
**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES**

**SCHOOL OF INFORMATION SCIENCE**

# A Framework for near real-time SIMbox Fraud Detection: The Case of Ethio Telecom

KALEAB ABEBAW

December, 2023

**ADDIS ABABA, ETHIOPIA**

**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES**

**SCHOOL OF INFORMATION SCIENCE**

# A Framework for near real-time SIMbox Fraud Detection: The Case of Ethio Telecom

Name and signature of Member of the Examining Board

| | | |
|---|---|---|
| Million Meshesha (Asst. Prof) | _____ | _____ |
| Advisor | Signature | Date |
| Dereje Teferi (Asst. Prof) | _____ | _____ |
| Examiner | Signature | Date |
| Michael Melese (Asst. Prof) | _____ | _____ |
| Examiner | Signature | Date |

KALEAB ABEBAW

December, 2023

**ADDIS ABABA UNIVERSITY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES**

**SCHOOL OF INFORMATION SCIENCE**

# A Framework for near real-time SIMbox Fraud Detection: The Case of Ethio Telecom

KALEAB ABEBAW

December, 2023

**ADDIS ABABA, ETHIOPIA**

## ACKNOWLEDGMENT

First, my gratitude goes to the almighty GOD who is in control of the existing and the coming world. He has been with me in all depraved and good times and will always be.

I am also thankful to Dr. Million Meshesha (PhD) for his support, inspiration, guidance and patience given in his tight schedule. I consider myself extremely lucky to work with the most prominent scholar in the field and am deeply grateful of his help in the completion of this thesis.

I would like to express my gratitude towards Mr. Tigabu Dagn, Anti-fraud department staffs Solomon Haile, Aragaw, network security specialists Yeshinigus, and Firehiwot Mola for their valuable experiences, and for their welcoming approach.

Finally, I would like to thank my family, my friends, classmates and Ethio Telecom Staffs especially Wr. Eskedar Habtemariam, Melak Zegeye and Antehungn Genetu for their kind help and support during the study.

# Table of Contents

**List of Acronyms**

BSC:____Base Station Controller

BTS:____Base Transceiver Station

CCB:____Customer Care and Billing

CRM:____Customer Relation Management

CDR:____Call Detail Records

EIR:____Equipment Identity Register

FMS:____Fault Management System

GSM:____Global Stations for Mobile communications

GMSC:___Gateway Mobile Switching Centre

HLR:____Home Locator Register

MSC:____Mobile Switching  Center

NSS:____Network Switching Subsystem

OSS:____Operation Support Subsystem

PBX:_____Private Branch Exchange

PRF:____Premium Rate Fraud

SIM:____Subscriber Identity Module

## *List of Tables*



## *List of Figures*



## *List of Appendixes*

# Abstract

Telecom fraud is a major concern for telecom operators as well as for governments especially in Africa and Asia. Bypass fraud is one of the most fertile and costly frauds in today's mobile industry making mobile operators and telecom regulators face a staggering annual revenue losses due to these fixed/VoIP to GSM/CDMA/Fixed line gateway equipment's, which are used to terminate international inbound calls to local calls to local Subscribers by deviating traffic away from the legal interconnect gateways.

Bypass fraud is more rampant in the countries where the cost of terminating international call exceeds the cost of a national call by a considerable margin or the countries where government carriers monopolize international gateways. Fraudsters through the use of different bypass mechanisms, sell capacity to terminate calls cheaply in these countries, on the open market or through direct connections with interconnect operators. Operators sending outbound international traffic are then attracted by these interconnect operators with lower interconnect rates. This leads to lost revenue for terminating network operators.

While several attempts have been made to fight against Bypass Frauds the common approaches have been the use of monitoring calling patterns and profiles through Fraud Management Systems and the use of Test Call Generators. Both approaches have their own set of limitations coping up on frequently changing fraudster's techniques and have short shelf life. In addition, those approaches took couple of months to detect a single fraudulent service number.

The general approach used to perform this research is a design science research methodology.

The proposed system works on processing the near real-time data using Spark Streaming. The objective is to build features to process the near real time data with Spark Streaming to reduce the workload on the node(s), achieve low latency to provide a better execution plan for a scalable and fault-tolerant processing of data. The Proposed framework targeted to improve response time and to give real-time solution to real time problem.

Domain experts made evaluation In order to assure the proposed system has met the requirement needed. In this work, we intended to create a fraud detection framework, which detects frauds based on big data technology, precisely Apache spark and using its Machine learning libraries in order to minimize the latency and to process transactions in a real-time.



Keywords: Simbox Fraud, Big data, Stream Processing, Real-time data Processing, spark,

# CHAPTER ONE

# INTRODUCTION

## 1. Background of the study

Ethio telecom is the oldest public telecommunication operator in Africa. The first long-distance telephone line was established in 1894 between Addis Ababa and Harar [1].

Ethio telecom is government owned and the sole telecom operator in Ethiopia. The operator has passed through different names (brand names) and logos, by different governments that came in power, since the beginning. The name Ethio telecom was coined in 2010 after France telecom took the management of Ethiopian Telecommunication Corporation due to government transformation plan. Fixed telephone (both wired and wireless), Internet (wireless and broadband), mobile (pre-paid and post-paid) including 3G (voice and data), 4G LTE (voice and data) and other value-added services are among the major telecom services provided by the company [1].

The company placed mobile service division in its structure beginning from 1996. Mobile service in Ethiopia has existed since 1999 and at that time, the network coverage was limited to Addis Ababa with a network capacity of not more than 60,000 subscribers. After the introduction of mobile service in Addis Ababa, in April 1999, network expansion was a necessity, not only because of the demand from the subscribers but also due to government policy that is in place in the country [2].

Telecommunications fraud usually focuses on obtaining free telecommunications services and gaining financial benefits [3]. Yalland [4] discussed seven types of frauds. The top three types that cause a major loss are the following:

- International Revenue Share Fraud (IRSF) :when a Fraudster makes an agreement with a local carrier in high cost destination to share profit for increasing traffic, then the fraudster hacks into any organization's public branch exchange (PBX) and gets illegal access to generate calls [5] .
- Premium Rate Service: Fraud Premium rate service is an agreement between some service provider and telecom companies to share revenues generated by traffic to the premium service number [5].The culprits try to exploit this by sending random deceived sms to get a call back with so they can share the call back revenue.
- Bypass fraud: this type of fraud is committed over VoIP, installing Simboxes with multiple low-cost prepaid SIM cards Simbox tools contains SIM slots, and antennas. Simbox fraud is done in a way that

any international call goes through a certain way, before getting from subscriber A to subscriber B. In order to deliver the call quickly, efficiently and inexpensively, large mobile operators use the services of transit operators (intermediaries), whose task is to deliver the call to the operator of another country, which in turn will be sent to subscriber B. Therefore, transit operators buy the route from people or organizations that own Simbox fraud system. Thus, people who own such a system receive money for the route at international rates, and pay only local cost, which is actual within the network. Antrax [6] ,describes that Simbox are devices that acts as a repository of SIM cards, which after converting calls from international cards, make local calls. This is a way in which fraudsters re-route international calls by using Simbox device and local SIM cards [7] .

These days, fraud has been contest for telecommunication industry. The problem with telecommunication fraud is the massive loss of revenue and it can affect the credibility and performance of telecommunication companies. Telecommunication fraud also involves not only the theft of revenues but also services and deliberate abuse of voice and data networks. In such cases, the culprit's intention is to completely avoid or at least reduce the chargesfor using the services. Telecommunication fraud, which is the focus, is particularly appealing to fraudsters, as calling from a mobile terminal is not bound to a physical location, and it is easy to get a subscription. This provides a means for an illegal, high-profit business requiring minimal investment and a relatively low risk of being caught. Fraud is usually initiated by a mobile phone theft, by collecting SIM cards used for Simbox by cloning the mobile phone card, or by acquiring a subscription with false identification. Globally, according to a survey by the communication fraud control Association (CFCA) [8] telecommunications fraud is estimated to be around 55 billion US dollars and in the United States of America alone, 'telecommunication fraud has a share of 2% of network operators' revenue in 2015. However, it is difficultto 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 [9].Another survey by the Communication Fraud Control Association (CFCA) [10], found that the mobile telecom industry lost more than 38 billion dollars in 2015 alone due to telecom fraud. Besides those big losses, telecom fraud causes other indirect losses to mobile operators, suchas, decrease in quality of service, denial of service and network congestion.

The situation is significantly worse especially for mobile operators in Africa; because of fraud, they become liable for large hard currency payments to foreign network operators. Interconnection bypass fraud is so rampantin Africa and Asia due to high tariff on international call. Thus, telecommunication fraud is a significant problem,

which needs to be addressed, detected and prevented in the strongest possible manner. Popular examples of fraud in the telecommunication industry include subscription fraud, identity theft, Simbox or voice over internet protocol (VoIP) fraud, cellular cloning, billing and payment fraud on telecom accounts, prepay and postpaid frauds and PBX fraud [11]. Among the revenue sources of Ethio Telecom, international traffic takes the lion share. As Asfaw [12] indicated40% of Ethiopian Telecommunications Corporation revenue is from international traffic. During July 2016 the Ethio-Telecom annual conference report, the company announced that it has lost over one billionbirr due to fraud. The fraud was commonly bypasses the incoming international call into internet and charges it as local call with VoIP. A number of methods have been tried to prevent this type of fraud so far such as Test Call Generation (TCGs), Fraud Management Systems (FMS), and SIM Card Distribution Control (SDC) [13].

One such method is referred as Fraud Management System (FMS) where it detects Simbox for GSM mobile operators based on Call Detail Records (CDR) analysis. This primarily operates by looking at patterns of usage on SIMs. For example, when too many mobile originated calls are being made with no mobile terminated calls, then presence of Simbox can be detected. However, this pattern analysis is not done in real time. In addition, the bypass fraudsters can fake usages to fool the FMS. Hence, the FMS method either can have many false positives if it is too loose in its implementation logic or can have too few bypass detections if too strict in its implementation logic. In addition, the FMS method cannot detect off-net bypass fraudsters. Furthermore, FMS does not perform immediate fraudster prevention. Moreover, FMS cannot prevent the bypass fraudsters from moving off net, which makes it effectively useless in eliminating termination bypass fraud, as off-net fraud, though less lucrative, can stillkeep the fraud business running. In addition, there are also various techniques available for managing and detecting telephone fraud [14]. The first technique is manual review of data; the problem with this technique is the fact that there are too many data records for a team to filter the fraudulent data. Typically, a telecom company will have in order of 1 millionor more records of telephone calls generated by their customers for a single month within a specific region. As aresult, this is a time consuming and laborious technique for detecting fraud.

The second one is Conventional analysis using a fixed rule based expert system together with statistical analysis. A rule-based system is a set of rules that take into account the normal calling hours, the called destinations as well as the normal duration of the call etc.

The latest one is adaptive flexible techniques using advanced data analysis like artificial neural networks (ANNs) fed with raw data, a neural network can quickly learn to pick up patterns of unusual variations that may suggest

instances of fraud on a particular account [15]. However, the nature of flexibility of fraud, and Fraudsters are smart enough to detect and design their own anti detection mechanism so that the above detection techniques can't handlethe problem. Moreover, they are batch-based: all the data to be used for the modeling must be available upfront for the model construction and many of them cannot work incrementally i.e., incorporate into the model information arrived after it has been built; the model construction is computationally costly. Therefore, the need for near real-time detection mechanisms are inevitable.

The word Real time stands for processing streamed data in motion and analyzing it on time, rather than storing the data as it arrives and analyze at some point of time. In addition, real-time fraud detection is the real-time execution of fraud-detection algorithms in order to detect fraudulent activities. This method highly applied on insurance credit cards and otherfinancial payment systems. It makes use of near real-time data analysis such as forensic analytics and predictive analytics to determine if an ongoing transaction is legitimate or not. Real time in memory computations can helpto solve in real time situations. Near real-time information of any distrustful or potentially fraudulent activity can be instantly identified and taken under control. This also serve to continuously updated to take into account newly identified fraud patterns so as to be able to detect bypass fraud before they make any huge damage.

## 1.1 Motivation of the study

Ethio-telecom is an integrated government owned telecommunications solutions provider operating in Ethiopia, and it is a monopoly company that distributes telecom infrastructure throughout the country with endless effort. During June2016, Information Network Security Agency (INSA) revealed that over the past nine months state-owned telecomcompany, Ethio Telecom, has lost over one billion birr to telecom fraud (www.reporternews.com). According tothe nine-month performance report in 2020, telecom fraud is becoming the major obstacle in expanding access totelecom services in Ethiopia. Furthermore, the report stated that telecom fraud is increasingly claiming huge amount of revenue and costing the nation's telecom sector. The report also noted that most of the fraud is made by installing illegal Simbox to cellular network, bypass the incoming international call to VOIP, and charge it asnormal local call.

Thus far, for all other existing solutions to interconnect bypass fraud took two or more months to give solutions. However, the problem is real time and it needs a near real time solutions. For this reason, real time simbox fraud detection is proposed. The main motive of this study is to prevent this severe revenue lose, improve customer satisfaction, and contribute to the domain by modeling a real time solution for real time problem.

## 1.2 Statement of the problem

International Direct Dialing (IDD) is one of the services based on the Telecommunications Operator clear channel access and Voice over IP (VoIP). In running this business, Operators face Grey Operators who do illegal practices by passing traffic of international incoming call without going through the official international service providers called Fraud Subscriber Identity Module Box (Simbox) [10].

A survey by the Communication Fraud Control Association (CFCA) show that, the mobile telecom industry lost more than 38 billion dollars in 2015 alone due to telecom fraud. [16] So far Ethio Telecom uses different techniques to control such damage. Ethio telecom uses FMS (fraud management system) to prevent simboxing. However, this system has limitations in addressing frequently varyingbehavior of fraudsters, timely solutions and stick to some common features only. More often than not, the existing system ended up blocking normal non-fraudulent especially potential customers Sim card of thecompany. This creates inconvenience among customers and make them not to fully rely on their telecom provider,which leads to customer dissatisfaction. In addition to that, Mobile operators lose important international call termination revenues due to poor quality.

Although bypass fraud has been damaging telecom companies brutally, there has been a shortage of published research to study it and solve it. This is for two reasons: first, fearing that fraudsters might use information published to their advantage. Second, a lack of publicly available data for research purposes. However, fraud detection methods are being developed to stop criminals who also adopt new strategies regularly. There are different approaches for fraud detection. Most researchers follow different data mining approaches to detect fraud. These techniques include machine-learning algorithms, artificial intelligence, and rule-based systems.

However, these methods have limitations and cannot contend with the swiftness of fraudulent activities. The above methods mainly focus on mitigation after severe damage happen on the organization, and are mostly offline and rule based. In addition, conventional methods are exposed to false positives and this further leads to termination of potential user's service number, which results in customer dissatisfaction, and may lead to customer churn as Safaricom telecom operator is the new competitor to Ethiotelecom. Hence, fraudsters easily stun traditional way of Fraud management systems (FMS) and the need for more refined ways is inevitable. Detecting fraud after the event is not nearly as useful as catching it in real time. This study focuseson exploring real-time techniques to effectively detect Simboxing and terminators activity. The fraud detection system has two parts: The collection of data is the first part. data collected from call detail records every time a call is made from an international incoming call, then using predictive analytics tools to analyze and discover patterns while simultaneously training the model when new patterns are analyzed. Once data gets analyzed, the proposed near-

real-time fraud detection system classifies the call as fradulent and legit before making a potential call.

