ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

Interconnect Bypass Fraud Detection Model Using Data Mining Technique

Bekele Haile Nassa

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

August, 2019

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

Bekele Haile Nassa

Advisor: Dr. Dida Midekso

This is to certify that the thesis prepared by *Bekele Haile*, titled: *Interconnect Bypass Fraud Detection Model Using Data mining Technique* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	_____		
Examiner:	_____		
Examiner:	_____		

ABSTRACT

Interconnect bypass fraud is a process by which official interconnect termination routes are being bypassed by using VoIP to route international call traffics into a SIM-Box device where calls are terminated and subsequently regenerated as local calls. According to communication fraud control associate (CFCA, 2017), it is categorized under a type of damage fraud along with subscription fraud. Telecom industry has been expanded dynamically as a result of the development of affordable technologies and an increasing demand of communications. However, the expansion in telecommunication industries in parallel motivated fraudsters to commit telecom fraud using different methods and techniques resulting in the decreasing of the revenue and quality of service in telecommunication providers.

This thesis work focuses on predicting interconnect bypass fraud using different classfication techniques such as multilayer perceptron (MLP), support vector machine (SVM), random decision forest (RF), and J48 algorithms. To achieve our objective, call detail records (CDR) are collected from ethio telcom billing system for two months, from 41 millions active mobile subscribers. We applied cross-industrial standard process for data mining (CRISP-DM) model to the collected raw data; extracted important features from customers CDRs, and derived additional new features so as to characterize the behavior of interconnect bypass fraud. In addition, we preprocessed, aggregated and formatted the datasets convenient for the selected ML algorithms. Each algorithm was trained with five different aggregated datasets such as 4 hours, 8 hours, 12 hours, daily and weekly using two training modes (10-fold cross validation and percent split).

The performance of the models were compared using confusion matrix and we proposed the best models for interconnect bypass fraud prediction. From our experiments, we found that J48 and RF models gave us the highest accuracy as compared to MLP and SVM by giving the classification accuracy of 99.99%, 99.99%, 99.84% and 95.61% respectively on 8 hours aggregated dataset.

Keywords: Telecom Fraud, Bypass Fraud, SIM-Box, Fraud Detection, Data Mining, Knowledge Discovery, CRISP-DM Process Model, Supervised Machine learning, Multilayer Perceptron, Support Vector Machine, J48, Random Forest.

DEDICATION

This research work is most dedicated to all those who pay ultimate sacrifice to make the world is a peace place and to ensure our freedom and security, especially Holy Mother St. Mary and Son.

ACKNOWLEDGMENTS

First and above all, my gratitude goes to the Almighty God for everything.

I am also Thankful to my advisor, Dr. Dida Midekso for every possible effort he has been made to reach this level and provided me valuable support I needed in all situations throughout the research time. Without his support this research would not have been successful.

My deepest gratitude goes to my family for their treatment throughout my life and my studies.

I am also gratitude to ethio telecom management and staffs for their cooperation to give me valuable resources which are relevant for my research and for providing me appropriate professional knowledge.

Finally, I would like to thank all my friends for their encouragement and support.

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

ADT	Alternative Decision Tree
ANN	Artificial Neural Network
ARFF	Attribute Relational File Format
BF	Bypass Fraud
BPNN	Back Propagation Neural Network
BTS	Base Transceiver Station
CBS	Convergent Billing System
CDR	Call Detail Records
CFCA	Communications Fraud Control Association
CFS	Correlation-based Feature Selection
CRISP	Cross Industrial Standard Process
CRM	Customer Relationship Management
CSV	Comma Separated Values
DM	Data Mining
DT	Decision Tree
FCC	Feature to Class Correlation
FDT	Functional Decision Tree
FFC	Feature to Feature Correlation
FFNN	Feed Forward Neural Network
FMS	Fraud Management System
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HPMN	Home Public Network Operator
IBF	Interconnect Bypass Fraud
IMEI	International Mobile Equipment Identity
IRSF	International Revenue Share Fraud
IQR	Interquartile Range
KDD	Knowledge Discovery from Data
ML	Machine Learning
MLP	Multilayer Perceptron
MMS	Multimedia Message Service
MSC	Mobile Switching Center

OSS	Operational Support System
PRSF	Premium Rate Service Fraud
RBF	Radius Base Function
RFA	Random Forest Algorithm
RMSE	Root Mean Square
ROC	Receiver Operating Characteristic
SEMMA	Sample Explore Modify Model Assess
SF	Subscription Fraud
SIM-Box	Subscriber Identity Module Box
SOM	Self-organizing Map
SVM	Support Vector Machine
TCG	Test Call Generator
VoIP	Voice over Interne Protocol
VPMN	Visited Public Mobile Network
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1 : Introduction

1.1 Background

The telecommunication industry has expanded dynamically as a result of the development of affordable technologies and an increasing demand of communications. The development of information and communication sector plays a vital role in the overall development aspect of a population of a given country. However, operators are confronted with a big challenge known as telecom frauds which include technical fraud, subscription fraud, distribution fraud, business fraud, and prepaid fraud [1].

The expansion in telecommunication industries in parallel motivated fraudsters to commit telecom fraud using different methods and techniques [2]. The mobile communication industry has attracted many fraudsters especially for those the subscription methods which are easy to get, and the mobile terminal is not bound to a physical place. Illegal high-profit business can be set up with minimal investment and technical skills as well as very low risk of getting caught [3].

The Global System for Mobile communications (GSM) telecom fraud defines as it is perpetrated where process, control or technical weaknesses are intentionally exploited, resulting in a financial or other loss [1]. In addition, according to the Communications Fraud Control Association (CFCA, 2017), it is that estimated the revenue loss of telecom fraud is $29.2 billion, and that of interconnect bypass fraud is around $ 4.27 billion in 2017 [1]. An interconnect bypass fraud is a technique by which fraudsters re-route international traffic calls by using interconnect bypass device, on which local SIM cards are installed on it and delivered as local calls. Interconnect bypass fraud is the interconnect agreement between mobile network operators are bypassed and the interconnect fee is not paid to the destination mobile operator. The traffic is billed as local (on-net) calls instead of international calls (off-net) where the operator receives only the fee of local calls.

A Subscriber Identity Module Box (SIM-Box) fraud is a bypass fraud that has emerged with the use of Voice over Internet Protocol (VoIP) technologies which is identified as the most damaging fraud type for operators, because it is the main cause for interconnect bypass revenue loss [1, 3]. In addition, this fraud is a reason for mobile networks traffic congestion that leads to quality of service degradation on the base station (cells) where interconnect bypass device is deployed. To commit bypass fraud a fraudster hijacks legal international traffic calls (off-net) and transfers them over the Internet and then injects back into the cellular mobile network using a smart device called SIM-Box. As a result, the calls become local at the destination network, and the terminal cellular operators of the intermediate and destination mobile networks do not receive payments for the international traffic calls routing and termination as per legal rated prices of the international traffic calls.

An interconnect bypass fraud can be operated by many SIM cards that are inserted into SIM-Box device, re-route and terminate international traffic calls to a local home public mobile network, as if it were a call from the same mobile network (on-net) call. The fraudsters can use a virtual GSM gateway where SIM cards are stored and administrated in a central location. The central administration tool communicates with many GSM gateway via TCP/IP protocol and the SIM can be redirected or swap any of the GSM terminals directly attached to the server which is dynamically configured by the fraudster using special software.

Therefore, the common platform for committing of interconnect bypass fraud is using SIM-Box device, connected with VoIP GSM gateways. This bypass device receives international traffic calls through VoIP, then inject back to local cellular mobile network via a collection of SIM cards inside of it.

A telecom fraud detection method has been continuously improving to tackle the fraudsters [4]. The three main approaches used to tackle interconnect bypass fraud are Test Call Generator (TCG), which is a method for testing different international traffic routes of the mobile networks; rule-based Fraud Management System (FMS) which uses exact expression matching predefined by domain experts [5], and controlling distribution of SIM cards through retailers. However, fraudsters are dynamic and work hard to bypass those detection and prevention firewalls. Whenever fraudsters are aware or feel being detected, they immediately change their techniques of committing fraud. For instance, fraudsters

avoid TCG by analyzing the incoming voice call traffic and based on defined patterns, they could determine whether the calls are legitimate or test generated calls. In addition, they use smart SIM-Box devices that can imitate the activities of normal subscribers' behavior to avoid being detected by rule-based FMS. Besides, evolving of new technology and complex integration of communication mobile network elements open unseen vulnerability for the operators and brings new opportunities for the fraudsters.

1.2 Motivation

The researcher is an employee of ethio telecom who has been observing and attending different telecom fraud survey reports. Through the reports, he came to know the severity of the telecom fraud.

The impacts of fraud are not only the revenue decline, but also damages the reputation of the operator and low-quality of service. In addition, interconnect bypass fraud may cause national security threat, and loss of customer confidence. These are the major motivations to do this study and tackle interconnect bypass fraud and protect operators against the fraudsters.

