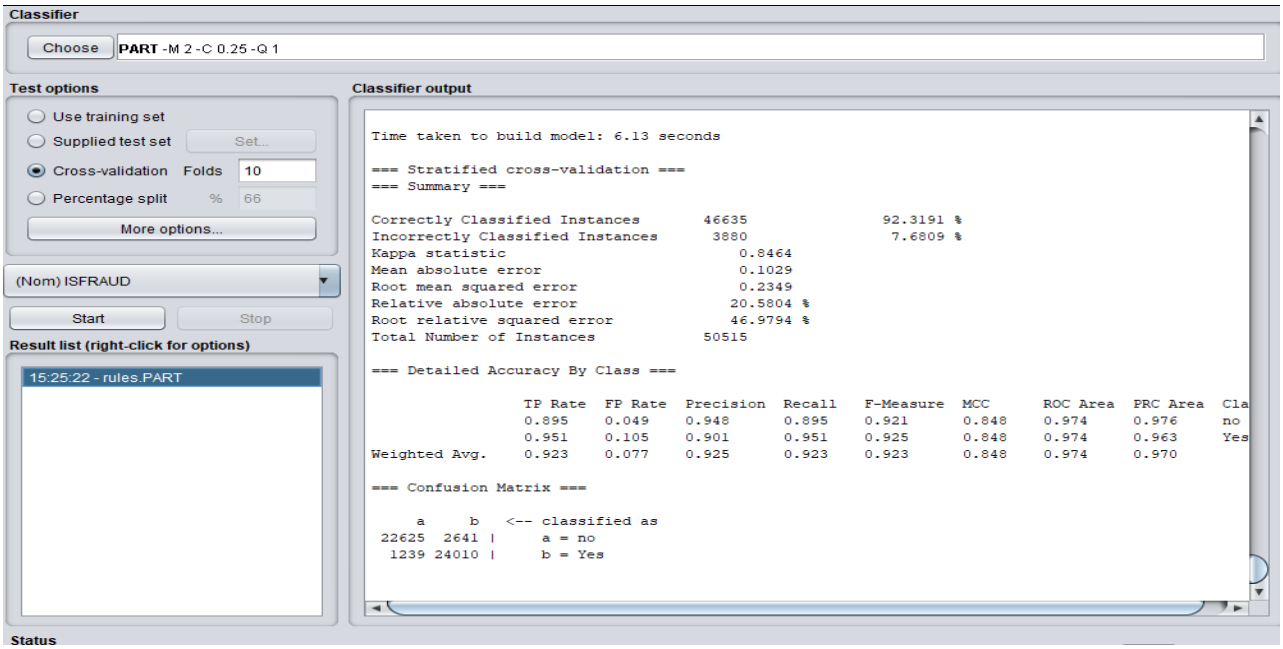
Annex-10: The snapshot running information of voting (J48, PART and MLP) algorithm with 10-fold cross validation technique