Interconnect bypass fraud is selected for this study indifferently than other telecom frauds because it is rampant in Africa in general, and Ethiopia in particular;

This is Because African countries are known to have higher interconnection tariffs compared to other regions. Bypass fraud is one of the major frauds affecting the dynamic telecom marketin Africa. The impact is huge in terms of the loss in revenues to Telco's and taxes to the government. It is estimatedthat Africa loses up to 150 million US dollars every year to interconnection frauds [17].

Therefore, it is importantto find techniques that can detect this type of fraud efficiently. Real-time fraud detection is an Ideal solution, which could prevent Simbox fraud instantly. This solution allows telecom companies revenues to grow and their costs to shrink. Ethio Telecom will be protected from being affected by its revenue, quality of service, network congestion, and denial of service.

## 1.3 Research Questions

This study attempts to explore and address the following research questions.

1. What are the suitable attributes that are useful to build a good predictor?
2. What suitable framework should be used for designing a near real-time fraud detection system?
3. Evaluate what extent the proposed system works in detecting Simbox fraud?

## 1.4 Objective of the study

### 1.4.1 General objective

The main objective of this thesis is to design a framework that enables to detect and prevent inter connect bypass fraud in near real- time telecommunication.

### 1.4.2 Specific objectives

- To review literature on related research works in order to explore and understand related research works in international interconnect bypass fraud.
- To select the machine learning algorithms which is used in identification of fraudulent Sim cards behavior.
- To analyze and model the domain knowledge and construct structured domain knowledge to gain new knowledge.
- To develop a prototype for simbox fraud detection.
- To evaluate the performance of the proposed detection system.

## 1.5 Scope and limitation of the study

Due to sophisticated technology-aided frauds, telecom operators all over the world are now suffering severe revenue losses, frauds can be in various modes (IRSF, Premium Rate Service, and bypass fraud) regardless of location Telecommunication operators Perpetrators can steal telecom services, misuse them to suffer losses or defraud innocent subscribers, resulting in massive bills or the loss of personal data. However, some types of frauds such as bypass fraud apparently are rampant for telecom operators with high termination fee, and Ethio Telecom is one of the victims of such fraud so this study focuses on this type of fraud only.

The study is intended to evaluate the proposed system in real world scenario. Nevertheless, due to Ethio telecom's business secret, the seriousness and risk of disclosing their existing system to external testing purpose they are not willing to allow testing of the new proposed system. Accordingly, as Ethio Telecom is the only internet service provider in the country, this study cannot make Simulations of the proposed system. However, strong Logical arguments and expert evaluations are undertaken. In addition to that, while conducting of this study, a request for IMEI (international mobile equipment identity) number and IMSI (International Mobile Subscriber Identity) was made. Unfortunately, the data was exclusive due to security reason. This research is intended to be done based on calling number, called number, SMS, GPRS, date and time, call fee, location number and duration of call detail records. However, due to the change of governance and compliance of the information security division they are not willing to give data. Thus, this thesis uses domain knowledge and call-detail record attributes. The study is also limited to pre-paid mobile customers. However, similar frauds could be found on post-paidmobiles as well which fellow researchers could conduct.

Telecom companies are sitting on a gold mine, as they have plenty of data. Nevertheless, what they requireis a proper digging and analysis of both structured and unstructured data to get deeper insights into customerbehavior, their service usage patterns, preferences, and interests in near real-time. For this purpose, this study followbig data analytics frameworks to detect fraudulent service numbers in real time.

## 1.5 Significance of the study

This research enables telecom industries in general and, Ethio telecom in particular to detect frauds in relation to international incoming call in real time bases. The findings of this study can be used for firstly, to provide relevant information to telecom companies to take necessary action and maximize their revenue, by closing the back doorsof revenue loss. Secondly, Ethiopian government retains more hard currency from Ethio telecom incoming international calls, and minimizes the cost for fraud prevention. Thirdly, benefit customer from quality service,

and no subscription suspension due to false alarm**.** Fourthly, other organizations such as banking industries can use this in credit card fraud detection. Finally, this study is an important step forward to retain revenue and mobile end users, and researchers can benefit from this new way of fraud detection method since it shows a new way whichis one step up a head in grappling with financial fraud. Other than focusing on batch data processing.

This research is anticipated to be done based on the following: calling number, SMS, GPRS, date and time, call fee, location number, and duration of call detail records. However, due to the change in governance and compliance of the information security division, they are not willing to provide data. Thus, this thesis uses domain knowledge and call-detail record attributes. The study is also limited to pre-paid mobile customers. However, similar frauds could be found on post-paid mobiles as well, which fellow researchers could conduct

Telecom firms have a plenty of data, which makes them a gold mine. However, in order to gain deeper insights into customer behavior, service usage patterns, preferences, and interests in almost real-time, businesses need to properly mine and analyze both structured and unstructured data. In order to achieve this goal, this study uses big data analytics tools to quickly identify bogus service numbers.

Numerous fraud schemes negatively impact carrier providers not only monetarily but also in terms of large voice bandwidth, service level, and network capacity. The purpose of this research is to support the industry—and ethio-telecom in particular—in its endeavor to minimize revenue loss and safeguard clients from attrition.

## 1.4 Organization of the Thesis

This thesis is organized in a manner assuming the sequential activities of the study. Following this introductory chapter, the principles and concepts of data mining, data mining methodology, fraud detection methods and related works are discussed under literature review in chapter two. Then chapter four covered the discussion on the problem identification and motivation of the research. Chapter Five is about design and development of the proposed framework, the chapter six discuses demonstration and evaluation of the artifact. The last chapter covered conclusions and recommendations from this study.

# CHAPTER TWO

# LITERATURE REVIEW

In this chapter, overview of fraud in telecommunication, types of telecommunication frauds, fraud detection in telecommunication industry and some previous work related to Simbox fraud detection are reviewed, the gaps identified in previous works, and approaches which significantly fill and overcome those gaps are presented.

## 2.1 Fraud in Telecom Industry

Fraud is an unceasing risk to network operators' revenue and it remains difficult to predict exactly how, when, or where new fraud settings will attempt to attack services. Within the telecommunications industry, fraud is an ever-increasing and most prolific threat. Nowadays, it is becoming more pervasive and sophisticated. Telecommunication fraud encompasses a variety of illegal activities on telecom operator network. The dynamicity of fraud types and misuse of technological advancement made the operation of telecom industries more challenging. The industry is not simply watching what fraudsters are doing rather doing its best to keep their reputation to stay as a single point of sale for customers and minimize revenue leakages [1].

There are different types of frauds, which adversely affect the carrier providers, not only financially but alsoin terms of extensive voice bandwidth, service quality and network resources. This research is intended to help the industry, specifically ethio-telecom, in the effort to protect customers churn and lessen revenue loss.

### 2.1.1 Definitions

In legislation, the term fraud is used broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences [18].

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 [19]

The term fraud can also be referred to as the abuse of a profit organization's system without necessarily leading to direct legal consequences [20].

In general, telecommunications fraud can simply described as obtaining telecommunication service with no intention of paying for the service [21]. The major characteristic that makes telecommunications fraud more attractive to fraudsters is that the danger of localization is minimal. This is because all actions are performed from a distance, which makes the process of localization time-consuming and expensive. The simple knowledge of an access code, which can be acquired even with methods of social engineering and advancement of technological progress make the implementation of fraud feasible. Finally, it is possible to say that the product of telecommunications fraud can easily be converted to cash [22].

## 2.2 Fraud Types

Different types of fraud exist in the telecom sector. Different scholars list and categorize the fraud types into different manner. According to Tesfaye [23], there are about six fraud scenarios. These are subscription fraud, PBX fraud, free phone call fraud, premium rate fraud, handset theft and roaming fraud. These fraud types are explained with some detail in next sections.

### 2.2.1 Call and SMS Spamming Fraud

Call and SMS spamming is like email spam [24].Subscribers receive unwanted calls and SMSs about a deal. In the case of SMS spam, the message will have a text to call a specific number or visit a website, which will promote the subscriber to redeem the offer. After that, the subscriber presses or calls the provided link, which will result in premium charges. What distinguishes call and SMS spamming from email spam is that a subscriber might be charged for receiving a spam SMS from websites. In addition, once the subscriber replies to the spam number, he or she will be charged regardless of the subscribed plan. Contrary to email, there is no filtration on call and SMS replies, unlike the case of junk email. Some operators created a mechanism to fight SMS spam. The subscriber can report the spam SMS by forwarding it to specific numbers, but still, there is no built-in mechanism to separate spam SMS on an industry level. [25].

As of March 2005, internationally SMS spamming is illegal to send to users who have not specifically asked for them [26]. However, there is a loophole in the law: solicitors are only prohibited from sending unwanted messages to cell phones from Internet domains. They can still send these messages from a cell phone.

### 2.2.2 Private Branch Exchange (PBX) Fraud

Private branch exchange (PBX) fraud happens when the fraudster takes over the private switching network and uses linked external phone lines to make calls to premium numbers owned by the fraudster. Private branch exchange fraud occurs when the internal network of an organization is not secure enough from outside attacks. Many ways are used to take control of a PBX. Companies might leave default passwords unchanged or they could be corrupted through social engineering, another option might be the attack comes from an internal employee or a vendor [27] .

### 2.2.3 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. Subscription fraud occurs in the phase of signup. The fraudster uses stolen information (SSN, address, or credit card account) to login to services provided by an operator [21]. After signup is complete, the fraudster will commit the fraud and will be billed for general usage. Once the fraudster does not pay the outstanding amount, the amount will be sent to collection agencies, which will rely on the account information that was fake or stolen. In this case, the account was opened using stolen information and now the original owner of the information will be required to pay the outstanding amount. Since the information was mostly fake, no one can be required to pay the outstanding amount, so that amount will be accumulated in the operator’s account as a bad debt [19].

According to GSM Association and the Communications Fraud Control Association [26], 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.

### 2.2.4 Premium Rate Fraud

Premium rate service fraud is the second largest contributor to the $46.3 billion problem of mobile fraud in 2013. It rakes in $4.73 billion globally and $1.35 in North America of losses for subscribers annually. This type of fraud directly attacks subscribers by getting them to make calls to a premium rate telephone number [26].

The most common occurrences of premium rate service fraud directly attack phone companies through the subscription fraud method [26]. It is a basic scheme that takes advantage of phone billing cycles. Fraudsters set up a premium - rate phone number through a carrier and subscribe for one or multiple phone lines through a

different carrier using false information. They then run auto dialers on the subscriber lines that call the premium rate numbers, running up extremely large bills. They do not pay the subscription bills, but receive the profits from the premium -rate line. This goes on until the phone company begins to investigate a bill for non -payment, and then the fraudsters simply close out their services leaving the bills unpaid at the expense of the phone company [7].

### 2.2.5 Domestic revenue share fraud

Domestic revenue share fraud pertains to the abuse of carrier interconnect agreements and is very similar to international revenue share fraud and premium rate service fraud [26]. In all three scenarios, there is an artificialinflation of traffic to a premium rate phone number. The scheme is fairly simple: A fraudster gets hold of a premium rate service number, a phone number, where a portion of the charges goes to the operator and not only the phone carrier, like with regular phone numbers, and inflates the traffic to the service to generate more revenue. There are many different ways this is done and they range from very simple to calculated and organized[11]. One of the simplest methods is by dialing a phone number just long enough to place a missed call on victims' phones but not long enough for them to pick up, to lure them into calling back. This fraud method has become popularly known as "One Ring" or, in its more advanced variant, Wangiri fraud [21]. More sophisticatedmethods of artificial traffic inflation like PBX and voicemail hacking are very common today. Moreover, there are Bluetooth- based attacks that can replace mobile phone numbers with premium rate phone numbers. This not only increases the fraud revenue, but also enables fraudsters to listen into phone conversations. VoIP hacking is also common, where hackers introduce their premium number into victims' communications as a call-through service. With the evolution of technology, it's only normal that fraud becomes more organized [17].

### 2.2.6 Simbox Fraud

Fraudulent Simboxes hijack international voice calls and transfer them over the Internet to a cellular device, which injects them back into the cellular network [28]. By-Pass Fraud occurs when in-bound off-network traffic is disguised as on-network traffic (By-pass) to avoid high costs of terminating traffic. Most By-pass operations are performed on a large scale utilizing advanced Simboxes that can be managed from anywhere.

Content Service Providers that are attacked can experience significant losses in their in-bound interconnect revenues. Service providers should constantly monitor in-bound and on-net traffic in order to detect any indications associated with Simbox fraud, such as suspected calling numbers or suspicious call pattern tendencies.

Simbox is a hardware, which is used to bypass the legal or normal route for international incoming call [1] [29].

Figure 2.1 shows Simbox device, which have SIM slots, antennas, and Ethernet ports that can be used to get the Simbox equipment connected to the internet. Simboxes are used as part of voice over IP gateway installation and the function of Simbox is used to make and terminate international incoming call to local call. The fraudsters can forward international calls through local phone numbers in the respective country to make it appear as the call is a local call [1].

Figure 2.1: Two models of Simbox device (New module 32 SIM card GSM Simbox and 128 SIMs cards call center Simbox device) [1].

Current Simbox equipment have advanced features that help to fraudsters while forwarding the calls like SIM automated rotation, changeable international mobile equipment identity (IMEI), behavior pattern setup and etc. Simbox equipment can be found and purchased online easily throng companies like EBay and Amazon.

A typical Simbox has 32 modems and antennas make calls continuously, it causes congestion problems [14]. Simbox voice fraud mostly occurs where the cost of terminating international call exceeds the cost of mobile-to-mobile call in the country. Fraudulent Simboxes hijack international voice calls and transfer them over the internet to a cellular device, which injects them back into the cellular network. Moreover, fraudsters make a profit by offering low cost international voice calls to the operators and to bypass call routing fees they buy or hijack large amounts of SIM cards and install them into hardware (Simboxes). Then the fraudsters transfer a call via the

internet to a Simbox in the area of call recipient to deliver the call as local. As a result, the operators serving the called party do not receive the corresponding call termination fees.

Simbox fraud also creates many quality issues, like delay, echoes and noise on the line. This quality issues, cause people to make shorter duration calls. The caller telephone number is not visible on the receiver phone, so someone is not sure who is calling. The fraudsters enjoy some portion of the difference between the international termination rate and local tariff. Countries, especially developing countries, in Africa like Ethiopia, Ghana, Congo and Asia suffer this loss due to high incoming traffic of international call for different reasons [19]. The following Figure 2.2 shows how the Simbox works and the routes for both legal and illegal ones.