1.3 Statement of the Problem

An interconnect is one of the major services offered by telecommunication mobile operators using VoIP. However, while running this service bad transient operators bypass the traffic of international incoming call without going through the legal international gateway.

An interconnect bypass fraud is a significant source of revenue loss for the telecom operators [1]. Even though telecommunication operators apply different methods and techniques such as TCG, rule-based FMS and control SIM-cards distribution to tackle the problem, fraudsters easily defraud both TCG and rule-based FMS [6]. Besides, to generate test calls using TCG to the entire international and national mobile network routes is costly to the operators. Also, fraudsters designed smart SIM-Box device which can imitate the normal customer's behavior to avoid being detected by rule-based FMS. One of the major drawbacks observed in rule-based FMS is scalability. The more data the system processes, the more the system performance downfalls [6]. In addition, the ability to adaptation and learning to the current huge data sets in the telecom industry is limited and restricted.

The major limitation observed in the related work is that researchers in the area conducted their experiment used unlabeled dataset, limited Call Detail Records (CDR) and features which may not reflect the actual behavior of the fraudsters. Similarly, in telecom there are more legitimate subscribers than fraudsters, so that the model trained with limited and unlabeled dataset might not catch and/or reflect the real behavior of the interconnect fraud in production environment.

Therefore, to overcome this limitation, there is a need to explore data mining techniques that learn the dynamic change of subscriber's behavior from the big data to identify fraudulent subscribers.

1.4 Objectives

General Objective

The general objective of this thesis is to develop a model for telecom interconnect bypass fraud detection using data mining technique.

Specific Objectives

To meet the general objective, the following specific objectives will be carried out:

- Understand the domain area through discussions and interviews with domain experts.
- Explore telecom fraudulent cases by reviewing literature and related work.
- Review different literature on data mining techniques and their applications in the field of telecom industry.
- Identify appropriate data mining algorithms.
- Build a model, and
- Evaluate the performance of the model.

1.5 Methods

The goal of this research is to develop a model for detecting interconnect bypass fraud using data mining techniques. To achieve this goal, the following methods will be used.

Literature Review

A review of literature will be done to acquire a deeper understanding of the problem domain. In addition, discussion with domain experts will be conducted and some of the major discussion points are presented in Appendix 4.

Data collection and Analysis

Subscribers call detail records (CDR) data will be collected and then analyzed.

Data preparation

Potentially important attributes will be extracted from the analyzed raw data and new features will be derived and preprocessed to make ready for conducting experiment. WEKA data mining tool will also be used.

Model building

A model will be developed using different supervised Machine Learning (ML) algorithms and training methods using different aggregated datasets.

Model Evaluation

The developed models will be tested and evaluated through confusion matrix.

1.6 Scope and Limitations

The focus of this study is to develop a predictive model to detect telecom interconnect bypass fraud so that the study is limited to a data mining technique based on customers toll tickets for prepaid subscribers and the data collected for a period of two months only because a CDR contains millions of records within a few days in a specific region.

1.7 Application of Results

The result of this study will enable telecom operators to detect interconnect bypass fraud and to show the limitation of existing rule-based FMS and TCG. In addition, it gives the operators a better qualitative understanding of subscriber fraudulent behavior and it maximizes correct prediction and minimizes incorrect prediction of fraudulent subscriber at a satisfactory level.

1.8 Organization of the Rest of the Thesis

The rest of the thesis is organized as follows:

Chapter two presents literature review on telecommunication frauds, data mining and machine learning. Chapter three states a review of related work specifically on telecom interconnect bypass fraud. Chapter four deals with the proposed solution for the problem which includes knowledge discovery, process model, data collection, data preprocessing and data modeling. Chapter five focuses on experimentation and evaluation, which includes selecting modeling methods, building models, testing and evaluating the models, and discussing the outcomes of the experiment. Chapter six is dedicated to conclusion and future works to be carried out.

Chapter 2 : Literature Review

2.1 Telecommunication Fraud

A telecommunication fraud can be simply described as any activity by which telecommunications service is obtained with no intention of paying [3]. Telecom operators are the major sectors that are attacked by fraudsters [7]. Telcom frauds are a combination of variety of illegal activities like illegitimate access, identity theft, and revenue share. It occurs because there are three fundamental things that happen together, namely the pressure of cheating or negative financial growth, the opportunity to commit fraud, and attitude or rationalization to justify the act of cheating [8].

Telecommunication operators use different devices and technologies to facilitate their services. However, fraudsters deceive, or deliberately misuse services offered via a telecom system in order to avoid payment for services used or earning money. The crime of telecom fraud also increases with advancing communications and Internet technologies and causing huge losses of revenue every year [1].

2.2 Telecommunication Fraud Types

There are different types of telecommunications fraud and these can occur at various levels. Different authors categorized frauds in different ways. For instance, according to Yufeng *et al.* [6] and Iquebal and Gulan [7] subscription fraud and superimposed fraud are the most common fraud types, and according to Rechard A *et al.* [9], they are classified into seven groups as superimposed, subscription, technical, internal fraud, social engineering, fraud based on loopholes in technology, and fraud based on new technology. Similarly, Ibrahim and Hussamedin [10] list the top three types of telecommunication fraud that cause a significant loss such as International Revenue Share Fraud (IRSF), Premium Rate Service Fraud (PRSF) and Bypass Fraud (BF). Whereas, Phil and Mark [11] classified the types as contractual, hacking, technical and procedural frauds.

The most damaging frauds for telecom operators are stated as follows.

A. Subscription Fraud (SF)

It is an enabler for most damaging fraud type and the most common fraud types along with the SIM cloning [12]. Fraudsters obtain an account with no intention to pay the bill. In such cases, abnormal usage occurs throughout the active period of the account. The theft account is usually used for call selling or intensive self-usage [13].

B. International Revenue Share Fraud (IRSF)

IRSF is the largest contributor to the overall fraud losses according to CFCA [1]. It occurs when an operator makes an agreement with other operators which will generate calls to premium rate number to generate revenue for increasing traffic. IRSF often involves a combination of multiple fraud schemes. One of the techniques is exploiting roaming SIM cards or dialer malware, call divert and call forwarding, and social engineering techniques [2]. Fraudsters generate high traffic calls to costly destinations and get revenue from the sharing agreements.

C. Interconnect Bypass Fraud (IBF)

It is an illegal injection of traffic onto legal carrier operators. To commit this fraud, the fraudsters must have access to advanced technology like VoIP, and SIM-Box device, which enable them to convert international traffic calls to be domestic calls. Thus, SIM-Box device is the platform for committing interconnect bypass fraud and local SIM cards are used for routing international traffic calls masking from carrier operators, terminate them over the Internet through VoIP gateway device and the calls appear to be local. Operators of the intermediate and destination networks receive payment at the price of local calls.

The success of committing interconnect bypass fraud depends on how easy it is to obtain many SIM cards and the impacts of bypass fraud vary from country to country [3]. In some countries, where unregistered SIM cards are not allowed, and SIM-Box devices are illegal, the effect is less as compared to countries where obtaining of SIM cards by individual retailers is easy and there is no law enforcement for unregistered subscribers.

The main cause of interconnect bypass fraud is tariff difference between international traffic call and local call, specially the termination cost of international traffic calls is high. The terminator operators lose a toll free of the difference between international and local calls, so that the fraudster benefit from this revenue difference.

Evolving SIM-Box technology along with VoIP is the reason for interconnect bypass revenue loss, traffic congestion, and quality of service degradation [14]. The network cells where SIM-Box device operated, the voice calls becomes overloaded, and poor voice quality which results in customer dissatisfaction.

2.3 Interconnect Bypass Fraud Scenario

To explain how interconnect bypass fraud is committed, it is important to show the legitimate international calls route as shown in Figure 2.1.

Suppose caller A and recipient B, are residing in different countries A and B respectively [10, 12].

In a legitimate route of an international call the traffic path is:

- A caller (country A) makes a call to a call recipient (country B) over the legitimate home mobile operator and paying a charge for service usage.
- Now, the call is forwarded through cellular networks to the international gateway of caller (country A).
- The international gateway of country A redirects the received call to the interconnect carriers (transient operators) and paying a toll for the service provided by the transient carriers.
- The transient operators then route this call to the international gateway of a destination operator (country B) and paying a toll to the destination operator.
- The international gateway of country B terminates the call through its cellular networks and delivers to call recipient.