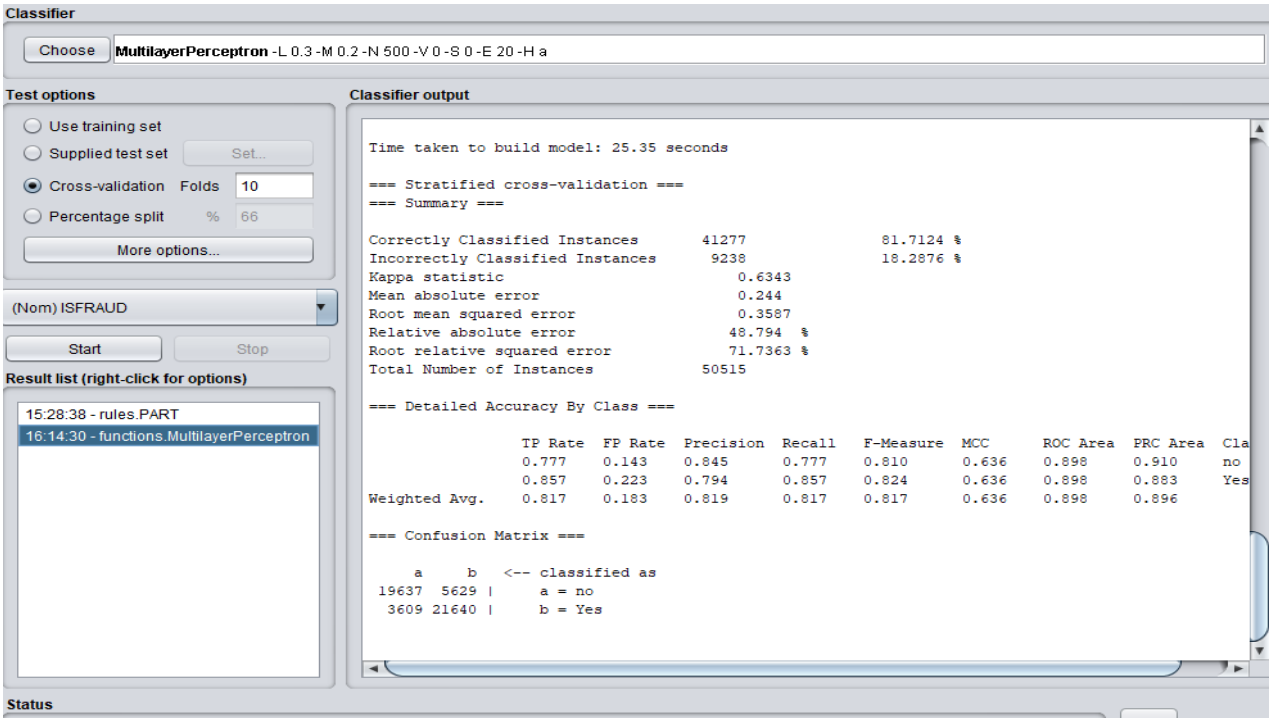
Annex-11: The snapshot running information of J48 with 10-fold validation technique



Annex-12: The snapshot running information of PART with 10-fold cross validation technique



Annex-13: The snapshot running information of MLP with 10-fold cross validation technique



Annex-14: Sample code for implementing rules

```
Private void button2_Click (object sender, EventArgs e)
        {
            comboBox1.Text = "";
            comboBox2.Text = "";
            textBox5.Text = "";
            comboBox5.Text = "";
            comboBox6.Text = "";
            textBox1.Text = "";
            textBox6.Text = "";
            textBox3.Text = "";
            textBox4.Text = "";
            textBox2.Text = "";
        }

        Private void button1_Click (object sender, EventArgs e)
        {
            Decimal callRatio = Convert.ToDecimal (textBox3.Text);
```

```
            Int cellID = Convert.ToInt32 (textBox4.Text);
            Int outCall = Convert.ToInt32 (textBox1.Text);
            Long cella = Convert.ToInt64 (comboBox2.Text);
            Int inCall = Convert.ToInt32 (textBox5.Text);
            Int dis = Convert.ToInt32 (textBox6.Text);
            If (cella == 636011500614211 && cellID >= 7 && outCall >= 26 &&
inCall > 1 && callRatio >= Convert.ToDecimal (0.75)
&&comboBox6.Text=="High"&&dis<=1&&comboBox5.Text=="High")
            {
                textBox2.Text = "YES";
            }
            else if (cella == 63601221432992 && cellID >= 7 && outCall >= 26
&& inCall > 1 && callRatio >= Convert.ToDecimal (0.75) && comboBox6.Text ==
"High"&&dis<=1&&comboBox5.Text=="High")
            {

                textBox2.Text = "YES";
            }
            else {
                textBox2.Text = "NO";
            }

        }
```

Annex-15: Questioner form

St. Mary's university school of graduate studies, department of computer science

Direction:

Dear respondents,

This questionnaire is designed to check for the validity of the prototype which we designed for predicting voice traffic termination fraud status. The truthfulness of your responses will donate much to the validity of this prototype. So you are requested to be honest to give

accurate information. No need to write your name on any part of this questioner. Thank you for your cooperation.

Gender of the respondents 2 female 4 male

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency: o In terms of time (reducing time to accomplish the tasks) o In terms of accuracy					
Effectiveness': o In terms of the output result (quality of work output) o In terms of performance o In terms of error tolerance					
Easy to understand: o In terms of the feature of user interface (easy to learn) o In terms of platform					
Easy to Remember: o In terms of remembering the way to use the prototype					
o Int termsof users productivity					