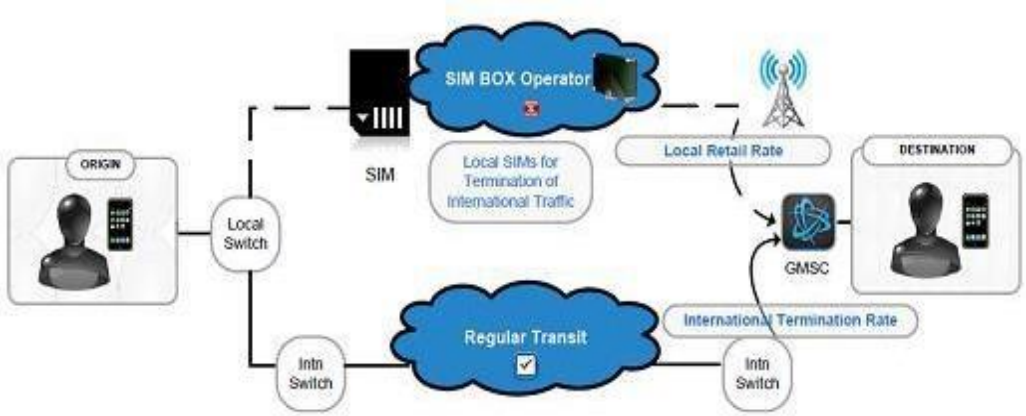



Figure 2.2: Example of Legitimate call and Simbox fraud hijacking of an international call Simbox fraud analysis [24]

In order to identify the Simbox numbers or SIM cards that are used by fraudsters, the numbers have different behaviors as mentioned by different researchers [24] and Ethio telecom experts. The fraudulent numbers did not have outgoing call detail records, most of Simboxes are clustered around two areas without any legitimate account, and legitimate accounts are clustered around another area. It can be observed that most Simboxes have originating calls more than terminating ones while legitimate accounts have comparable number of originatingand terminating calls. That is because Simboxes are used mainly to regenerate the calls received from

the VOIP branch and make them GSM calls again. This feature is very useful to distinguish between Simboxes and legitimate accounts [30].

Simboxes are installed in one location and could be moved from time to time, legitimate users are usually not tight to a specific location. This feature is also very attractive to utilize in order to detect Simbox accounts. It can be noticed that most Simboxes don't send SMS and do not use internet but most of the time legitimate accounts have SMS and GPRS call Details. Mostly fraudulent calls happen on peak of hour and weekend days. Based on the analysis above can conclude that from the total call of incoming and outgoing call, SMS, GPRS, call location, call fee are explored features. Call location number and total number of call per day, SMS and GPRS feature [31] Can give the highest distinguish rate between Simboxes and legitimate accounts.

## 2.3 The Effect of Fraud on Telecom

According to Estévez, Held & Perez [32] Fraud is one of the major revenue leakage sources for telecom industry. Globally, telecommunications lose tens of billions of dollars per year due to fraud. In addition, telecom fraud has a negative impact in terms of quality of service, lost income and wasted capacity. These frauds have either direct or indirect loss of money for a service provider. The direct loss is when resources are consumed and the service provider does not receive payment. If a user succeeds in damaging the reputation or market value of the service provider, then we call this indirect loss [33].

According to Akhter & Ahamad [34], in addition to multi billions of dollars of revenue loss, fraud can also affect the credibility and performance of telecom companies. It involves theft of services and deliberate abuse of voice and data networks. The intent of fraudsters is to avoid or at least reduce the charges for using the service. The negative impact of fraud on the telephone company is described in four ways such as financial, marketing, customer relations and shareholders perceptions. According to Akhter & Ahamad [34], in addition to multi billions of dollars of revenue loss, fraud can also affectthe credibility and performance of telecom companies. It involves theft of services and deliberate abuse of voiceand data networks. The intent of fraudsters is to avoid or at least reduce the charges for using the service. The negative impact of fraud on the telephone company is described in four ways such as financial, marketing, customer relations and shareholders perceptions.

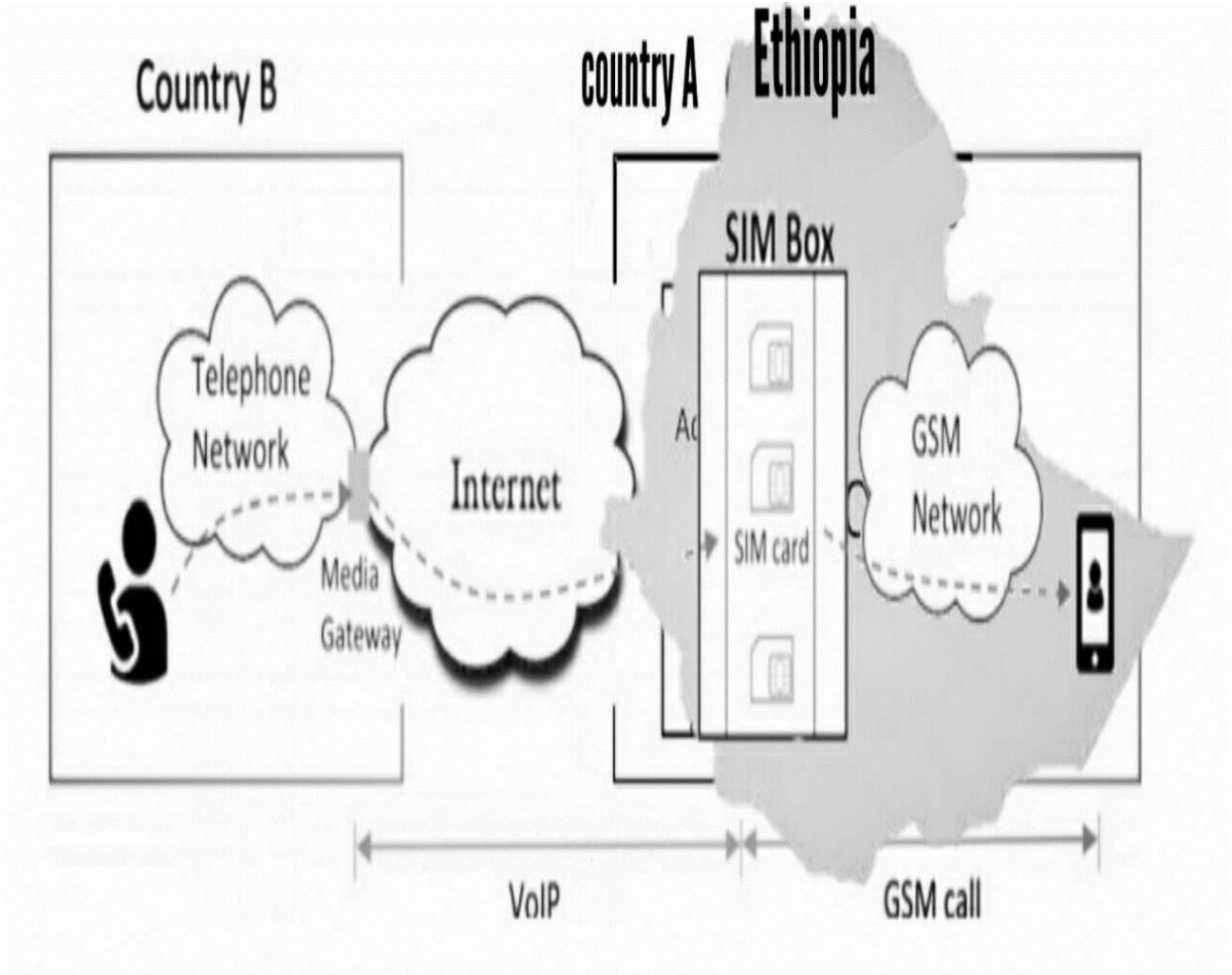



Fig 2.3 interconnect bypass fraud in Ethiopia [21]

## 2.4 Methods Used to Detect Simbox Fraud in Telecommunication Industry

Operators and regulators have devised several fraud detection schemes for generated revenue assurance. There are different techniques and tools to detect the Simbox fraudulent internationally [26]. Test call generation is one of them, it has proven with the identification of Simbox fraud, and it is an effective method for detecting fraudulent numbers. Test calls are initiated to those numbers from various countries by using different interconnect voice routes worldwide. By using test call generation, they identify the paths followed to reach the Simboxes in the home country. Test call generation is a probabilistic method in which the number of fraudulent Simboxes identified increase as more calls using more routes are generated. These methods are relatively new to network operators; an understanding of these methods is the key to managing revenue assurance [5].

The other detection technique is CDR analysis and analytics. CDRs are used to identify fraudulent activities through extensive analysis while performing analytics on fraud indicators by comparing different fields of the CDR like calling number, called number, call type, IMEI, IMSI, time and call duration. In Simbox detection the fraud management system (FMS) uses CDRs user based profiling that distinguish between fraudulent SIMs installed in Simboxes and legitimate users. Call generation providers and FMS tools providers collaborate to pool their alerts in order to more efficiently detect Simbox fraud [4].

## 2.5 Related works

In this section, an effort has been made to review different works which are related with interconnect bypass fraud detection and prevention for telecommunication sectors are presented.

Shawer and Burge [1] who investigated detection of fraud in mobile communication, European Project ASPecT (Advanced security for personal communications technologies), have conducted a research on fraud detection. The ASPecT fraud detection tool is based on investigating sequences of call detail records (CDRs), which contain the details of mobile call attempt for billing purpose. The information produced for billing contains usage behavior for fraud detection. A differential analysis is performed to identify a fraudster through profiling the behavior of a user. ASPecT fraud detection tool utilizes a rule-based system for identifying certain frauds and neural networks to deal with abnormal scenarios.

Bolton, Hand [35] Study based on the statistical and machine learning technology for fraud analysis and detection including their application to detect activities in credit card fraud, telecommunication fraud and computer intrusion. Their study has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and they used different methods to detect them. Such methods have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms.

According to Ilona Murynets and Adam Panagia [5], study the fraudulent traffic from Simboxes operating with a large number of SIM cards. They processes hundreds of millions of anonymized voice call detail records (CDRs). In addition to overloading voice traffic, fraudulent Simboxes are observed to have static physical locations and to generate disproportionately large volume of outgoing calls. Based on these observations, novel classifiers for fraudulent Simbox detection in mobility networks are done.

Farvaresh and Seperi [36] , Applied decision tree (DT), neural network (NN) and support vector machine (SVM) in order to identify customer with residential subscription of wire line telephone service but used it for commercial purposes to get lower tariffs, which is classified as subscription fraud. The employed data mining approach consist of preprocessing, clustering and classification phases. Combination of SVM and K-Means were used in the clustering phase and decision tree (C4.5), Neural Network, SVM as single classifiers were examined

in the classification phase. The results are presented in terms of confusion matrix. DT, NN and SVM as single classifiers were able to correctly classify 88.1%, 84.9% and 88.2% respectively. Therefore, SVM has shown best performance among all the classifiers.

Global fraud survey [9] proves that using bi-directional Neural Network (bi-ANN) in predicting generic mobile phone fraud in real time gave high percentage of accuracy. Bi-ANN is used in prediction the time series of call duration attribute of subscribers in order to identify any unusual behavior. The results show that bi-ANN is capable of predicting these time series, resulting 90% success rate in optimal network configuration. However, call duration is the only parameter used; therefore other relevant parameters are missing to accurately predict customer behavior.

Abdikarim and Roselina Sallehuddin [17] outline the Artificial Neural Network (ANN) and Support Vector Machine (SVM) to detect Global System for Mobile communication (GSM) gateway bypass in Simbox fraud. The suitable features of data obtained from the extraction process of Customer Database Record (CDR) are used for classification in the development of ANN and SVM models. The performance of ANN is compared with SVM to find which model gives the best performance. From the experiments, it is 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.

Bülent [18] Examine the call detail records (CDR's), demographic data and payment data of mobile subscribers in order to develop models of normal and fraudulent behavior via data mining techniques. First, they have done some Exploratory Data Analysis (EDA) on the data set and discovered that some variables like Account length, Package type, Gender, Type, Total Charged Amount showed important tendency for fraudulent use and then they applied k-means cluster method to cluster the customer, based on their call behaviors. Standard variables with ranked attributes and variables obtained from factor analysis due to some correlated variables were used as two different set of variables performed the data mining techniques. Decision trees and Neural Networks for both training and test sets and then discussed the collected results based on performance measures such as accuracy, sensitivity, specificity, precision and RMSE.

After reviewing the above listed related literature we found that CDR data is mostly used and it is a relevant source for analyzing and determining fraudulent calls and also it is known that different

Features (attributes) from the CDR data can be used. We also found different data mining techniques could be used to detect fraudulent activities by analyzing call patterns.

There are different ways of uncovering Bypass fraud being used by operators and regulators. Up to now, a lot of effort has been done on detection of Simbox fraud, however, limitation of the above listed works is that there is no way of updating the existing patterns which makes fraudsters to know the existing patterns and come up with

a committing fraud, detection is done after severe damage on revenue of the company is happened. The fraudsters, once learning the tactics of the detection even the harder one because the fraudsters will make every attempt to make their calling patterns indistinguishable from those of the legitimate callers. This study tries to address this gap, this means the detection algorithm needs to evolve too in order to catch up with the fraudsters, eventually making their cost of hiding challenge the economic gain. Here is where real time analysis needed. Real-time data processing is the execution of data in a short time period, providing near-instantaneous output. A real-time fraud detection model helps telecom companies to detect and prevent bypass fraud in real time, minimize loss, improve customer satisfaction and improve revenue. This approach also allows a telecom provider to deactivate the associated SIMs rapidly, and virtually eliminates the economic incentive to conduct such fraud.

ANN has been used in many business applications for pattern recognition, forecasting, prediction and classification [14]. the neural networks will become an increasing presence in major aspects of telecommunication networks improving efficiency, adapting to changing calling patterns, and providing betterinformation about the use of networks.

# CHAPTER THREE

# RESEARCH METHODOLOGY

The aim of this chapter is to discuss the research methodology used to carry out this research. In this study, the methodology followed to solve the research problem is presented to define methods and techniquesused so as to design the deliverables.

## 3.1 Study design

This study follows design science research method. Design science research is a "lens" or set of synthetic and analytical techniques and perspectives (complementing positivist, interpretive, and critical perspectives) for performing research in Information System. Design science research typically involves the creation of an artifact and/or design theory as a means to improve the current state of practice as well as existing research knowledge [34]. We follow the Information Systems Design Research (ISDR) approach as proposed in [30].

Design science research is a systematic way how research is done scientifically to solve the research problem. It consists of phases of research methods, techniques, procedures, tools, etc., that might be appropriate at each stage of the research. Design science (DS) is a problem-solving paradigm that involves building and evaluating innovative artifacts in a rigorous manner to solve complex, real world problems, make research contributions that extend the boundaries of what is already known, and communicate the results to appropriate audiences [37].

Design science research method is selected because this approach is an improvement research. Aim of this research method is to design an artifact, which can significantly improve an existing method. Design Science is a creative research paradigm that informs multiple audiences: [37] for Researchers It give design principles and mid-range design theories, whereas, For Practitioners it used as an Artifact (product and process) instantiations, it also brings for Managers Work and application system controls and for Government leverages Economic and social welfare.