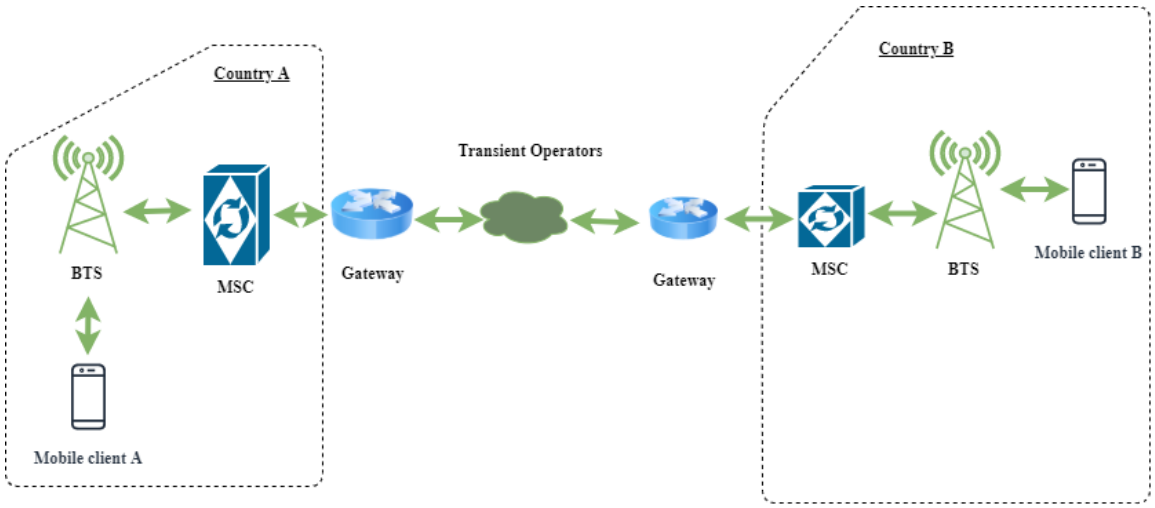

Figure 2.1: Official International Calls Traffic Path

In an interconnect bypass fraud route of an international call traffic path is presented in Figure 2.2:

- A mobile client A (country A) makes a call to a mobile recipient (country B) through home mobile operator and paying a charge for the service usage.
- The call generated by caller is forwarded to the international gateway of country A through cellular networks.
- The international gateway of country A routes the received call to a transient operator and paying toll.
- The transient bad operators then re-route this international call traffic to a special device called SIM-Box device, it can be placed anywhere in country B of call recipient.
- Now, the SIM-Box device acts as a GSM gateway connected to transient operator using VoIP and transient operator paying a toll to the SIM-Boxer with less price than the legitimate calls terminator.
- SIM-Boxer makes a separate call on home mobile network in country B to a call recipient using local operator SIM card and paying a charge for the service usage with local price scale. In fact, the SIM cards are often prepaid, with no customer identification.
- Since, they are using local SIM cards, the call appears to be local for call rating and billing purpose at home public operator.

- The fraudsters exploit the tariff difference between international call traffic (off-net) and local call traffic (on-net) price rate and avoids interconnect revenue of destination operator.

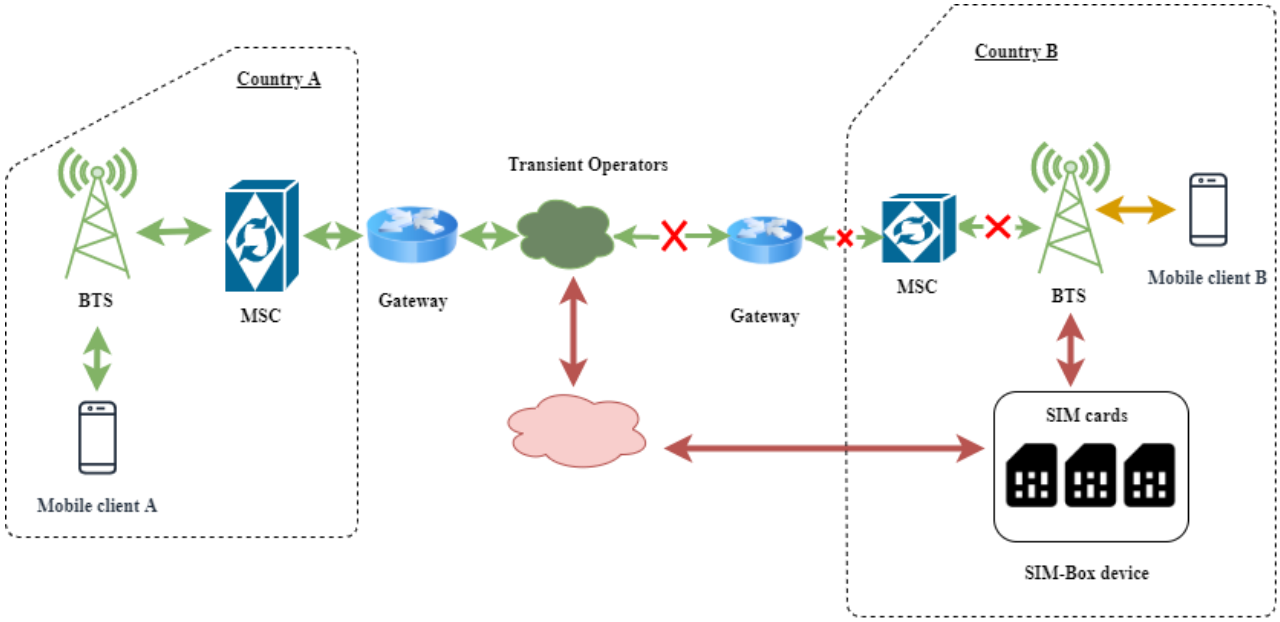

Figure 2.2: Bypass International Calls Traffic Path

2.4 Interconnect Bypass Fraud Detection

Currently, the common methods used to detect interconnect bypass fraud includes TCG and rule-based FMS [10]. TCG is a signal-based interconnect bypass fraud detection mechanism. A test call is generated for that specific route and the fraudsters assume that the call is a legitimate international call traffic and forward it to the interconnect bypass routes. It is used as a fast method of detecting interconnect bypass fraud, where operators test international routes to their network and analyze whether the calls are coming through an international number or local number. This is an effective method of identifying interconnect bypass fraud, but the method is costly to the operators to generate test calls to their entire network so that highly dependent on the number of test calls. For instance, if N number of test calls are generated, then only N number of interconnect bypass frauds can be detected provided that the call is not detected by the fraudster.

The FMS is based on predefined rules and thresholds. The rules are manually defined by domain experts targeting interconnect bypass fraudsters or any related fraudulent activities. The rules are relying on data features that are extracted from the CDR. However, advanced SIM-Box devices are designed to imitate the actual activities of a normal subscriber behavior [10]. This device makes fraudulent calls to bypass the predefined rules which avoids being detected by FMS.

The fraudsters are dynamic in terms of techniques and approaches. Thus, in the dynamic environment of today's technological evolution, best expert rules system even quickly deteriorates, and fraudsters start to behave differently to deviate the setup pattern of the rule [15].

An interconnect bypass fraudster uses the following techniques for human behavior imitation [10, 16]:

- SIM Card Migration: fraudsters implement SIM-Box devices attached with international gateways in different locations, and they swap the SIM cards between those gateways to imitate the mobility of the SIM card holders. The technical process could be done manually or automatically using software tools.
- SIM Card Automatic Rotation: Fraudsters are smart enough, if they operate one SIM card repeatedly, usage becomes high and they will be easily detected by rule-based FMS, so that they limit their usage for one SIM card by rotating other SIM

cards. They operate each SIM card for limited hours of a day to imitate normal customer's behavior.

- Short Message Service (SMS): unlike advanced SIM-Boxer, the SIM cards deployed in the traditional interconnect bypass device are used only for voice call service. However, this practice exposed them to be detected by rule-based FMS. In order to protect themselves from predefined rule-based FMS, they came up with advanced device with additional features and they can send and receive SMS in order to imitate normal customer behavior.
- Family Lists: advanced SIM-Box device has a feature of SIM cards segregation that assigns list of numbers to a specific SIM card to avoid detection setup.

All these efforts are done by fraudsters to simulate a normal subscriber behavior so that interconnect bypass fraud detection becomes more difficult. For any security measures put in place by the operators, the fraudsters find a way around to avoid being detected.

2.5 Data Mining (DM)

Basically, storing data in the data warehouse does not have any relevance for the operators. To realize the value of the data in the data warehouse, it needs to extract the hidden knowledge within the data. However, as the volume and variety of the data in the data warehouse grows, it becomes increasingly difficult for business analysts to identify trends and relationships between data using simple query and reporting tools.

According to Witten *et al*. [17], DM can be defined as the extraction of implicit, previously unknown, and potentially useful information from data. Whereas, Han *et al*. [18] defines it as Knowledge Discovery from Data (KDD) with the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, and other massive information repositories or data streams.

DM is an analytic designed process for exploring large amounts of data. Searching of consistent patterns and relationships between raw data. The ultimate objective of data mining is prediction. DM techniques are used by experts to quickly identify relevant patterns, transform the processed data into meaningful information to enable to make fast and accurate decision.

A. Knowledge Discovery from Data (KDD)

KDD is a method, trying to make sense of data and extract useful knowledge from it. It is a multistep process such as data collection, target data selection, data preprocess, data transformation, data mining, interpretation and evaluation. The most important step in the entire KDD processes is DM. It is the application of specific algorithms for extracting patterns from raw data [18]. The process of DM consists of several stages for knowledge discovery from big data, and Han *et al*. [18] give a more detailed summary of the KDD processes.