## Design science research methodology

In order to apply design science research methodology, we use the six-step design science process Model suggested by peffers et al. [37]. An overview of this process model is shown in figure 3.1.

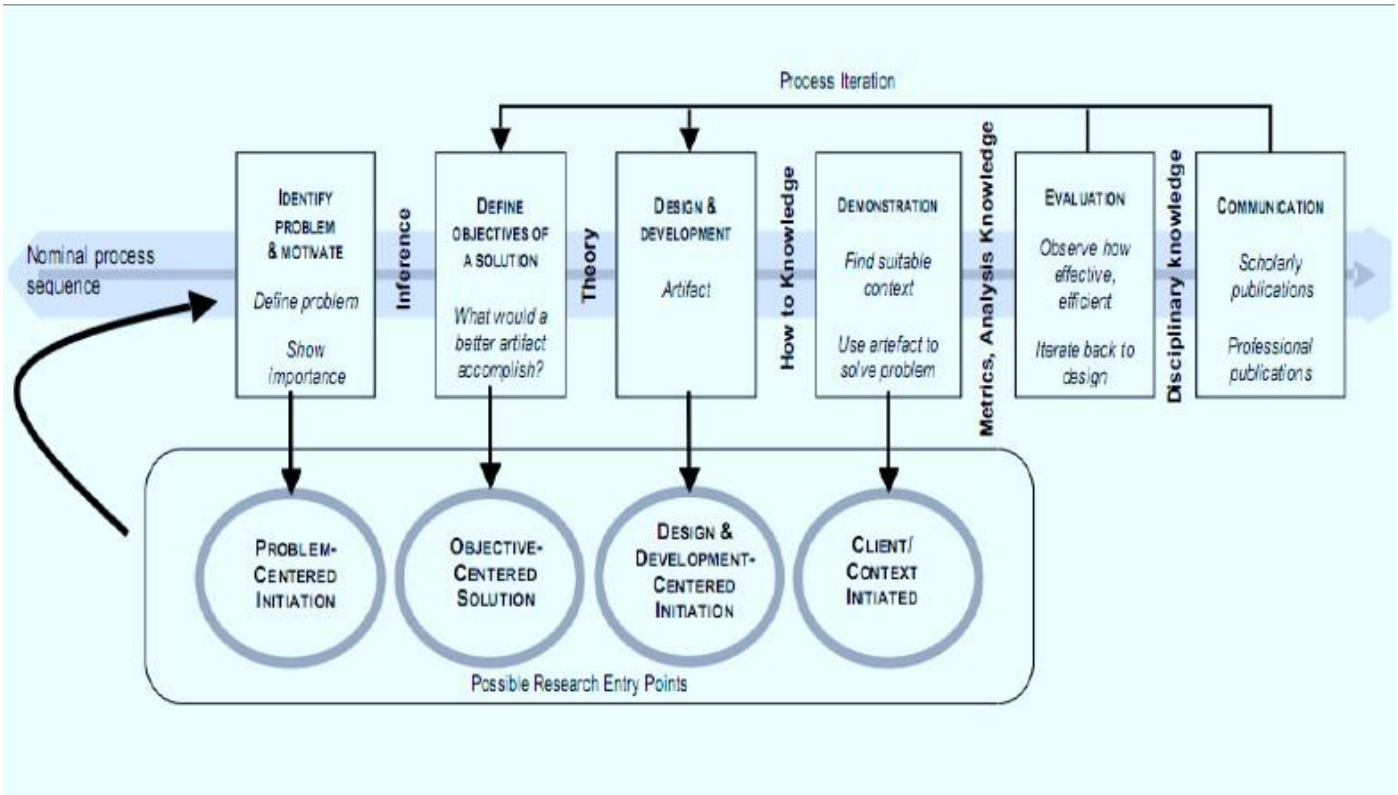



*Figure 3.1* design science method process model) source: [38]

According to Peffers, Tuunanen & Rothenberger [38] DSRM comprises of six subsequent activities namely:

1) Identify problem and motivate

2) Define objectives for a solution

3) Design and development

4) Demonstration

5) Evaluation

6) Communication

| DSRM activities | Activity description | Knowledge base |
|---|---|---|
| Problem identification and motivation | *What is the problem?* Define the research problem and justify the value of a solution. | Understand the problem's relevance and its current solutions and their weaknesses. |
| Define the objectives of a solution | *How should the problem be solved?* In addition to general objectives such as feasibility and performance, what are the specific criteria that a solution for the problen defined in step one should meet? | Kowledge of what is possible and what is feasible. Knowledge of methods, technologies, and theories that can help with defining the objectives. |
| Design and development | *Create an artifact that solves the problem.* Create constructs, models, methods, or instantiations in which a research contribution is embedded. | Application of methods, technologies, and theories to create an artifact that solves the problem. |
| Demonstration | *Demonstrate the use of the artifact.* Prove that the artifact works by solving one or more instances of the problem. | Knowledge of how to use the artifact to solve the problem. |
| Evaluation | *How well does the artifact work?* Observe and measure how well the artifact supports a solution to the problem by comparing the objectives with observed results. | Knowledge of relevant metrics and evaluation techniques. |
| Communication | Communicate the problem, its solution, and the utility, novelty,and effectiveness of the solution to researchers and other relevant audiences. | Knowledge of the disciplinary culture. |

Table 3.1 DSRM activities, activities description and knowledge base. Source: [39].

## 3.3 Problem identification and motivation

Researchers in this section dig out resources required for this activity and knowledge of the state of the problem from literatures and the importance of its solution. To define the problem, we used both primary and secondary sources. Secondary sources such as ethio telecom annual report, literatures (books, journal articles, the Internet) are reviewed to have an understanding of concepts, theories and methods used for solving real-life problems.In addition, primary sources are applied using in - depth and exploratory interviews of domain experts. This isalso done to define the research problem and justify the value of the solution.

### 3.4 Define objectives of a Solution

The objective of design-science research is to develop technology-based solutions to important and relevant business problems.

Our overall objective for the study is to propose a real time Simbox fraud detection artifact to Ethio Telecom in particular and for Telecommunication industry in general. Which significantly outfox the existing fraud management system of a company in the following ways. Customer complain response enhances to in near real

time bases, prevents damage on Ethio Telecom revenue, builds a good reputation for a company and boosts customer journey in the company. The proposed artifact brings solutions to problems not hitherto addressed. Objective of a solution for the current problem is defined based on the identified requirement of the solution using the result of the interview and the specific criteria a solution for the problem should meet.

## 3.5 Design and Development

As Hevner and Chatterjee [40] define "Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence.Conceptually, a design research artifact can be any designed object in which a research contribution isembedded in the design. This activity includes determining the artifact's desired functionality and its architecture and then creating the actual artifact. Resources required moving from objectives to design and development include knowledge of theory that can be brought to bear in a Solution. The researchers use different mode of requirement gathering to solve organizational, people and technology problems. One is using literature review, qualitative methods like interviewing domain experts, on the job observation and from document review.

## 3.6 Demonstration

The essence of Information Systems as design science lies in the scientific evaluation of artifacts" [41].

Rigorous evaluation methods are required to demonstrate the design artifact's utility, quality and efficacy. The evaluation process helps researchers to understand the nuances in their design and contribute to the body of knowledge to facilitate learning by future researchers [40].

## 3.7 Evaluation

According to Hevner [42], a framework, as a model artifact, needs to be evaluated in order to demonstrate its quality, utility and efficiency. This helps to improve the framework in an iterative manner to ensure the quality of the proposed solution so that it can solve real world business problems. Peffers et al [43] suggested the following evaluation methods these are Logical Argument, Expert evaluation, Technical experiment, Subject based experiment, Framework, Case study, Illustrative scenario, and Action research.

| | Logical Argument | Expert Evaluation | Technical Experiment | Subject-Based Experiment | Prototype | Action Research | Case Study | Illustrative Scenario | none | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm | | | 4 | | | | | | | 4 |
| Construct | 1 | | | | 1 | | | | | 2 |
| Framework | 1 | 1 | | | | | 1 | 1 | 1 | 5 |
| Instantiation | | | 3 | | | | | 1 | | 4 |
| Method | 1 | | 2 | 2 | | | 6 | 1 | | 12 |
| Model | | | 1 | | 1 | 1 | | 1 | | 4 |
| Total | 3 | 1 | 10 | 2 | 2 | 1 | 7 | 4 | 1 | |

Table 3.2 Distribution of Evaluation Methods by Artifact Type (IS journals) source [37]

Based on the above table the researcher intends to combine expert evaluation with illustrative scenario as evaluation method. To apply illustrative scenario evaluation method it needs an application of an artifact to a synthetic or real-world situation Anticipated at illustrating appropriateness or utility of the artifact. However, because of the high risk of exposing their data to public use the researcher cannot find any synthetic data from international big data repositories or Ethio Telecom. Thus, researcher uses domain expert evaluation on the clarity, completeness, usefulness, correctness and of the proposed solution. Number of experts participated in this evaluation.

### 3.8 Communication

Design-science research must be presented effectively both to technology-oriented as well as management-oriented Audiences. This work will be presented to audiences through case studies how the proposed system will function to Ethio Telecom community. The result or the proposed system presented to Ethio Telecom information and network security division, coaches, and advisors in contact center, on knowledge base of Information system security division, to top managements of the company, for innovation and support division of the company, and would be available in Addis Ababa university research work repository.

# CHAPTER FOUR

## Problem identification and Setting the objective of a solution

This section focuses on the description and analysis of the data collected to assess the existing fraud management system in Ethio Telecom in order to clearly identify for requirement artifact.

Data collection techniques are listed here why they are used. Data is collected from domain experts, and analysis is done. This is the first step in the process model of DSRM. This phase includes defining the research problem with the proposed solution.

In this study, the researcher states the research problem and justifies the value of a solution through assessment of gaps of prior research related to this study. Here, the study used primary data. The researcher collected data through on the job observation in addition to in-depth interviews of domain experts in order to identify problem and understand the existing simbox fraud detection framework. Interviews are selected as a data collection method to understand in depth the existing FMS because it protects privacy of employee in exposing information for third party.

The responses obtained through in depth interviews are supplemented with on the job observation, ethio telecom annual reports, contact center daily call reason through 994 calls logged, (It's one of the top call reasons customers are calling through 994), and the analysis is done using thematic analysis for interviews taken. Frequency, average and percentage values are used for discussion of the data collected. These methods are used for describing the data collected to investigate the existing simbox fraud detection framework.

In this study, the analysis is made based on the data collected from interview, observation, document analysis. Accordingly, by analyzing the data collected, the study presents a requirement to construct the proposed framework for fraud detection.

## 4.1 Research Motivation

Ethio-telecom is an integrated telecommunications solutions provider operating in Ethiopia, and it is a monopoly company that distributes telecom infrastructure throughout the country with endless effort so far. However, due to political and economic transformations in the country takes place these days, Ethiopia has planned in privatization of giant industries in the country, and telecom industry is one of them. This brings lots of mess for a company whose customer satisfaction rate is below 7.0 [44]. Retaining its customers, reduce those revenue losses due to simbox fraud, improve quality of service, generate more hard currency from international gateways for country which now get pressured to do so, minimize its expenses and to stand competitive the need to reduce the impact of interconnect bypass fraud to do so designing a system which can be a solution to a real time problem is inevitable. Other than the above benefits listed so far this real-time solution will also give telecom companies to make call drops analysis and improve their network service.

## 4.2 Expert interview

Interviews are a systematic way of talking and listing to people [49] There are three different types of interviews structured, semi structured and unstructured.

According to [46], structured interview the researcher asks a predetermined set of questions, using the same wording and order of questions as specified in the interview schedule and structured interviews consist of a series of pre-determined questions that all interviewees answer in the same order.

On the other hand unstructured interviews allow flexibility in objectives, design, sample and the questions that you plan to ask of respondent‘s aspects of the process. This interview is usually the least reliable from research viewpoint, because no questions are prepared prior to the interview and data collection is conducted in an informal manner.

According to [50], semi structured interview offers sufficiently flexibility to approach different respondents differently while still covering the same areas of data collection.

Semi-structured interviews simply a conversation in which you know what you want to find out about and so have a set of questions to ask and a good idea of what topics will be covered but the conversations is free to vary and is likely to change substantially between participants.

Semi-structured interviews contain the components of both, structured and unstructured interviews. In semi-structured interviews, interviewer prepares a set of same questions to be answered by all interviewees. It provides a clear set of instructions for interviewers and can provide reliable, comparative qualitative data.

After carefully understanding each types of interview method semi-structure interview method applied for this

study because it is most suitable to give the researcher flexibility and this flexibility was important to obtain rich information from the interviewers about the subject [51].

Interviews were conducted between June 10 and June 27, 2020 and responses were obtained from all selected informants. The objectives and concepts of the study were briefly explained for interviewees and the researcher interviewed each informant individually. Interview questions are prepared by adapting others work from various literatures and modifying for the specified case

This research has preferred to use primary data collection, method observation and semi-structured interview for different reasons. Firstly, primary data is up to date/fresh and helps to gather appropriate data from the respected bodies. Secondly, primary data collection method for identifying the real requirements of telecom industries, to get in-depth opinion from participants, to gain a better understanding of the participant‘s environment, respondents to have adequate time and to give well thought answer.

For the purpose of this study interviewing the best suitable respondents selected through purposive sampling are used as information acquisition method. Purposive sampling is used in data collection, which means that the best suitable respondents are chosen in order to understand some activity or phenomenon better and discover new viewpoints instead of making statistical generalizations. Nevertheless, Gathering data from the total population is time taking and expensive. Thus, sampling method is preferable. Sampling techniques provide a range of methods that enable to restrict the amount of data needed to collect by considering only data from a subgroup rather than all the population and finally generalizing the result of on whole population.

According to Kothari [45] there are two types of background research primary and secondary research. Primary research involves the study of a subject through firsthand observation and investigation and secondary research involves the collection of information from studies that other researchers have made of a subject.

To collect primary data observation is one way and it is a purposeful, systematic and selective way of watching and listening to an interaction or phenomenon as it takes place [46]. In addition to that, Kumar [46] suggests interview as an essential ingredient in primary data collection techniques and commonly used method of collecting information from people.

## 4.3 Observation

Observation, as the name implies is a way of collecting data through observing. It is the researcher observes phenomena of interest in the environment studied to draw information, which was not obtainable from other methods [47].

Observation as a data collection method can be structured or unstructured. In structured or systematic observation, data collection is conducted using specific variables and according to a pre-defined schedule.

Unstructured observation, on the other hand, is conducted in an open and free manner in a sense that there would be no pre-determined variables or objectives [48].

## 4.4 Sampling method

Sampling means selecting a given number of subjects from a defined population as representative of that population [52] there are two main types of sample probability and non-probability sampling [53].

In probability sampling, all people within the research population have a specifiable chance of being selected whereas non-probability samplings are used if description rather than generalization is the goal. According to [45] in non-probability, sampling the investigator may select a sample, which shall yield results favorable to his point of view. Therefore, for this study the researcher used non-probability sampling methods in order to understand the phenomena and the best suitable respondents to be chosen from selected banks instead of making statistical generalization. In addition, purposive sampling is better for this study to gather data from the experts who are familiar with the topics in detail, helps for the researcher to select information rich cases for study and it is a method whereby a researcher selects sample based on experience or knowledge of the group to be sampled. For conducting purposive sampling, a researcher has something in mind and participants that suit the purpose of the study are included. Purposive sampling enables to use judgment to select case that best enable to answer questions and to meet the objectives. For this study ten experts are selected for interview from 20 information system and network division using purposive sampling. These are two fraud management supervisors, five anti-fraud management analysts, two contact center advisors, and one contact center coach. Of which in turn helped the researcher to acquire in depth insight for the study.

| Divisions of respondents | Job title of respondents | Number of respondents |
|---|---|---|
| Security division | Information security manager | 1 |
| Customer service | Contact center advisors | 2 |
| Information systems | Business and operation expert | 1 |
| Customer service | Contact center coach | 1 |
| Network security | Network security expert | 2 |
| Information security | Anti-fraud analyst | 3 |

Table 4.1 **Sampling for Interview respondents**

## 4.5 Data Analysis Techniques

According to [45] researcher may review two types of literature: the conceptual literature concerning the concepts and theories, and the empirical literature consisting of studies made earlier, which are similar to the one, proposed. In this study, literatures on concepts of interconnect bypass fraud technology and other similar studies were reviewed. Accordingly, interview and review of relevant literature are done for this study to develop the knowledge bases required for developing the proposed near real- time simbox fraud detection architecture.

The researcher collected relevant information after interviewing all selected informants for this study. The collected data was described, organized, analyzed and interpreted for better understanding the current situation. Thematic analysis technique is applied, which means that the result of the interview is grouped into main categories for formulating a problem and developing the working framework from an operational point of view.In this study, the analysis is made based on two components. First, data collected from interview, observation, document analysis and second one is literature review conducted in related works. Accordingly, by analyzing thistwo this study presents a requirement to construct the proposed framework.

### 4.5.1 Findings from interview

The Respondents of the study have a direct relation with the domain knowledge in telecom industry who are directly involved in fraud detection and customer services who are impacted by interconnect bypass fraud. Those are contact center coach, contact center advisor, anti-fraud section, business and operational support analysts, and network security experts.

The data collection method employed was semi-structured interview which helped the researcher to acquire in-depth Insight for the study. The outcome of interview is presented by comparing interview finding against the related literature findings. For this reason, this section contributes a lot for the design and development phase of design science research methodology.

### 4.5.2 Interview Interpretation

The main purpose of this interviews interpretation is to present the views and ideas of the respondents on the current Fraud management system deficiencies. The result of the interview finding helps for identifying each problem of the study. Analysis of the collected data from interview done by the researcher and Interview questions were prepared by adopting from literatures and modified according to this study.

As the interview findings suggested, most of the employees agreed with the advantage of using real-time interconnect bypass fraud detection to solve real-time problems. In addition, related works on the domain problem suggest that even if there are several methods to partially remedy this problem, they remain limited, hence the need for big data technology. Big data is an adequate and efficient technology that has real-time processing capability and is proving to be very useful for fraud detection problems.

According to Gartner [54] result of interview interpretation was consistent with the related literatures. Therefore,the deployment of real-time fraud detection and prevention allows Ethio Telecom to generate more revenues, reduce cost of operation, improve problem solving response time, scalability, advances reliability, enhance qualityof service, and contribute in becoming customer centric telecom provider.

### 4.5.3 Findings from on the job observations

According to on the job observation, the existing fraud management system has been found to be in a poor condition, they are doing the analysis using simple statistical tools. The output from the analysis is used to enhance the quality of the FMS in fight against culprits and to take appropriate actions to deliver the service to its customers as well as sent as a report for higher officials for further decisions. It has its own drawback to analyze huge amount of data using simple statistical tools. In addition, the current fraud management system has many limitations.

Firstly, the right and esteemed customer's service number is frequently suspended because of the false positives of the existing fraud management system (FMS). This happens whenever the existing auto system tries to handle the fraudulent service number based on the rule-based system. Too often, the potential Ethio Telecom customer service or mobile number is also suspended.

It has about 35% success rate. What makes the situation the worst nightmare for customer is the interruption is too frequent for the same service number, this lead customer not to rely on Ethio telecom services, and there is a long response rate for customers complain.

secondly, the current fraud management method has a high false positive rate, which leads in suspension of genuine subscribers of Ethio Telecom, from the call reasons logged at contact center section 994 20% logged call reasons are directly related to sim blocked due to fraud suspension activities. This means that call reasons blocked the chance of Ethio Telecom getting profits from calls that need service of 994.

In addition, security managers use rule-based tools which is good for short period of time only as the drastically changing fraud technology changes to bypass every coming detection mechanisms.

The other thing observed was Ethio Telecom work with vendors in fight against fraud. Thus, to make right decisions on any suspected service number it took longer response time, it takes two to three months to confirm whether a given service number is fraudulent or not.

## 4.6 Gap Analysis

Based on the exploratory and in-depth interviews, on the job observation, document analysis from call reasons of CRM system and from literature reviews, the following gap has been identified.

The existing system has operational delays, difficult to debug, reactive, and it also need dedicated staff to handle all the issues. Also the problem has been identified after the fraud happens and huge damage occurs on the organization telecommunication service.

The gap analysis shows the need for a real-time fraud detection system, which is inevitable to cope up with the dynamically changing fraudster's behavior.

### 4.8.1 Design Requirements

The proposed framework included the following design requirements. In order to enhance the existing Fraud Management System.

Functional requirements determine if the proposed system is to be integrated with the ET system, report latency time, and who has the privilege to view the workflows. And the non-functional requirements include reliability, maintainability, flexibility, scalability, and the near-real-time nature of the proposed system.

Based on the stated functional and nonfunctional requirements the researcher started to identify the type of framework suitable for protecting bypass fraud deployment model.

## 4.9 Objective of the solution

This is the second phase in design science research process models. In this stage, the study aims to identify the requirements for developing the proposed predictive framework from the state of problems. After identifying a problem and pre-evaluating its relevance, a solution has to be developed in the form of a framework.

After data was collected from respondents, it has to be recited and abridged. This involves data preparation, analysis, and finally data interpretation. The analysis technique used in this study is thematic analysis technique in which the result of interview was grouped in to main categories. Finally, appropriate generalization is made and presented accordingly for the qualitative data by way of narrating and interpreting the situations.

Different types of monitoring and predictive models have been proposed so far in Identifying fraud in telecom industry. One of the major problem in those system was they are not proactive. Ethio telecom has its own fraud detection system called fraud management system. However, it is reactive and rule based. There are number of inadequacies in this system. To tackle this problem telecom companies need to deploy the proposed system, which is using real time big data processing engines. Real-time time data processing involves a continual input, process and output of data in minor period nearly in milliseconds to seconds. The proposed system apache spark, which has in memory and iterative computing, speed, dynamic, and it, has machine-learning libraries. The proposed framework gives the following benefits.

Firstly, the proposed system gives solutions in real time, this boost the efficiency of fraud prevention by instantly updating changes nearly as soon as a change of behaviors observed in the fraudster. Secondly, since its decision are made in real time it blocks culprit's service numbers before they make a call, which prevents revenue loss. Thirdly, the proposed system significantly minimizes false positives and network congestions, which prevent customer, churn. Fourthly, it helps telecom companies to reduce their revenue loss, make them to generate more revenue from international gateways. Finally, companies with this service would have a good customer experience and reputation.

# CHAPTER FIVE

## DESIGNING A SIMBOX FRAUD DETECTION FRAMEWORK

The intention of this chapter is to present the results of analysis the researcher identified to interconnect bypass framework for Ethio Telecom. The analysis process based on telecom expert's interview requirements, observation, document analysis and reviewing literatures best predictive model practices. Following the results of the analysis the researcher, proposed higher level architecture that could best meets organizations functional requirements, services, security and deployment model.

### 5.1 Existing fraud management systems: limitations and possible solutions

Since the 1990s, different approaches have been used by telecommunications based on statistical analysis and heuristics methods to help them detect and categorize fraud situations. It was recently that they adopted and explored data mining and knowledge discovery techniques for this task. [55].

Techniques available for managing and detecting telephone fraud include manual review of data, conventional analysis using rule based expert system, artificial intelligence, advanced flexible techniques using data mining (advanced data analysis) , and real time big data stream analysis.

In manual review of data, the problem with this technique is the bulkiness of the data that makes almost impossible for a team to filter the fraudulent calls manually. Especially telecom companies will have millions of call detail records generated by their customers for a single month within a specific region. As a result, this makes it a time consuming and laborious technique for detecting fraud [34].

The second technique is conventional analysis using a fixed rule based expert system together with statistical analysis. A rule-based system is a set of rules that take into account the normal calling hours, the called destinations as well as the normal duration of the call etc. [34]. Rule-based is described as something, which is very difficult to manage because of the proper configuration of rules, requires precise, laborious, and time consuming programming for each imaginable fraud possibility. The dynamicity of new fraud types requires constantly updating the rules to adapt to the existing, emerging and future fraud options. This will introduce a major obstacle for scalability. There will also be a drastic performance downfall of the system when more data is processed by the system [56].

The third technique is adaptive flexible techniques using advanced data analysis like artificial neural networks

(ANNs). Neural network can quickly learn to pick up patterns of unusual variations that may suggest instances of fraud on a particular account by feeding the raw data [34]. Supervised and unsupervised neural networks are the two main forms. Unsupervised learning neural network is one future system that will reduce the processing load for both rule based system and supervised neural based system [57].

Among several approaches proposed for fraud detection, where data needs to be processed on different stacks; one for batch processing and other for stream processing [7], [10]. Traditional systems can perform only one function at one instance; stream either processing of hundreds of MBs with low latency or batch processing of Terabytes with high latency. It is extremely expensive and risky to maintain two separate stacks for two processing modes - such as separate programming model, thus increasing the implementation effort for both stacks. Existing systems uses the pipeline of nodes for data distribution, which causes a problem of latency.

In the entire study of the simbox fraud detection mechanisms in Ethio telecom, it is noticed that, the existing systems are not efficient to deal with huge amount of real time data. Due to the pressing need of processing large volumes of streaming data, the proposed system in this study is aimed at processing massive amount of data in a specified time.

## 5.2 Data Mining

The fast growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. The widening gap between data and information calls for systematic development of data mining tools that will turn “data tombs”, data archives that are seldom visited, into “golden nuggets” of knowledge [58].

Bearing in mind that defining a scientific discipline is controversial and accepting that others might disagree about the details, data mining is defined as: “The analysis of (often large) observational data sets to find unsupervised relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [59]. Similarly, Jiawei & Kamber [58].viewed data mining as a natural revolution of information technology. An evolutionary path has witnessed in database industry in developing functionalities like data collection and database creation, data management, and data analysis and understanding (involving data warehousing and data mining).

Data mining is becoming a mainstream technology used in business intelligence applications supporting industries

such as financial services, retail, healthcare, telecommunications, and higher education, and lines of business such as marketing, manufacturing, customer experiences, customer service, and sales. Nowadays, it is becoming a common practice among business analysts, scientists and researchers to apply data mining on seemingly random data points.

## 5.3 Big data

Big Data has become the buzzword in the world of technology [52], "Big data" is a name given to collection of different types of data sets that are collected from and stored on clusters. Big data is used to describe large amounts of data (structured, unstructured, and semi-structured) that would take too long and are too expensive to load into a relational database for analysis. Therefore, the concept of big data is applied to any information that cannot be processed or analyzed using traditional tools or processes.

Analyzing such large amount of data and processing which needs to be processed instantaneously. It poses a great challenge for database and data analytics research. Conventional or existing database systems are often unable to deal with such large and complex data.

The three main characteristic of big data are volume (data quantity), velocity (data speed), variety (data type). Big data can handle more than 1 million customer transactions per hour.

Big data is made for the telecommunications industry. Thanks to their networks and the increase of smart devices, communications service Providers (CSPs) have access to a wealth of information about their customers' behaviors, preferences and movements. Big data is a tremendously valuable asset for these companies. It puts them in a prime position to win the battle for customers and create new revenue streams provided they could become organized.

Big data is divided into data at rest and data in motion [60].

Data at rest:

This refers to data that has been collected from various sources and is then analyzed after the event occurs. The point where the data is analyzed and the point where action is taken on it occur at two separate times.

Data in motion: The collection process for data in motion is similar to that of data at rest; however, the difference lies in the analytics. In this case, the analytics occur in real-time as the event happens.

Every time call is placed on an Ethio telecom network, descriptive information about the call is saved as a call detail for future record. This data naturally comes as a never-ending stream of events,

For this particular research, different MLlib Spark's machine learning (ML) library used. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as: ML Algorithms: common

learning algorithms such as classification, regression, clustering, and collaborative filtering, factorization: feature extraction, transformation, dimensionality reduction, and selection, Pipelines: tools for constructing, evaluating, and tuning ML Pipelines, Persistence: saving and load algorithms, models, and Pipelines, and Utilities: linear algebra, statistics, data handling.

Due to the pressing need of processing large volumes of streaming data, the proposed system in this study is aimed at processing massive amount of data in a real time based. This study follows an approach of processing live data streams with Spark Streaming.

### 5.3.1 Big Data Technologies

One of the biggest challenges for fraud detection systems is the tremendous growing amount of transactions and streams within a second. Current fraud detection systems need to be more effective and scalable in order to handle such large amount of incoming data. Hence, using Big Data technology is the best solution for this problem. Many Big Data platforms are released to store and process data in recent years. The MapReduce framework was proposed by Dean and Ghemawat [61] in 2004.

Apache Hadoop is a scalable fault-tolerant distributed system for data storage and processing (open source under the Apache license). It is composed of two main subsystems: Hadoop Distributed File System (HDFS) and MapReduce. Hadoop distributes our files through HDFS and then processes them through the MapReduce programming model.

Open- source implementation of MapReduce and DFS can be used for largescale data processing and storing. However, Hadoop has a poor performance on iterative and online computing. Apache Spark2 allows users to persist the data in memory and is the most popular batch-processing platform for iterative computing Storm3 is the most widely used real-time streaming processing system. Storm's applications are submitted as topologies. These topologies usually contain two components, which are called spout and bolt. Spout is the source of streams in topology. It reads tuples from external source and sends them into the topology. Bolt processes the data once a tuple. HBase4 is an open source distributed key-value store developed on top of the distributed storage system HDFS.

### 5.3.2 Apache spark:

Apache spark is a cluster-computing framework built in Scala. It is a fast and general engine for large-scale parallel data processing an in-memory, streaming-enabled, Map-Reduce implementation which automatically distributes the computation among the assigned resources and aggregates the results on a distributed file system. One of the most important characteristic of Spark is that works in memory. This fact implies that is more efficient

doing computations like iterative algorithms, interactive queries and stream processing due the avoidance of the disk read or write bottleneck.

The significant idea of this tool is to organize data in a distributed object, the Resilient Distributed Dataset (RDD) [62]. In case of partition lost, the RDD object contains sufficient information to retrieve the data structure [63]. Spark includes a built-in library for machine learning (package MLlib [40]), as well as one for streaming (package Streaming). A strong point of Spark is its capacity to enable batch and streaming analysis in the same platform. The proposed model relies on Spark Streaming which processes data stream in mini-batches trailing latency of the order of seconds.