- Data cleaning to remove noise and inconsistent data.
- Data selection, where data relevant to the analysis task are retrieved from the database (selecting target data from raw data).
- Data integration, where multiple data sources may be combined.
- Data transformation, where the data are transformed and consolidated into forms appropriate for DM by performing summary or aggregation operations (preprocessing the target data).
- Data mining, which is an essential process where intelligent methods are applied to extract data patterns (strong pattern discovery from transformed data).
- Pattern evaluation to identify the truly interesting patterns representing knowledge based on interesting measures.
- Knowledge presentation, where visualization and knowledge representation techniques are used to present mined knowledge to users.

Therefore, one of the goals of this study is to discover a strong pattern from target data. It is a process that is comprised of several steps that include data selection, data preparation, data transformation, application of machine learning algorithms, evaluation of models and presentation of results.

B. Data Mining Process

Several standards have been developed for DM processes, such as sample-explore-modify-model-assess (SEMMA) and cross industrial standard process (CRISP-DM) [19]. In fact, each process model provides a life cycle of a data mining project. It comprises of the corresponding phases of a project, their respective tasks, and relationships between those tasks. The models usually emphasize independency from industrial application and technology used, but they can be generally divided into two categories, those that consider

industrial issues and those that don't. However, the academic models usually are not concerned with industrial issues [19]. CRISP-DM is the most frequently used and convenient for industry application domain [10, 20].

C. Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM provides a great framework for delivering data mining projects. As shown in Figure 2.3 [21].

CRISP-DM process model identified six phases within a typical data mining project such as:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation and
- Deployment

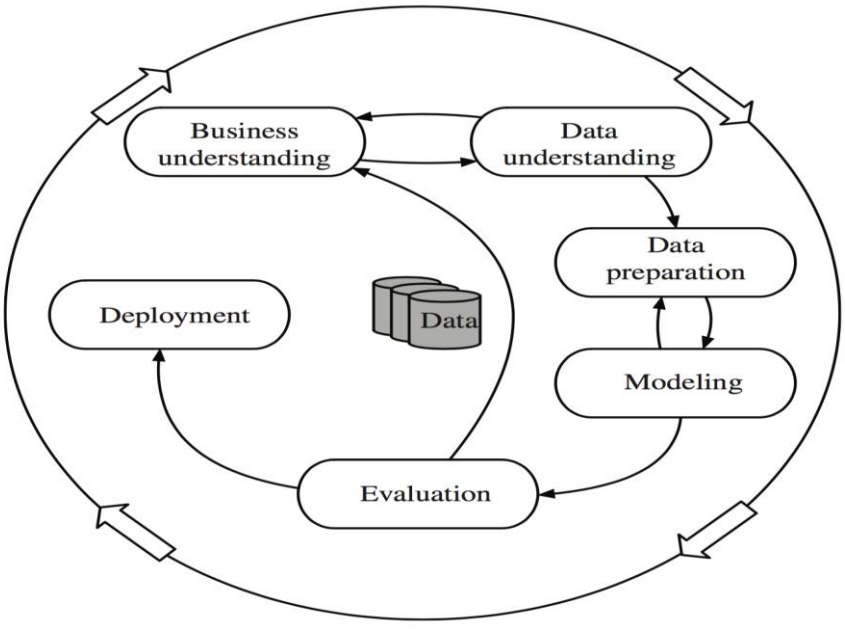

Figure 2.3: Phases of the CRISP-DM Process Model

I. Business Understanding

This phase focuses on understanding the project objectives and requirements from the business point of view. This phase converts the business problem to a data mining problem. There are various tasks involved in this phase such as determining a business objective, determining the data mining goal and producing a project plan. The goal of this phase is to uncover the factors that could influence the result of the project, miss understanding this step can lead us to trying to provide answer to the wrong questions. Hence, deep understanding of the business which provides us a comprehensive information about the services and products providing to the customers and which help us to understand the application domain.

II. Data Understanding

This phase is concerned with establishing the main characteristics of the data. It includes the data structures, data quality, and identifying any interesting subsets of the data. The tasks involved in this phase are as follow [22]:

- Data Collection
- Data Description
- Data Exploration
- Data Quality Verification

III. Data Preparation

This phase involves all the activities for building the dataset on which modeling tools can be applied.

The major tasks involved in this phase are stated as follows [22]:

- Data Selection
- Data Cleaning
- Data Construction
- Data Integration
- Data Formatting
- Feature Selection

ML algorithms are designed to learn appropriate dataset features to be used for making accurate decisions. The challenge is identifying a representative set of features (attributes) from the aggregated dataset to build a predictive model. Even though most learning techniques can select appropriate features from the given dataset and ignores inappropriate ones, but the performance in practice might be affected [20]. This is due to inappropriate selection of features by the ML algorithms so that the learning process must proceed after appropriate features selection from a given dataset using dimension reduction which yields simple interpretable and more dense features. In fact, both methods (attribute selection and dimensionality reduction) need to reduce the number of attributes in the given datasets, but a dimensionality reduction method do by creating new combination of attributes, whereas attribute selection method do by include and exclude attributes presents in the datasets with without changing them.

There are three general methods of feature selection algorithms such as filter methods, wrapper methods and embedded methods which are discussed below [20].

Filter method: applies a statistical measure to assign a score to each feature based on an independent assessment on the characteristics of the attributes. The attributes are ranked by the score and either selected to be kept or removed from the dataset. Widely-used filter method is Correlation-based Feature Selection (CFS) or correlation coefficient scores.

For this research work we applied CFS technique. It is basically a metric to evaluate the efficiency of features subset. According to Hall Mark [23], CFS is a good feature set that has features which are highly correlated to the output class yet unrelated to each other. The objective is to reduce Feature to Feature Correlation (FFC) and increase Feature to Class Correlation (FCC). The criteria defined using Pearson coefficient, which is essentially a ratio of FCC to FFC (FCC/FFC), a higher ratio indicates a better subset. It measures the worthy of as subset of features, by inspection of the individual predictive capability of the classifier and along with the degree of redundancy between them.

Wrapper method: evaluates a feature subset using ML algorithms that will be employed for learning process, i.e., selection of a set of attributes as a searching problem, whereas different combinations are also prepared, evaluated and compared to other combinations. A predictive model constructed by each combination is evaluated and assign a score based on model accuracy. It's is a recursive feature elimination algorithm.

Embedded Method: learns which attributes best contribute to the accuracy of the model while the model is being created.

IV. Model Building

This phase realizes the actual data mining operation involved and selecting modeling techniques, identifying modeling parameters, and evaluating the models developed.

a. Modeling Methods

The K-fold cross validation and percent split are the two mostly used methods for training ML algorithms.

K-Fold Cross-Validation: the dataset is partitioned into mutually exclusive and equal size K subsets. The model trained with $K - 1$ times and tested on the K^{th} subset**.** The processes are repeated iteratively changing the test subset from the 1^{st} to K^{th} subset, to get an average error rate of each subset which is, therefore an estimate error rate of the classifier [24]. Based on the results across each iteration, the test error of the developed model can be determined by root mean square error (RMSE), and misclassifications error rate.

Percent Split: the datasets are split into two parts, then the first part of the dataset is used for model building and the remaining part is used for testing.

b. Performance Assessment Measurement

Accuracy, precision, recall, f-measure, RMSE, and receiver operating characteristic (ROC) are the commonly used standard evaluation techniques.

The evaluation measures depend on the results of a confusion matrix which shows a comparison between the actual and predicted outcomes for a set of labeled instances. Table 2.1, shows the structure of a confusion matrix. The rows represent the actual values of positive and negative classes whereas, the columns represent the predicted values of positive and negative classes.

Table 2.1: Confusion matrix of model classification

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The confusion matrix terms and values can be described as follow:

I. True Positive (TP): The actual number of fraud instances that are classified as fraud instances.
II. False Positive (FP): The normal instances that are classified as fraud instances.
III. True Negative (TN): The normal instances that are classified as normal instances.
IV. False Negative (FN): The actual number of fraud instances that are classified as normal instances.

Therefore, the confusion matrix values such as TP and TN are correctly predicted results of the classifier. Whereas, FP and FN are incorrectly predicted results [20]. So, the model performance can be measured as follows.

Accuracy: realizes the ratio of correct classification capability of the model. It computes summation of TP and TN divided by all data instances. Mathematically, it can be expressed by equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: focuses only on positive instances. It realizes that those predicted positive are actually positive. Mathematically, it can be expressed by equation (2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: realizes the true positive rate. It measures the proportion of actual positives which are correctly identified as positive by the model. Mathematically, it can be expressed in (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F-Measure: it is the combination mean of precision and recall. Mathematically, it can be expressed by equation (4).

$$F - Measure = 2 * \frac{Precision \ . \ Recall}{Precision \ + \ Recall} \tag{4}$$

Root Mean Squared Error (RMSE): measures the average magnitude of the error. It's the square root of the average of the squares of the difference between predicted and corresponding observed values. Since, the errors are squared before they are averaged, it gives relatively high weight to large errors. It is most useful when large errors are particularly not required. Mathematically, it can be expressed by equation (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} \xi i^2} \tag{5}$$

Where $i = 1$, $\xi = (\text{predicted} - \text{observed value})$ and $\text{n} = (\text{predicte} + \text{observed values})$ *or sample size*.

Receiver Operating characteristics (ROC): it's a graph of FP versus TP rate. It measures the area; the ability of the model correctly classifies the test data. A model that cover a larger area in the graph plot has a better classification performance. According to Ravichandran *et al*. [25] a good predictive model, the RMSE values should be between 0.3 and 0.5, and the RMSE greater than 0.5 is related to a bad predictive model.

2.6 Types of Machine Learning (MLP)

There are four machine learning methods such as supervised, unsupervised, semi-supervised and reinforcement. In the following sections, we describe different ML types and their respective algorithms.

2.6.1 Supervised Machine Learning

Supervised machine learning can be defined as a learning function $f(x) = y$ where y is the class of the sampled data, and x denotes the features of this class. Here, we have input variables (x) and an output variable y and we use an algorithm to learn the mapping function from the input to the output.

In supervised learning, the models are trained with data that have been pre-classified [21]. However, during supervised learning a great caution must be taken in order to ensure the training data for the experiment should be picked proportionally and correctly classified. The supervised learning methods are categorized based on the structures and objective functions of learning algorithms. Popular categorizations include artificial neural network, support vector machine, and decision trees [26].

1) Artificial Neural Network (ANN)

ANN is a powerful data modeling tool that can represent complex input to output relationships [27]. The motivation behind for the development of neural network technology is coined from the desire to implement an artificial system that could perform intelligent tasks like human brain does.

According to Williams et al. [28], neural networks represent the brain symbol of human being for information processing. ANN has the capability to learn from its environment through an iterative process of adjustments applied to its synaptic weight and bias level. It's also able to improve its performance through learning.

There are various number of learning algorithms for designing ANN. One differs from the others in the way in which the adjustment to a synaptic weight of neurons is formulated. Learning algorithms such as error-correction, memory-based, competitive and Boltzmann learning are among the popular learning algorithms for ANN. Its learning paradigm is either supervised (associative learning) or unsupervised (self-organizing). In the case of supervised modeling, there is a need to train the model by input and output pattern. However, for the case of unsupervised modeling, only requires input patterns from which it develops its own representation of the input stimuli [20].

ANN can be classified in to three neural network architectures such as, feed forward neural network (FFNN), recurrent neural network (back propagation) and self-organizing map (SOM) [28]. In FFNN activation is propagated forward to the output units all the way

through the intervening input-to-hidden and hidden-to-output weights. It is the first and simple type of ANN and recurrent neural network on the other hand is a dynamical network with cyclic path of synaptic connections which serve as the memory elements for handling time-dependent problems. SOM mainly is applied for cluster analysis, especially for unsupervised learning. ANNs are considered as one of the most efficient pattern recognition, regression, and classification tools [28, 29].

Perceptron

A perceptron serves as a basic building block for creating a ANN. It is the simplest kind of ANN which consists of a single neuron that can receive multiple inputs and produces a single output. Perceptron is used to classify linearly separable classes [20]. As shown in Figure 2.4 [30], a perceptron takes a list of values and calculates a linear combination of these inputs. The output can be 1 if the result is greater than some threshold value or -1 otherwise based on the selected function.

Each of the input vectors received by the perceptron has been weighted based on the significance of its contribution for obtaining the final output. However, the essential learning problem is to determine a weight of the vectors that cause the perceptron to produce the correct output for each of the given training example data. The common way that the perceptron algorithm is used for learning from a bulk of training instances is to run the algorithm repeatedly through the training dataset until it finds a prediction vector which is correct across all the training sets.

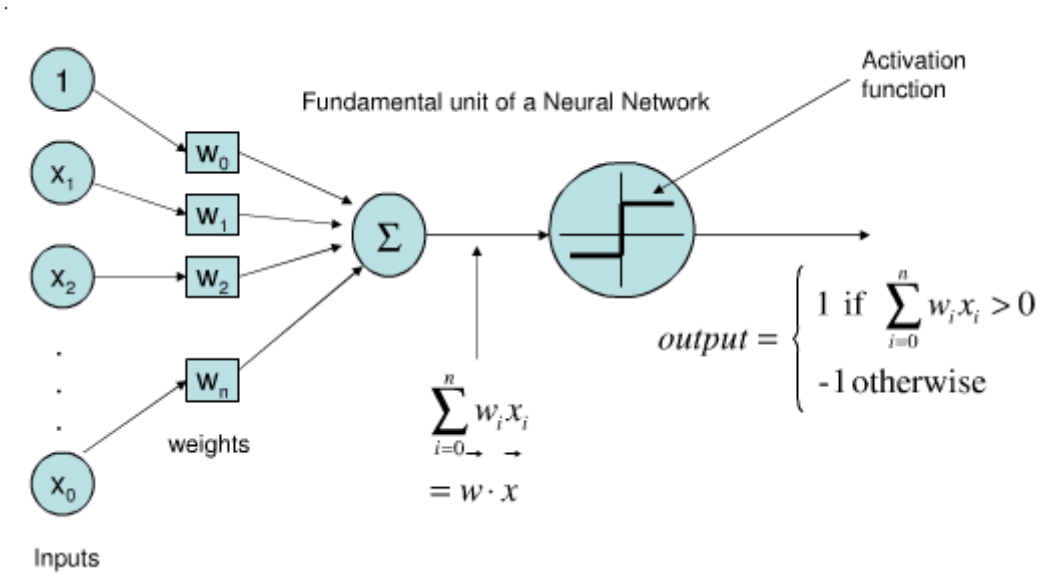

Figure 2.4: Perceptron with multiple inputs and single output

In a perceptron model, the input vectors received by the perceptron is first multiplied by their respective weights and then, all these weighted inputs are summed together. This summing junction is then fed to an activation function, which compares it to a predetermined threshold Θ, for obtaining the final output using Equation (6). If the weighted sum is greater than the threshold value, then the outputs of the perceptron can be 1, otherwise, it can be 0 (-1) depending on the activation function we provided to perceptron.

$$\sum_{i=1}^{n} \text{wixi} = w1x1 + w2x2 + \cdots + wnxn \tag{6}$$

An activation function is a function used to transform the activation level of a neuron into an output result. There are various number of activation functions that can be used with in the perceptron model, but the *step(unit)*, *sign*, *linear*, and *sigmoid* functions are the most commonly used function [20].

- *Step(unit)* $f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$
- Sign $f(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$
- Linear $f(x) = x$
- Sigmoid $f(x) = \frac{1}{1+e^{-x}}$

All the above listed activation functions are triggered at a threshold value of Θ = 0. However, it is more convenient to have a threshold other than zero. For that cases, we have a bias (b) is added to the perceptron model inputs. The role of bias (b) in the perceptron model enable us to shift the decision line so that it can separate the inputs into desired classes. One of the aims of training the perceptron is to determine the optimal weights and bias value at which the perceptron is triggered.

Multilayer Perceptron (MLP)

MLP is a composition of perceptron, they are connected in different ways and operating on different activation functions. A single perceptron cannot solve any classification problem for non-linearly separable data instances [20].

The two major problems observed in a single perceptron model are:

- Cannot classify non-linearly separable data instances.

- Complex problems, that involve a lot of parameters cannot be solved by single-layer perceptron.

Therefore, a non-linearly separable problem is solved by using MLP [31]. In MLP neural network, each perceptron receives a set of inputs from other perceptron, and multiplied with their respective weight, then the sum of the weighted inputs is above or below threshold value called Θ to transform the activation level of the neuron to the expected output.

MLP comprises of three layers, the input layer, the hidden layer, and the output layer as show in Figure 2.5 [30].

- Input layer consists of input nodes which represent the system's variable, which receive information from the outside world to the NN.
- Hidden layer consists of nodes which facilitate the flow of information from the input to the output layers, have no connection with outside world, just performs the computation. In addition, MLP has one or more hidden layers.
- Output layer responsible for computations and transferring information from the hidden network to the outside world, which represent the system's classification classes.