The Spark project contains multiple integrated components. Spark SQL, Spark Streaming, MlLib and GraphX. These components have been designed to work together. Thus, they can be combined like libraries in a software project. Where Spark is their core, the one that is responsible for scheduling, distributing, and monitoring applications over cluster [64].


Figure 5.1: Spark stack graph [64].

### 5.3.3 Apache spark for real time solutions

Apache spark streaming is an open source real time processing platform that used computation of big data sets in real-time. It works as stream of data is entered in streaming way to handle this streaming data apache kafka used as data store then useful abstractions will be performed to produce sparks abstraction data RDD (resilient distributed dataset).

Once this Spark is a cluster-computing framework that uses a read only collection of objects called Resilient Distributed Datasets (RDDs) that let users perform in-memory calculation on large clusters. Resilient distributed datasets: A fault-tolerant abstraction for in memory cluster computing.

Processing real-world data requires the ability to analyze data in real-time. Data processing engines like Hadoop come short when results are needed on the fly. Apache Spark's streaming library is increasingly becoming a popular choice as it can stream and analyze a significant amount of data.

Hadoop is the most popular MapReduce framework today, but it has its limitations. The most prominent shortcoming of Hadoop lies in the iterative data processing.

Unlike Hadoop's batch processing, Spark streaming's functional APIs provides horizontal scaling of data across the nodes of cluster in a distributed manner, by which the data can be processed hundred times faster. The statistical measures shows the benchmarks of Hadoop and Spark; i.e. the time (sec) taken to perform iteration(s). This library allows applications to stream data from different sources and with its general code base; it boasts that if it can be stored, then it can be streamed. Some of the most popular streaming sources include Kafka, Flume, Twitter, HDFS, etc. Data can be streamed into the streaming job from one source or multiple sources as they can be unified into a single stream. For the application designed for this research, data is streamed from the Hadoop File System (HDFS).

In addition, apache spark is cluster-computing framework that uses a read only collection of objects called Resilient Distributed Datasets (RDDs) that let users perform in- memory calculation on large clusters [65]

An RDD is a read-only, partitioned collection of objects, which can be stored either in memory or in disk. RDDs are fault-tolerant, parallel data structures, which makes it possible to explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators RDDs can only be created in two ways: by loading an external dataset, or by distributing a collection of objects using the command parallelize.

In many real-world applications, data can often get stale very quickly as it is time sensitive. So, to make the most of such data, it must be analyzed on time. Traditional MapReduce is not a viable solution for such cases as it is mostly suited for offline batch processing where results are not associated with any latency [66].

If the input data is repeatedly produced in discrete sets, multiple passes of the map and reduce tasks would create overhead which can be eliminated by using Spark instead. Apache Spark Streaming lets the program store results in an intermediate data-form within the memory, and when new data arrives as another discrete set, it is batched to perform transformations on them quickly and efficiently [66].

| Hadoop MapReduce | Apache Spark |
|---|---|
| Fast | 100x faster than MapReduce |
| Batch Processing | Real-time Processing |
| Stores Data on Disk | Stores Data in Memory |
| Written in Java | Written in Scala |

Table 5.1 Hadoop MapReduce vs apache spark [62]

### 5.3.4 Data integration

Data collection layer is the very first layer where data is ingested from real-time streaming data when a call is made in Ethio telecom network. Data can be streamed into Apache Spark streaming framework from various sources like Kafka, flume, twitter or HDFS. Two of the most commons data integrators are Apache Flume and Apache Kafka. However, Kafka is a more general-purpose system than flume.

### 5.3.5 Apache Kafka

It is a distributed publish-subscribe messaging system that is used for ingestion of real-time data streams and makes them available to the consumer in a parallel and fault-tolerant manner. Kafka is suitable for building real-time streaming data pipeline that reliably moves data between different processing systems.

The figure below shows how Kafka solves problems in system with many sources of excess pipelines. To overcome these complications, Kafka, will use a system of producer / consumer messaging. The first concept and the core of this system is the topic. Topics are categories where Kafka maintains feeds of messages and producers are processes that publish messages to a Kafka topic. Then, the processes that are subscribed to topics and take the published messages are the consumers. Finally, the Broker is one of the servers that comprise a Kafka cluster.

#### 5.3.5.1 Important parts of Kafka

## Topics

A topic is a category or feed name to which messages are published. Topics are always multilayer subscriber; they can have zero, one, or many consumers that subscribe to the data written to it. For each topic, Kafka maintains a partition log. Zookeeper usually manages Metadata for the partition’s logs and topics. [65]

**Producers**

The second key element explained about apache Kafka are the Producers.

The producer is responsible of publish data to the topics whom they are assigned. Besides, the producer is also responsible for choosing which message to assign to which partition within the topic.

**Consumers**

Messaging traditionally has two models: queuing and publish-subscribe. In a queue, a pool of consumers may read from a server and each message goes to one of them; in publish-subscribe the message is broadcast to all consumers. Kafka offers a single consumer abstraction that generalizes both of these consumer groups.

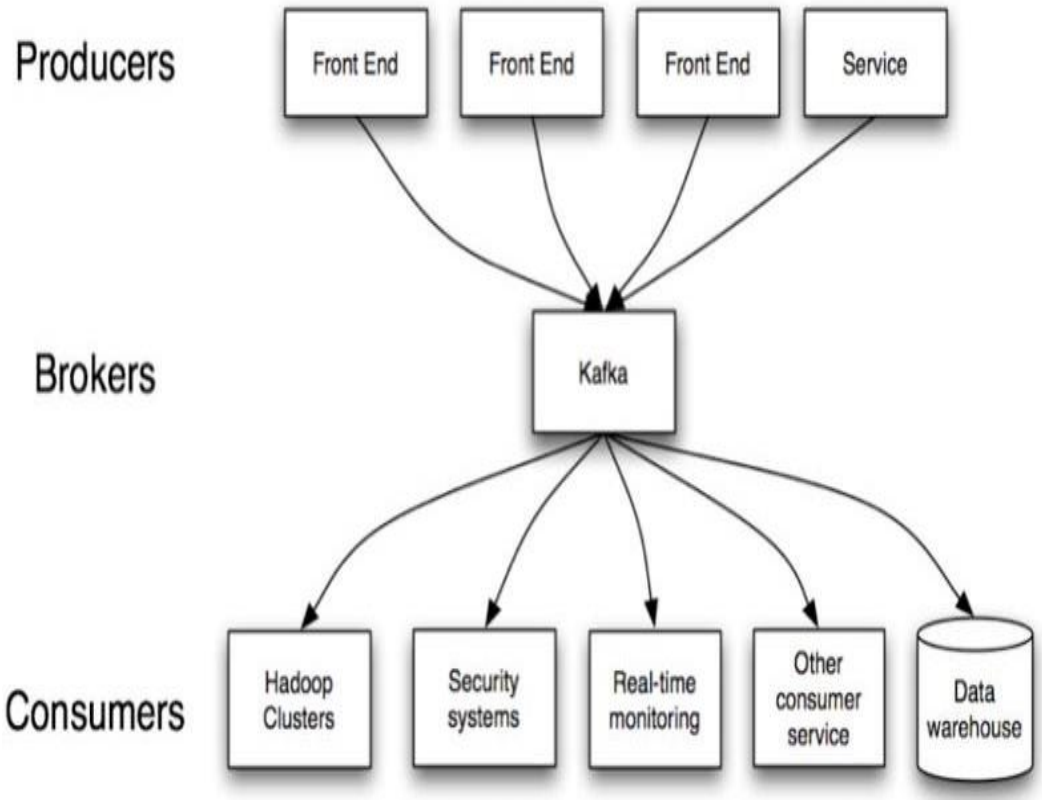



Fig 5.2 Kafka decoupling of the data-pipelines [65]

## 5.4. Machine Learning Library (MLLIB)

The most important tools in Spark is its machine-learning library. MLlib is the Spark's library that contains machine-learning algorithms meant to work in a cluster, classification, and regression. This library is part of the Spark core, therefore can be used from any other of the other libraries. MLlib's design and philosophy are simple there are various algorithms on distributed datasets, representing all data as RDDs. MLlib introduces a few data types (e.g., labeled points and vectors), but at the end of the day, it is simply a set of functions to call on RDDs. [67]

| | |
|---|---|
| basic statistics | summary statistics |
| | correlations |
| | stratified sampling |
| | hypothesis testing |
| | random data generation |
| Classification and regression | linear models (SVMs, logistic regression, linear regression) |
| | naive Bayes |
| | ensembles of trees (Random Forests and Gradient-Boosted Trees) |
| | isotonic regression |
| | decision trees |
| Clustering | k-means |
| | Gaussian mixture |
| | latent Dirichlet allocation (LDA) |
| | streaming k-means |
| | power iteration clustering (PIC) |

Table 5.2 the algorithms that are currently available in spark MlLIB [67]

## 5.5 Spark Streaming

Spark streaming is stream processor, which is a library of Spark that deals with streaming data. It enables scalable, high-throughput, fault-tolerant stream processing of live data [67] [68].


Fig 5.3 spark streaming diagram

The main advantages of using Spark streaming:

- It is scalable.
- Achieve second-scale latencies.
- It is integrated with batch and interactive processing.
- It has a simple programming model.
- It is efficient fault-tolerance. That is possible because it works with batches that are replicated over the cluster.

  Spark Streaming receives data streams and divides this data into batches. After the division, the data is processed by the core of Spark which will generate the final stream of results in batches.


**Fig 5.4 Spark Streaming process of creating of *Discretized stream*** [67]

5.6 Programing language: Spark has currently three main types of programming languages APIs such as Scala, Python and Java. Apache Spark framework is written in scala. Scala has advantages over python and java in the following parameters performance Scala is x10 faster than Python because it works on JVM, on documentations all the documentation of Spark is made in Scala, whereas Python and Java are not used all the examples. Scala also have greater number of apache spark libraries than java and python. However, Python has better machine learning libraries than Scala and java [68].

For this study, the main programming language to be used is python for the following reasons.

**Easy to learn:** Python is comparatively easier to learn because of its syntax and standard libraries. Moreover, it is a dynamically typed language, which means RDDs can hold objects of multiple types.

**A vast set of libraries:** Scala does not have sufficient data science tools and libraries like Python for machine learning and natural language processing. Moreover, Scala lacks good visualization and local data transformations.

Apache spark streaming gives unlimited ability to build cutting-edge applications spark (the python API for spark)

## 5.7 The Proposed Framework

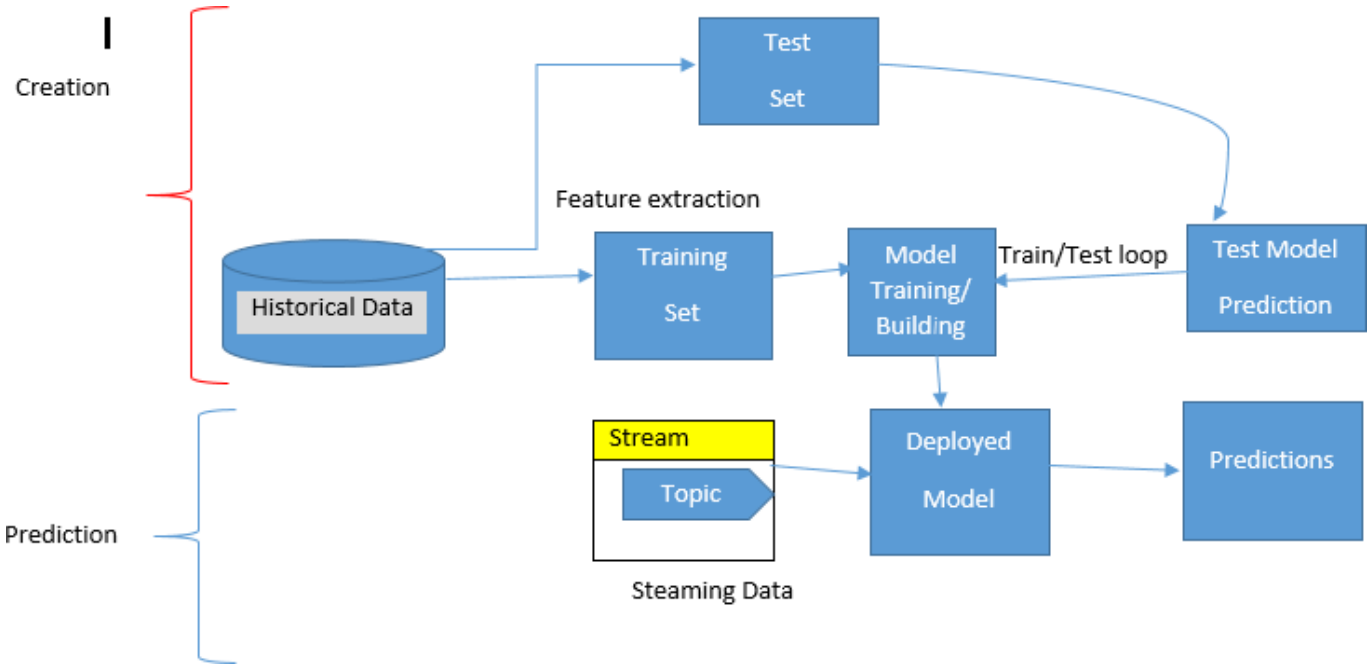



Fig 5.5 near real-time Simbox fraud detection framework

### 5.7.1. Components of the proposed framework:

The proposed framework has two major components creation and prediction. The creation phase performed using Apache sparks Machine learning libraries.

The first phase in creation of the predictive model is building the predictive model. One of the most important

tools in Apache Spark is its machine-learning library. Machine Learning library is the Spark's library that contains machine-learning algorithms meant to work in a cluster. Supervised Learning is a type of machine learning algorithm that uses a training dataset to train a model, based on known outcomes. The model is then applied to new data that it has not seen before to make predictions.

This library is part of the Spark core, therefore can be used from any of the other libraries. MLlib's design and philosophy are simple there are various algorithms on distributed datasets, representing all data as RDDs.

MLlib introduces a few data types (e.g., labeled points and vectors), but at the end of the day, it is simply a set of functions to call on RDDs. Machine learning, on the other hand, learns continuously from new coming data which facilitates processing. a huge collection of past transactions from CDR is uploaded the past data from Cassandra to our Spark Data frame and trained it using Logistic regression which is a supervised learning algorithm, then saving this ML model to our local system to be used later [67].

After the model is created, the predictive phase determines whether the call is legitimate or not before they allow the service number any terminating call.

The data Steam Apache Kafka can be used for data streaming. The simulator created will act as a producer, generating multiple transactions per second and sending it to telecom consumer.

The consumer catches a transaction information at a time and converts this streaming data it gets from the producer to a Spark Data frame. This data frame will be used to prediction whether the transaction is fraudulent or not. Finally, in deployment in Real-time: **a**s the new labelled data gets generated, the model performance can be improved by training the model with new data and by deleting the effects of old data in the model in real time.