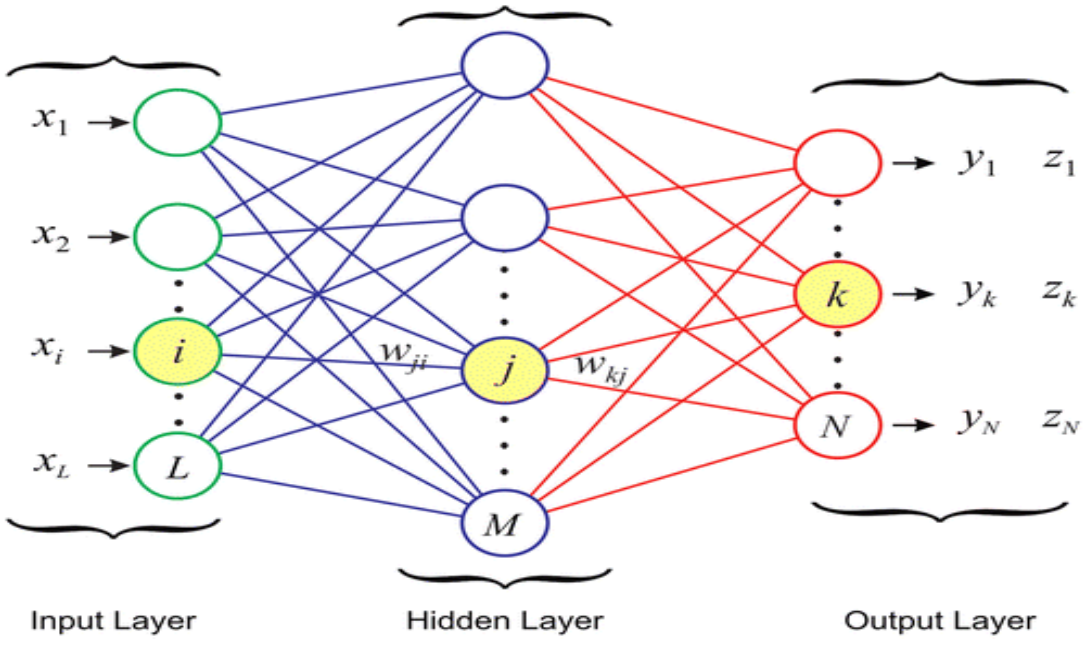

Figure 2.5: A multilayer perceptron neural network

2) Random Decision Forest (RF)

RF is a method that operates by building multiple decision trees during training phase. The decision of most of the trees is chosen by the random forest as the final decision. A RF algorithm is a supervised classification algorithm. As the name suggests, this algorithm creates the forest with several trees. In RF classifier, the higher the number of trees in the forest gives the high accuracy results [36]. A RF builds multiple decision trees and combines them together to get a more accurate and stable prediction. A RF algorithm randomly selects features from sample dataset to build several decision trees and averages the results. Most of the time it is trained with the bagging method. The basic idea of the bagging method is a combination of learning models increases the overall model result. This classifier grows independently and identically distributed random vectors, then each vector casts a unit vote for the most popular class at the input [36].

The output of random forest is decided by the votes given by all individual trees (nodes). Each decision tree is built by classifying a random sample of the input data using decision tree algorithm. The RF model decides the classification result of the testing data after collecting the votes of all the tree models. The working principle of decision tree is somehow like the rule-based system. Providing the training dataset with features and targets, the decision tree algorithm will come up with some set of rules. The resulting set rules can be used to perform the prediction on the test and evaluation dataset. One of the advantages of the random forest classifier will handle the missing values and more trees in the forest, it won't overfit the model. However, many trees can make the algorithm to slow for prediction and ineffective for real-time demanding predictions. So that, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees which results in a slower model. The RF algorithm works on different parts of the dataset, result in high accuracy. It is predominantly the bagging approach [37].

The RF algorithm builds many decision trees, the number of trees to build is input variables, which can be selected during the learning phase. Every decision tree is learned with a random subset of features from a sampled training set with replacement. The output is decided by the votes given by all individual trees. Each decision tree is built by classifying the samples of the input data using a tree algorithm. Then, every tree will be used to classify testing data. Each tree has a decision to label any testing data. This label

is called a vote. Finally, the forest decides the classification result of the testing data after collecting the votes.

The RF algorithm is very robust to noise and missing values, i.e. small changes in the training dataset will have little impact on the final decisions made by the resulting model [20]. While building a decision trees, the model builder algorithm may select a random subset of the features available in the training dataset. So that, during the process of building each decision tree, only a small number of the available variables is selected. This essentially reduces the computational cost and makes suitable when there are many input variables and little features and it can handle large dataset with higher dimensionality and for building patterns and detect outliers [26]. The strength of individual trees in the forest and the correlation between them determines the generalization error of the forest and its trees [36].

The RF algorithm is a great choice for model building for several reasons such as less training time, runs efficiently for large data, it produces highly accurate prediction, no need variable selection RF model builder is able to target the most useful variables, each decision tree is not influenced by the other decision trees while constructing and the model builder uses multiple trees reduce the risk of overfitting.

3) J48 Decision Tree (C4.5)

A J48 decision tree model is developed in a top-down approach using recursive divide and conquer manner. The training dataset is recursively separated into smaller subsets, a tree like structure. It is a supervised ML algorithm that decides the target value (dependent variable) based on the values of the input variables. A decision tree (DT) is used to learn a classification function which concludes the value of a dependent attribute given the values of the independent attributes as input [38].

Decision tree offers many benefits to data mining tasks, some are as follows [38]:

- It is easy to understand by the end user.
- It can handle a variety of input data type such as nominal, numeric and textual.
- It selects the attribute with the highest normalized information gain.
- Able to process erroneous datasets or missing values.
- High performance with small number of efforts.
- It can be implemented data mining packages over a variety of platforms.

The size and complexity of the decision tree has an effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains; the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [39].

4) Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm. It constructs a hyperplane or set of hyperplanes in a higher (maybe infinite) dimension space, which can be used for classification and regression. However, there are many hyperplanes that might classify the data points, but the reasonable choice is the optimal hyperplane the one that represents the largest separation, or maximum margin between the two classes. Optimal separation can be achieved by this hyperplane that has the largest distance to the nearest training data points of any of the data class. The larger the margin, the lower the generalization error of the classifier [20].

SVM goal is to find a hyperplane in an N-dimensional space (N number of features) that classifies the data points (hyper vectors) in the best possible way. It is a method for classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension, where it finds a hyperplane that separates the data groups using essential training tuples called support vectors [32]. SVM was the first proposed kernel-based algorithm [40]. It uses a kernel function to transform data from input space into a high dimensional feature space in which it searches for a separating hyperplane.

As show in Figure 2.6 [20], to split the two classes of training data points, there are many possible hyperplanes that could be selected. The objective is to find a hyperplane that has the maximum margin, i.e. the maximum distance between support vectors of both classes. Maximizing the margin distance provides optimal separation so that future data points can be classified with more accurate.

Hyperplanes are decision boundaries that help us to split the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Important thing the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then

the hyperplane becomes a two-dimensional plane. Otherwise, it becomes difficult to imagine when the number of features exceeds three.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we can maximize the margin of the classifier, i.e. avoiding the support vectors will change the position of the hyperplane. On the other hand, these are the points that help us build our SVM [40]. The training vectors Xi are mapped into a higher dimensional space by using the function θ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. So, support vectors are elements of the training dataset that would change the orientation and position of the separating hyperplane. Support vectors are the critical elements of the training dataset because they determine the weights and decision boundary.

SVM differ with other ML it always finds a global minimum [40], and it is a geometric interpretation. The training time is extremely slow, but they are highly accurate, can model complex nonlinear decision boundaries. It is less prone to over-fitting than other methods [32].

Margin Maximization

SVM is just like a NN, important difference is SVM use optimization techniques for maximizing the margin ('street width') this is achieved by reduce the number of weights that are ranging from non-zero to a few that correspond to the important features that matter in deciding the separating line(hyperplane).

The main goal of SVM is to classify the support vectors through a hyperplane that has the maximum distance to the nearest support vectors (data points) on both side of class and extend this to non-linear boundaries using kernel trick [40]. As mentioned earlier, there are many classifiers (hyperplanes) that separate the data points. However, only one hyperplane achieves maximum separation and the SVM analysis finds a hyperplane the margin between the support vectors is maximized [40].

As depicted in Figure 2.6 [20], a separating hyperplane can be defined as the set of data points x expressed as, $w^T x + b = 0$, where w is a weight vector perpendicular to the hyperplane, x an input vector and b is the bias (offset) of the hyperplane $w^T x + b = 0$ from the original point along the direction of w . Given a label training data points x for

two classes (class 1 and class 2), the label data points can be expressed as $y_i \in 1, -1$ and given a pair of (w^T, b), then the data points x can be classify into either of the two classes (class 1 or class 2), according to the sign of the kernel function $f(x) = sin(w^T x + b)$.

Therefore, the linear separating of the data points x into these classes can be expressed as in Equation (7) and Equation (8) Therefore, the linear separating of the data points x into these classes can be expressed as in Equation (4.2) and Equation (4.3)

$$x_i . w + b \geq +1, \quad for\ y_i = +1 \tag{7}$$

$$x_i . w + b \leq -1, \quad for\ y_i = -1 \tag{8}$$

If Equation (7) and Equation (8) combined together to form, Equation (9)

$$y_i (x_i . w + b) \geq 1 \quad i = 1, \ldots, n \tag{9}$$

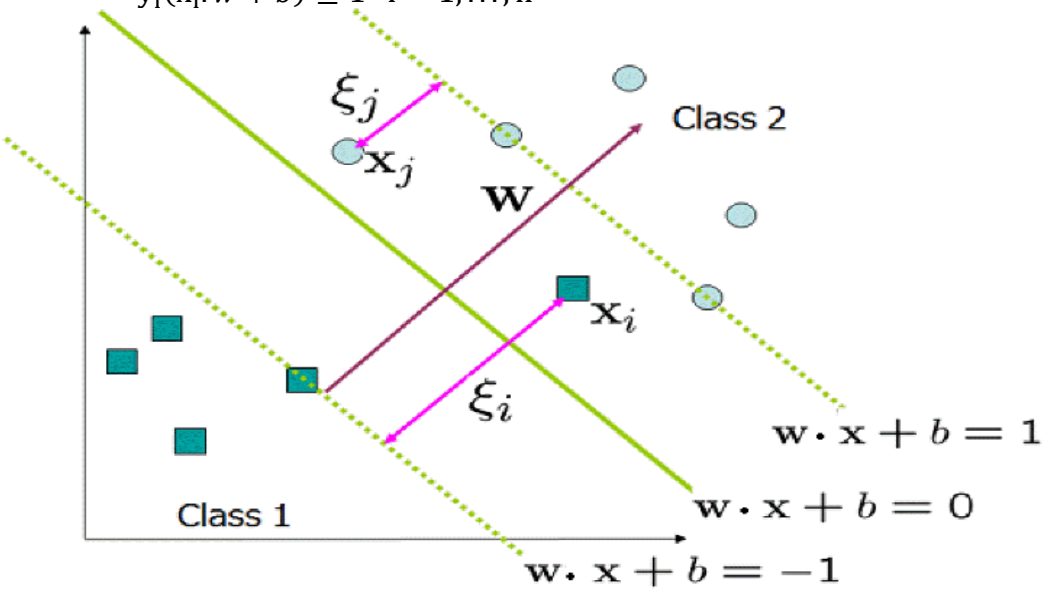

Figure 2.6: SVM classification

2.6.2 Unsupervised Learning

As the name suggests, training dataset are a collection of unlabeled examples without having a specific desired outcome or correct answer. The learner then attempts to automatically find patterns in the data set by extracting useful features and analyzing data structure.

The aim of unsupervised learning is to identify patterns in the data that extend our knowledge and understanding of the world that the data reflects [21]. The most famous

unsupervised learning methods include k-means clustering, hierarchical clustering, and self-organization map.

2.6.3 Semi-supervised Learning

In a semi-supervised learning, the given data are a mixture of classified and unclassified data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data [41].

This method is particularly useful when extracting relevant features from the data set is difficult, and labeling examples is a time intensive task for domain experts.

2.6.4 Reinforcement Learning

Reinforcement learning is a type of machine learning where an agent learns how to behave in an environment by performing actions and observing the results.

2.7 Data Mining Tools

There are a growing number of commercial data mining tools on the market. However, the important features of data mining tool to be choice include data preparation, available data mining operations (algorithms), product scalability and performance, and facilities for understanding results [22]. However, every tool has its own advantages and disadvantages. There are also open source tools that have been developed by different research communities and data analysis enthusiasts. For this research work one of the top open source tools available for data mining is stated in the upcoming section.

Waikato Environment for Knowledge Analysis (WEKA)

WEKA is a collection of machine learning algorithms for data mining operations. The tool first released in 1997 and after long journey now the latest version available is WEKA 3.8.11. It also has GNU general public license. It is also platform independent software because developed using java. The tool has different algorithms and these algorithms can either be applied directly to a dataset or can be called from java source code indirectly. The workbench contains a collection of several tools for visualization, algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality. It loads data file in the format of ARFF and CSV. Some of the major reason WEKA is the primary tool for this research, it is recommended by different literatures [3, 42], and full level of control because of open source, can be integrated into

other java packages, best suited for association rules, and comprises of stronger machine learning techniques.

After the datasets were transformed based on the requirements and suitable to the selected tool. Since the tool support both CSV and ARFF file formatting, but it prefers to load data in the ARFF format, it is an extension of the CSV file format where a header is used, that provides metadata about the data types in the columns. The directives start with at symbol (@) and a directive define the name and datatype of each attribute (@ATTRIBUTE SERV_NO INTEGER) and there is a directive to indicate the start of the raw data (@DATA). In addition, the lines in an ARFF file that start with a percent symbol (%) which indicate a comment and values in the raw data section that have a question mark symbol (?) which indicate an unknown or missing value, especially during testing.

Chapter 3 : Related Work

This Chapter reviews the various researches conducted on interconnect bypass fraud detection.

Ibrahim and Hussamedin [10] developed a model using data mining techniques that mines a huge number of mobile operators CDR data and detect SIM cards that were used to bypass international calls. They compared two classifier algorithms, Support Vector Machine (SVM) and Decision Tree Algorithms. Due to shortage of fraud labeled data, they used unsupervised learning algorithms to cluster SIM cards. However, accuracy of each classifier algorithm, and what CDR features extracted are not mentioned in the paper.

A research conducted by Ilona *et al*. [12] developed a model based on data mining techniques and they compared three classifier algorithms, Alternative Decision Tree(ADT), Functional Decision Tree (FDT), and Random Forest Algorithm (RFA) for detection of fraudulent SIM cards deployed in SIM-Box device. Based on their experimentation, RFA and FDT provide the lowest false positive and the lowest false negative respectively. Their dataset contains CDR of 500 International Mobile Equipment Identity (IMEI) of fraudulent SIM cards and 93,000 legitimate accounts. They took one-week customers data, extracted 12 CDR fields for characterizing customer's patterns of legitimate or fraudulent and then transformed into 48 features to characterize voice call communication patterns of legitimate and fraudulent subscribers. For experimentation, the data set was split into two parts, 66 percent and 34 percent of the labeled accounts were used for the training and testing respectively. WEKA data mining tool. The accuracy of their best classification rule was 99.95%.

Abdikarim *et al*. [15] developed a model using data mining techniques and they compared two classifier algorithms, ANN and SVM. Their data set consisted of 234,324 CDR calls made by 6,415 subscribers for a single cell location of two months' data. Among those data, 2,126 were interconnect bypass fraud subscribers and 4,289 were legitimate subscribers. The researchers extracted nine CDR features for characterizing patterns of legitimate or fraudulent and they used these extracted features to train the ANN classifier, where three architectures of neural network were constructed and three hidden layers. They found that the best design was the two hidden layers, each having five hidden neurons. The accuracy of their best classifiers was 98.7% with 20 accounts wrongly classified as false positive.

Kahsu Hagos [20] developed a model using data mining techniques based on subscribers CDR taken from ethio telecom and to detect SIM-Box fraud. The author compared three classifiers algorithms, SVM, ANN and RF. Also, the author took two months' customers data, extracted 8 CDR fields for characterizing customer's patterns and then transformed into 12 features to characterize voice call communication patterns of legitimate and fraudulent subscribers. In addition, the author used three user profiling data sets, 4 hours, daily and monthly data aggregation. These three algorithms along with the three data sets were applied in building the models. Results of the work show that RF performed better among the three algorithms with accuracy of 95.99% and a lesser false-positive on the 4 hours aggregated data set.

A research by AlBougha Redwan [16] compared four classifier algorithms, boosted (BT) classifier, SVM, ANN and Logistic Classifier Algorithm (LCA). Results of the work show that BT and LC performed better than the other algorithms with a false-positive ratio less than 1%. Neural networks had an accuracy rate of 60% with a false positive ratio of 40%. They concluded that BT and SVM classifiers are among the best algorithms to be used in the SIM-Box fraud detection because of their high accuracy and low false-positive ratios. However, the paper doesn't mention any CDR features extracted.

In summary, the major limitations observed in the related works are stated as follows:

First, some researchers conducted their studies through unsupervised learning with limited CDR datasets. However, in telecommunication, there are more legitimate subscribers than fraudsters so that the model trained with unlabeled and limited fraudster dataset might not catch the fraudster in the production environment. Second, some of the researchers also used limited number of attributes, which might not reflect the actual behavior of fraudsters and results in high false alarm rate (FP increases). Third, some researchers used a wider or no data aggregation period (hourly, daily, weekly, monthly), etc. Too broad granularity level has high risky revenue loss and too late to take preventive action.

Therefore, for this research we would add more fraud detection features to characterize subscriber's behavior based on attributes value and significance for improving the performance and accuracy of fraud detection. In addition, we will train and test our model on a larger data sample and four data aggregation levels (4 hours, 8 hours, 12 hours, daily, and weekly).

Chapter 4 : Data Collection and Preparation

4.1 Data Collection

For this thesis work, mobile subscribers CDR are collected for two months, February and March 2019 from ethio telecom billing system and prepared for the experiment. The CDR are stored in a dedicated server allocated for this research work.

The CDR contains complete information of the call, including subscriber service number, called number, duration of the call, call start time, call end time, amount charged, base station (cell id), destination location and the reset are show in Table 4.1. Hence, the CDR gives us the detail of each service transaction in a mobile network such as voice call, SMS, and data usage (GPRS).

As soon as the data is received, it is imported to a dedicated SQL server both labeled (fraud) and unlabeled raw data, and the process continued for two months starting from February 01, 2019 to March 31,2019. Due to massive size of the data, the two months' data is agreed upon by the experts at ethio telecom. Thus, each day around 300 million CDR were recorded including voice, data (GPRS), and SMS and more than 1 million equipment identity logs had been collected for about 41 million active mobile subscribers.