# CHAPTER SIX

## DEMONSTRATION AND EVALUATION

This chapter presents the evaluation of the proposed framework for interconnect bypass fraud detection. Demonstration is the process of using the artifact to solve one or more instance of the problem [43]. Furthermore, According to Peffers et al [43]. Demonstration of an artifact can be done using experimentation, simulation, case study, proof, or other appropriate activity. As frameworks are conceptual artifacts to use proof as a demonstration method is appropriate. Therefore, the researchers use proof of concept to show how the proposed framework will solve the problem.

## 6.1 Proof of concept

Peffers et al [43] stated that proof of concept is one of the methods in design science demonstration of an artifact. When cell phone A user dials the number B, presses the *send* or *talk* key, and the mobile phone sends a call setup request message to the mobile phone network via the nearest mobile phone base transceiver station (BTS).

The call setup request message is handled next by the mobile switching center (MSC), which checks the subscriber's record held in the visitor location register to see if the outgoing call is allowed. If so, the MSC then routes the call in the same way that a telephone exchange does in a fixed network.

If the subscriber is on a prepaid tariff, then an additional check is made to see if the subscriber has enough credit to proceed.If not, the call is rejected. If the call is allowed to continue, then it is continually monitored and the appropriate amount is decremented from the subscriber's account. When the credit reaches zero, the network cuts off the call. In the same way, when call was made to Ethio Telecom network as mobile terminating as a call reaches as a stream of call to ET at mobile switching center, then the proposed system checks the legitimacy of the system using Predictive model deployed. Since thepredictive model will incorporate all the attributes such as HLR, GMSC, EIR, AUC, and VLR the fraudulent activity haltedimmediately. Take the investigated culprits number suspended immediately and further actions. In addition, because of the information registered at EIR the predictive system in handing criminals address. Which highly enhance the power in combating againstculprits.

## 6.2 Evaluation of the Proposed Framework

A framework, as a model artifact, needs to be evaluated in order to check its quality, utility and efficiency. This helps to improve the framework in an iterative manner to ensure the quality of the proposed solution so that it cansolve real world business problems [42]. Evaluation is a central and essential activity in conducting rigorous design science research. Artifact can be evaluated using its goal, environment, structure, activity and evolution which are briefly described in the following figure 6.1 [69].

In this study, a total of 6 domain experts from Ethio Telecom specifically from network security, information security and Anti-fraud analysis sections are participated to evaluation survey. Each of the study 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 is presented in table 6.1 below.

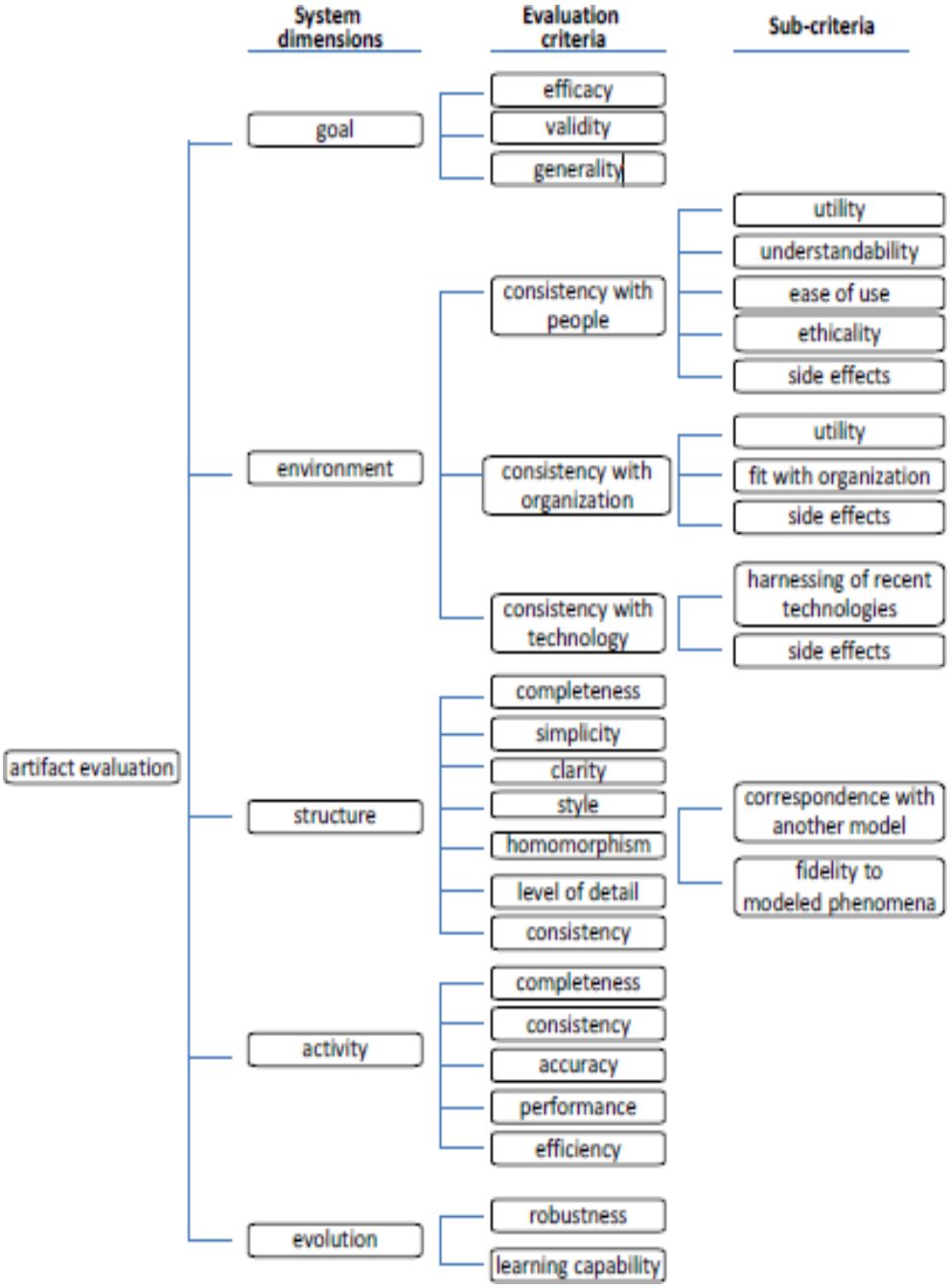



Fig 6.1 Artifact evaluation criteria

| Evaluation Criteria | Strongly Agree (5) | Agree (4) | *Neutral* (3) | Disagree (2) | Strongly disagree (1) |
|---|---|---|---|---|---|
| The proposed framework can improve resource utilization, scalability, flexibility and reduce operational complexity of organization. | 60% | 30% | _ | 10% | _ |
| The organization and presentation of the framework is suitable for Ethio Telecom. | 80% | 10% | - | 10% | - |
| The proposed framework is easy to be applicable | 70% | 10% | - | 20% | - |
| The content of the proposed framework is Scalable | 60% | 30% | - | 10% | - |
| The proposed framework is easy to use | 70% | 10% | 20% | | - |
| The content of the proposed framework is relevant. | 60% | 30% | - | 10% | |
| The content of the proposed framework is clear. | **70%** | **20%** | **10%** | **-** | **-** |
| The content of the proposed framework is complete. | **50%** | **20%** | **-** | **30%** | **-** |
| The objective of the framework is clear | **90%** | **10%** | **-** | **-** | **-** |
| The implementation of the proposed framework fits with the organization Problem. | **80%** | **20%** | **-** | **-** | **-** |
| AVARAGE | **76.0%** | **19%** | **15%** | **10%** | |

Table 6.1 summary of expert evaluation on proposed system

As presented in Table 6.1, the evaluation result by the domain experts revealed the following user acceptance result. The evaluation result indicated that, the applicability of the proposed framework can improve resource utilization, scalability and flexibility and on easy to use of the proposed framework 90% of the experts strongly agreed to it. Whereas, 10% of the experts disagreed with this issue. The number of domain expertise who has issues on easiness of the proposed fraud detection framework is not subtle. They claim that adopting big data technology needs highly skilled man power on the domain, high initial investment, and they added it is not easy to apply all at once: it has its own stages. Moreover, concerning the organization and presentation of the framework, clarity of objective of the framework and appropriateness of the solution to problem majority. 90% of domain experts agreed with 10% of them disagreed. However, concerning the extent to which the prediction framework is easy to learn and easy to be applicable, the respondents reply shows that the proposed system is user friendly and it is more explicit than before. On the other hand, about 20% of the experts have concerns on easiness of applicability of the proposed framework. According to 90% of the IT experts participated on the evaluation of survey of the proposed interconnect bypass fraud, the proposed system confirmed completeness, correctness and clarity, applicability of the proposed framework. Based on the above analysis evaluation result proves how the proposed framework can improve resource utilization, scalability and flexibility of the existing problem. The domain experts also suggested that their need to enhance the completeness of the content and on easiness on the applicability of the proposed framework. In general, the proposed system significantly eradicate the time taken to resolve trouble tickets from months to fraction of seconds.

## 6.3. DISCUSSION OF RESULT

The experts explicated that further attention should be given to the fault tolerance and adaptability of the system. Both test calls and the FMS by nature of their methodology, the fraud has already been committed by the time you detect it. Thus, you are already losing money before the detections occur. This is a major limitation of the previous methods in simbox fraud detection. The proposed system has a quicker response time which you can detect and block fraudsters immediately. Generally, the evaluation results show that the proposed system has believed in retaining its customers, reduce those revenue losses due to simbox fraud, improve quality of service, generate more hard currency from international gateways for country which now get pressured to do so, minimize its expenses and to stand competitive the need to reduce the impact of interconnect bypass fraud to do so designing a system which can be a solution to a real time problem is inevitable.

# CHAPTER SEVEN

# CONCLUSION AND RECOMMENDATIONS

## 7.1 Conclusion

For today's business institutions, fraud prevention and detection is all about speed. With the development of the telecommunication technologies and the big market size of telecom products has given rise to interconnect bypass fraud. Interconnect fraud is a bypass fraud that occurs when international traffic that should be routed through a legitimate international gateway is routed to bypass those gateways using voice over internet protocol.

In this study, an effort has been made to design Simbox fraud detection framework using big data technologies to detect and prevent international interconnect frauds before a potential revenue damage made .

The required knowledge was acquired from intensive empirical reviews, interviews from domain experts, on the job observations, document analysis and other relevant information through related works in domain area.

After the data analysis on the existing system problems, requirements identified.

The objective of this research was to develop an interconnect bypass fraud detection framework which can detect and prevent deceitful calls in real time. The researchers use big data processing engine called Apache spark, which is an effective technique that helps implement a simbox fraud detection system. It has the ability to work with largeamounts of real-time transaction data and helps reduce processing time to milliseconds. The predictive framework has two major components. The creation phase and the predictive phase. Once the predictive modelis created using supervised learning algorithms the model iteratively learns in order to extract new patterns madeby the culprits so as to update.

The proposed system meets the business requirements of Ethio Telecom and have significance to improve business agility, improve quality of service, generate revenues and reduce costs.

This study makes a significant contribution for the domain knowledge by formulating real time solution for near real-time problem. The proposed framework can also be used for credit card fraud detections, call drop analysis, insurance, health care systems, online billing systems and stock markets.

## **7.2** Recommendations

This research is mainly conducted for an academic purpose. This study can be used as a state of the art for real-time Simbox detection solution. This research has proven the applicability of Apache spark for real time analysis. Which automatically discover hidden knowledge that are interesting and accepted by domain experts. Based on the investigations of the study, the following areas are given as a recommendation for the future.

- In this research, we use apache spark technique was used, classification. The deep learning , case based classification can be used in future:
- This study has attempted to apply real time analysis for interconnect bypass fraud detection and prevention, but it could also be applied in other traffic analysis , financial industries such as , credit card , in telecommunication such as network fault isolation, customer profiling and event log analysis for decision making and other purposes.
- Future studies can also be done to select the best suitable fraud detection framework for Ethio Telecom using Apache storm data processing engine.
- This study only focus on interconnect bypass frauds. However, it is also possible to conduct similar researches on the other types of Telecommunication fraud in such as, subscription fraud and SMS spamming fraud.
- Anti-fraud detection staffs should get training on new big data analytic tools.
- Network security division should give trainings to its division employs on the purpose of using such frameworks.
- Managers at network and security division should convince top-level managers in the implementation of the predictive system.

# References


[1] M. H. S. B. A. &. L. L. Yigzaw, *Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian Telecom,* Addis Ababa, 2010.

[2] G. Gebremeskal, *Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Services,* Addis Ababa, 2006.

[3] P. G. a. M. Hyland, *Classification, detection and prosecution on mobile networks,* Belgium, 1999.

[4] M. Yelland, Fraud in mobile networks, 2013.

[5] A. Elmi, S. Ibrahim, and R. Sallehuddin, “Detecting SIM box fraud using neural network,” in IT Convergence and Security 2012, K. J. Kim and K.-Y. Chung, Eds.

[6] "How simbox works," Flames Group SIA, 2017. [Online]. Available: https://simbox.info/all-about-sim-box/. [Accessed 2018].

[7] N. Adu-Boafo, *The big issue: Perspective on SIM Box Fraud in Ghana. Africa Telecom & IT,* 2013.

[8] **Adam, "Global Fraud Loss Survey," December 2015. [Online]. Available: http://www.cfca.org/. [Accessed 2016].**

[9] M. Redwan, *Comparing Data Mining classification algorithms in Detection of SIM Box Fraud,* 2016.

[10] C. F. C. Association, *Global Fraud Loss Survey,* 2015.

[11] G. Jember, *Data Mining Application in Supporting Fraud Detection on Mobile Communication:,* Addis Ababa, 2005.

[12] N. Asfaw, *Challenges Facing International Telecom Business and the Way Forward, Ethiopian Telecommunication Corporation's Perspectives.,* Addis Ababa, 2006.

[13] Y. K. P. a. P. G. B. Igor Ruiz-Agundez, *Fraud Detection for Voice over IP Services on Next-Generation Networks.,* Basaque: University of Deusto Bilbao.

[14] D. M. G. A. M. I. Akhter, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *International Journal of Scientific & Engineering Research,* Vols. vol. 3, no. 3, 2012.

[15] P. Liatsis, "Recent Trends in Multimedia Information Processing, Proceedings of the 9th International Workshop on Systems, Signals and Image Processing, World Scientific," *World Scientific Publishing,* pp. 474-475, 2002.

[16] V. V. N. K. C. Lakshmi, "Survey Paper," *International Journal of Advanced Research in Computer Science and Software Engineering, MCA Department, RGMCET, .*

[17] N. Upadhyay, "https://www.subex.com/why-sim-box-fraud-is-rampant-in-africa/," february 2018. [Online]. Available: https://www.subex.com/why-sim-box-fraud-is-rampant-in-africa/. [Accessed 11 january 2019].

[18] N. Adu-Boafo, *The big issue: Perspective on SIM Box Fraud in Ghana.,* Ghana, 2013.

[19] A. Tariku, *Mining Insurance Data For Fraud Detection: The Case of Africa Insurance Share Company,* 2011.

[20] Phua, C., Lee, V., Smith, K. & Gayler, R. (2005). Comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review (2005) 1–14.

[21] F. Molla, "Analysis and Detection Mechanisms of SIM box Fraud in The Case of Ethio Telecom," 2017. [Online].

[22] Hilas and *Mastorocostas* (*2008*).

[23] Tesfaye, *Predictive Model to Subscription Fraud Detection using Data Mining Techniques,* Addis Ababa, 2013.

[24] V. Airn, "Analysis and detection of SIM box," [Online]. Available: www.ijariit.com.

[25] Hidalgo et al. Content based SMS spam filtering. In Proceedings of the 2006 ACM Symposium on Document Engineering.

[26] Y. Getanh, "PREDICTIVE MODELING FOR FRAUD DETECTION IN TELECOMMUNICATIONS: THE CASE OF ETHIO TELECOM," 2013.

[27] m. Redwan, *Comparing Data Mining classification algorithms in Detection of SImbox fraud,* ST cloud state university, 2016.

[28] G. Gebremesekel, *Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Services.,* Addis Ababa,2006.

[29] J. H. K. &. B. P. Shawe-Taylor, "detection of fraud in mobile telecommunications," *information security technicalreport,* pp. 16-28, 1999.

[30] A. H. a. R. Sallehuddin, *Classification of sim box fraud detection using support vector machine and artificial neural network,* Malaysia, 2014.

[31] G. Jember, *Data Mining Application in Supporting Fraud Detection on Mobile Communication: The Case of Ethio-Mobile.,* Addis Ababa, 2005.

[32] P. A. H. C. M. &. P. C. A. Estévez, "subscription fraud prevention in telecommunication using fuzzy rules and neural networks," *expert systems with applications,* pp. 337-344, 2006.

[33] L. &. K. Jonsson, *combining fraud and intrusion detection meeting, new requirements,* NORDIC WORKSHOP ONSECURE IT SYSTEMS, 2000.

[34] M. I. &. A. M. G. (. Akhter, "Detecting Telecommunication Fraud using Neural Networks through Data Mining.," *International journal of Science & Engineering Research, Volume 3(Issue 3).* 2012.

[35] H. a. Bolton, "stastical fraud detection," *a review stastical science,* pp. 235-249, 2002.

[36] H. a. S. Farvaresh, "A data mining framework for detecting subscription fraud in telecommunication," pp. 182-194, 2011.

[37] Peffers, "Design Science Research Methodology for Information system research," j*ournal of managementinformation system,* pp. 45-77, 2007.

[38] K. T. T. R. Peffers, "A design science research methodology for information systems research," *journal of managementinformation systems,* pp. 24-34, 2008.

[39] Ken Peffers et al, "A design science research methodology for information systems research," *information systems management,* 2006.

[40] A. C. Hevner, "Design Research in Information Systems," New York: 2010.

[41] J. Iivari, "Paradigmatic Analysis of Information Systems As a Design Science," *Journal of Information Systems,* vol. 19, (2007) ".

[42] A. R. Hevner, "Design Science in information systems research," *Management Information Systems,* pp. 75-105, 2004.

[43] M. R. T. T. a. R. V. Ken Peffers, "Design Science Research Evaluation," Oulu Finland, 2007.

[44] *Ethio telecom annual report.* [Performance]. ET, 2019.

[45] C.R.kothari, "Research Methodology Methods and Techniques." in *New Age International*, new Delhi, 1990.

[46] (Kumar, "research writing skills," new York, 2011.

[47] K. &. M. Mohd, "ase Study: A Strategic Research methodology." *American Journal of Applied Science,* 2008.

[48] S. C. &. R. R. P. Lynn, "asic Research Methods for Librarians," *ABC-CLIO,* 2010.

[49] Nyaoro, "A Framework to Guide Companies on Adopting Cloud Computing," *Faculty of Information Technology,* 2012 .

[50] B. &. M. Khairul, "ase Study: A Strategic Research methodology," *American Journal of Applied Sciences,* 2008.

[51] T. Yigremachew, *a framework for virtualized infrastructure as a service for Ethiopian banking industry,* Addis Ababa,2018.

[52] P. &. Meenu, *Research Methodology tools and techniques.* Romania, European Union: Romania. 2015.

[53] D. C. DAWSON, *practical Research Methods a User-friendly Guide to Mastering Research,* United Kingdom, 2002.

[54] Gartner, Inc., what is big data, 2013.

[55] M. B. Bella, M. Olivier, J. Eloff, "A fraud detection model for Next-Generation Networks. Paper presented at the Proceedings of the 8th Southern African Telecommunications Network," 2005

[56] Yufeng Kou, Chang Tien Lu, Sirirat Sirwongwattana, Yo Ping Huang, Survey of fraud detection techniques," in Networking, sensing and control, 2004.

[57] P Burge, J Shawe-Taylor, C Cooke, Y Moreau, B Preneel, C Stoermann , "Fraud detection and management in mobile telecommunications," in European Conference, 1997.

[58] Jiawei Han, Micheline Kamber, Jian Pei i, "Data mining: concepts and techniques," in San Francisco, CA, 2001.

[59] Hand, Mannila, and Smyth, “Principles of data mining," in London The MIT press, London, 2001.

[60] D. F. G. V. Denis Lehmann, "Technology selection for big data and analytical," 2016. [Online].

[61] Lehmann, D.; Fekete, D.; Vossen, G.: Technology Selection for Big Data and Analytical Applications. 2016.

[62] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Cluster computing," in Proceedings of the 9th USENIX, 2012.

[63] M. C. M. J. F. S. M. Zaharia, "Cluster computing with working set," 2010.

[64] A. K. ,. P. W. &. M. Z. Holden Karau, ] Learning Spark Lightning-fast data analysis..

[65] M. e. a. Zaharia, "Resilient distributed datasets: A fault-tolerant abstraction for in memory cluster computing," in *9th USENIX conference on Networked Systems*, 2012.

[66] M. C. M. J. F. S. S. I. S. Matei Zaharia, "spark cluster computing with working sets," in *Proceedings of the 2nd USENIX conference on Hot topics*, 2010.

[67] Nick Pentreath, Machine learning with Spark, 2015.

[68] "Spark.apache.org," 2015. [Online]. Available: https://spark.apache.org/docs/1.3.0/streaming-programming-guide.html.

[69] I. &. J. 2. Nicolas, *Artifact evaluation in information systems design research -holistic view,* paris, 2014.

[70] P. &. M. M. P. Prabhat, *research Methodology tools and techniques.* Romania, European Union: Romania. 2015.

[71] D. M. G. A. Mohammad Iquebal Akhter, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *International Journal of Scientific & Engineering Research,* vol. 3, no. 3, March 2012.

[72] M. J. Gichuki, "Mobile network Fraud Detection Using Artificial Neural Networks," September 2014.

[73] D. G. Edi Sukamto, "Dynamic detection system design of fraud Simbox to improve quality service of international incoming call," *Journal of Applied and Physical Sciences,* February 2016.

[74] S. I. A. M. Z. A. H. E. Roselina Sallehuddin*, "Detecting SIM Box Fraud by Using Support Vector Machine and Artificial Neural Network," *Jurnal of Teknologi,* March 2015.

[75] M. Z. R. P. J. a. A. P. Ilona Murynets∗, "Analysis and Detection of SIMbox Fraud in," *AT&T Security Research Center,* 2014.

[76] M. J. Gichuki, *Mobile network Fraud Detection Using Artificial Neural Networks,* 2014.

[77] D. G. E. Sukamto, "Dynamic detection system design of fraud Simbox to improve quality service of international incoming call," *Journal of Applied and Physical Sciences,* 2016.

[78] A. M. Z. A. H. E. R. S. Subariah Ibrahim, *Detecting SIM Box Fraud by Using Support Vector Machine and Artificial Neural Network,* 2015.

[79] R. P. J. a. A. P. I. M. Michael Zabarankin+, *Analysis and Detection of SIMbox Fraud in Artificial Neural Network,* 2014.

[80] H. E. Abouda Abdulla A. Marah, *Fraud Detection in International Calls Using Fuzzy Logic.*

[81] M. Z. R. P. J. a. A. P. I. Murynets, *Analysis and detection of simbox fraud in mobility networks. In INFOCOM,* 2014.

[82] " Agilis international simbox detection," [Online]. Available: //www.agilisinternational.com/solutions/ customer-analytics/risk-and-fraud-management/,.

[83] H. Berhanu, *Fraud Detection in Telecommunication Networks Using Self-Organizing Map: The Case of Ethiopian Telecommunication Corporatio,* 2006.

[84] M. W. &. B. Berry, *Lecture notes in data mining. Singapore:,* world scientific , 2006.

[85] I. H. &. F. E. Witten, Practical machine learning tools and techniques. Morgan Kaufmann., 2005.

# Appendix A

## Proposed Framework Evaluation Survey

## Addis Ababa University

## College of Natural Science

## Department of Information Science

Dear participant,

In partial fulfillment of the requirements for the Degree of Master of Science in Information Science, I am undertaking a research on near real-time **Simbox fraud detection framework at** Addis Ababa University. Based on the individual discussions, I have amended the proposed framework and accordingly prepared this questionnaire.The objective of the questionnaire is to evaluate the proposed framework with respect to its comprehensiveness,clarity, completeness, correctness, and applicability. There is no compensation for responding nor is there any known risk in participating this survey. In order to ensure that all your information will remain confidential pleasedon't include your name.

Thank you for your dedication to provide your genuine feedback regarding the proposed framework.

**Kaleab Abebaw**
**Kalrose27@gmail.com**
**924447888**

**General**

1. The proposed framework is comprehensive in terms of coverage

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

2. The organization and presentation of the framework is suitable for Ethio Telecom

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

3. The objective of the framework is clear

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

4. The content of the proposed framework is complete

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

Regarding the content of the framework

5. The content of the proposed framework is relevant

□ Strongly Disagree

□ Disagree
□ Neutral

□ Agree

□ Strongly Agree

6. The content of the proposed framework is clear.

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

7. The content of the proposed framework is Scalable.

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

**Regarding utility and applicability of the framework**

8. The proposed framework is easy to use

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

9. The proposed framework is easy to be applicable.

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

10. The applicability of the proposed framework can improve Resource utilization, scalability, flexibility and reduce internal operational complexity

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

11. The implementation of the proposed framework fits with the organization problems

□ Strongly Disagree

□ Disagree

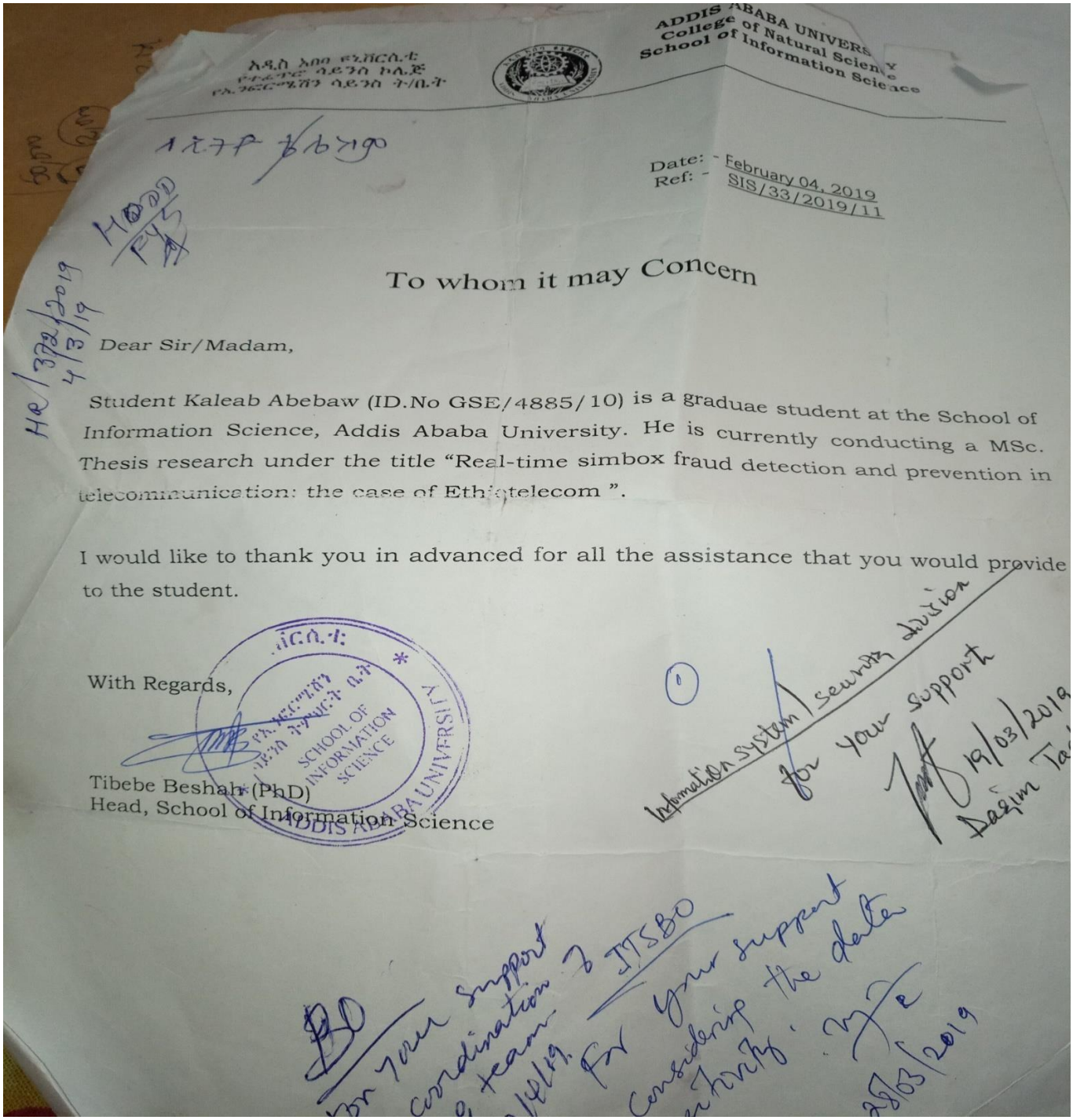
□ Neutral

□ Agree

□ Strongly Agree

# Appendix B

# Letter of Support



አዲስ አበባ ዩኒቨርሲቲ
የተፈጥሮ ሳይንስ ኮሌጅ
የኢንፎርሜሽን ሳይንስ ት/ቤት

ADDIS ABABA UNIVERSITY
College of Natural Science
School of Information Science

Date: - February 04, 2019
Ref: - SIS/33/2019/11

To whom it may Concern

Dear Sir/Madam,

Student Kaleab Abebaw (ID.No GSE/4885/10) is a graduae student at the School of Information Science, Addis Ababa University. He is currently conducting a MSc. Thesis research under the title "Real-time simbox fraud detection and prevention in telecommunication: the case of Ethiotelecom ".

I would like to thank you in advanced for all the assistance that you would provide to the student.

With Regards,

Tibebe Beshah (PhD)
Head, School of Information Science

SCHOOL OF INFORMATION SCIENCE
ADDIS ABABA UNIVERSITY