A sample of 30,000 bypass fraudulent service numbers were provided by the fraud management section under the ethio telecom security department. These service numbers were verified and blocked from accessing the communication network due to their fraudulent behavior. In addition, those service numbers with voice CDR, data (GPRS) CDR, SMS CDR and device identity (IMEI) history logs are collected from another data warehouse section are also imported into the SQL server.

4.1.1 Data Description

The primary objective of data understanding in the data mining process is to identify attributes, examining their corresponding data structure and evaluating their importance for characterizing the pattern of a subscriber. During this process, the domain experts were engaged for data assessment and exploring.

Each attribute name with its respective data type pair is used in the CDR file structure to record the communication information or service used. Table 4.1 depicts the CDR attributes and their descriptions.

Table 4.1: CDR attributes list and description

No.	Attribute	Description
1	CALLING_NBR	Calling Number
2	CALLED_NBR	Called Number
3	EVENT_BEGIN_TIME	Format：YYYYMMDDHH24MISS
4	DURATION	Duration, charging unit, in second
5	CALL_TYPE	Calling type: 0: Unknown; 1: Calling; 2: Call Forward; -1: Other
6	CDR_TYPE	CDR type
7	EVENT_END_TIME	Event End Time
8	BILLING_NBR	Bill settlement number
9	CALLING_IMEI	Caller device id
10	CALLED_IMEI	Called device id
11	LAC_A	Location area of caller
12	CELL_A	Caller base station transceiver (BTS)
13	LAC_B	Peer Party Area, location area of called
14	CELL_B	Peer Party Cell, called base station transceiver (BTS)
15	UP_DATA	Data usage, upload bytes
16	DOWN_DATA	Data usage, download bytes
17	CALLING_IMSI	Calling subscriber identity
18	CALLED_IMSI	Called subscriber identity
19	CALLED_AREA_ID	Called area ID
20	SERVICE_TYPE	service type, 1: voice; 2: data; 3:SMS; 4: VAS
21	TOTAL_FEE_MIN	Total fee paid in minutes for the service

The attributes in Table 4.1 have been selected based on the literature studied on the typical characteristics of interconnect bypass fraud behavior and further discussion held with domain experts.

4.1.2 Verifying Data Quality

The CDR data obtained are not directly used for data mining, since they may contain inconsistent, noise, irrelevant, and redundant data. Before the development of the model, the data must go through the preprocessing phases such as, data selection, data construction, data integration, data aggregation, feature selection, handling missing values, outlier removing and data formatting. Since interconnect bypass fraud detection is highly dependent on customers CDR, the quality of the data must be preserved, especially fraudulent subscriber CDR. However, the subscribers CDR records were stored in different distributed databases. Due to this reason, the data was extracted from different databases based on service type, so that for some service numbers, they don't have records. For instance, if one subscriber didn't send SMS within the required data access time span, then this customer will not have CDR in the SMS database and this will decrease the number of extracted records from the original dataset numbers and that makes the data incomplete, so that, null values are replaced by conventional value for any of the missing fields. The telecom CDR raw data have noise, massive in volume, and may originate from a bulk of heterogeneous sources and hence, understanding the data deeply is a vital prerequisite for data preparation.

4.2 Data Preparation

This phase involves all the activities for building the final dataset on which modeling tools can be applied directly. However, data preparation is the most time-consuming aspect of data mining, especially in massive amount of data.

The major tasks involved in this phase are:

- Data Selection
- Data Cleaning
- Data Construction
- Data Integration
- Data Formatting
- Data Aggregation
- Feature Selection
- Removing outlier

4.2.1 Data Selection

Data selection is a process where data relevant to the analysis task are retrieved from the database. Selecting the relevant subsets of records and attributes for data mining algorithms requires PL/SQL query capabilities.

For a pragmatic data mining application design and implementation, sampling is a critical step especially for massive telecom data, since collecting and processing all telecom massive data is difficult. This big data process using on hand platform is overhead task so, random sampling is required to reduce the size of the data and to make it manageable.

Due to the massive size of the telecom CDRs, a discussion was held with domain experts for the data selection and agreed to take the CDR randomly. The incorporated subscriber numbers were selected using random sampling for the normal subscribers whereas, one-week fraudulent service numbers were provided from the fraud management section of ethio telecom. The detail of the data selection is presented in Table 4.2.

Table 4.2: Raw data selection statistics from database

	Fraud	**Normal**	**Total**
Subscriber numbers	30,000	415,717	445,717
Voice Records	10,623,682	306,236,821	316,860,503
SMS Records	147,811	401,930	549,741
Data Records	60,319	2,041,839	2,102,158
IMEI History Logs	2,208,470	6,208,470	8,416,940

4.2.2 Data Cleaning

The goal of data cleaning is to remove or repair missing, unnecessary, and inconsistent CDR records from the collected raw data in order to provide high quality information to ML algorithms and to get accurate prediction. As such data cleaning needs a domain understanding to acquire important data features that represent the required information for the target ML algorithms.

The telecom massive CDRs records are prone to incomplete values, missing values, and noisy data. For instance, service numbers less than 12-digit (short code numbers) are removed from the data, records of other than mobile service numbers (fixed telephone) are removed, service numbers having prefix “251” which indicate country code are also

removed, null value of fields filled with zero (0) value, and encoded data are transformed into appropriate data format, especially date time fields.

4.2.3 Data Construction

At this stage, additional attributes which are necessary for this research are derived from the original attributes. The derived attributes are expected to characterize the behavior of interconnect bypass fraudsters. Accordingly, 26 attributes were derived which ultimately represent the behavior of a SIM-Boxer based on service type: voice, SMS, data and device identity. However, one of the challenges of the interconnect bypass fraud is that the attributes have extremely similar behavior with the legitimate users. Table 4.3 shows the derived attributes and their description.

Table 4.3: List of Derived Attributes

No.	Derived Attribute	Description
VOICE OUTGOING CALL CDR		
1	**VOUT_TOT_DIST_NBR_CALL**	Total number of distinct peer party called
2	VOUT_TOT_NBR_CALL	Total number of peer party called
3	**VOUT_CALL_DIV**	Ratio of total distinct outgoing calls to total outgoing calls (Outgoing calls diversity ratio)
4	**VOUT_CALL_DIF_SEC**	Average time different between two consecutive calls in seconds (call frequency)
5	VOUT_TOT_USAGE_MIN	Total number of times used in minutes
6	VOUT_TOT_FEE_MIN	Total fee paid for service used in minutes
7	**VOUT_TOT_DIST_CELL**	Total number of distinct cells used for outgoing calls
8	VOUT_TOT_CELL	Total number of cells used for outgoing call
9	**VOUT_CELL_DIV**	Ratio of total distinct number of cells to total number of cells used (Outgoing calls cells diversity)
VOICE RECEIVED CALL CDR		
10	**VIN_TOT_DIST_NBR_CALL**	Total number of calls received from distinct peer party
11	VIN_TOT_NBR_CALL	Total number of calls received from peer party
12	**VIN_CALL_DIV**	Ratio of total distinct received calls to total received calls
13	CALL_RATE	Ratio of total received calls to total outgoing calls
14	**VIN_TOT_DIST_CELL**	Total number distinct cells for calls received from peer party

No.	Derived Attribute	Description
15	VIN_TOT_CELL	Total number cells for calls received from peer party
16	**VIN_CELL_DIV**	Ration of total number of district cells to total number of cells used
	SMS SENT FOR PEER PARTY CDR	
17	**SMS_TOT_DIST_NBR**	Total number of SMS sent for distinct peer party
18	SMS_TOT_NBR	Total number of SMS sent for peer party
19	**SMS_DIV**	Ratio of total number of distinct SMS to total number of SMS sent
20	SMS_TOT_FEE	Total fee paid for SMS
	SMS RECEIVED FROM PEER PARTY CDR	
21	**SMS_R_TOT_DIST_NBR**	Total number of SMS received for distinct peer party
22	**SMS_R_TOT_NBR**	Total number of SMS received for peer party
23	**SMS_R_TOT_DIV**	Ratio of total number of distinct SMS to total number of SMS received
	MOBILE DEVICE LOG HISTORY	
24	**TOT_DIST_NBR_IMEI**	Total number of distinct mobile device used
25	**TOT_NBR_IMEI**	Total number of mobile devices used
26	**IMEI_DIV**	Ratio of total number of distinct mobiles to total number of mobiles used
27	**EVENT_CYCLE**	Service usage time

Legend:
DIF : Different
DIST : Distinct
DIV : Diversity
MIN : Minute
NBR : Number
R : Received
SEC : Second
SER : Service
TOT : Total
VIN : Voice incoming call
VOUT : Voice outgoing call

As shown in Table 4.3 those bolded attributes are our newly derived attributes. These attributes are based on closed investigation on the typical characteristics of interconnect

bypass fraud behavior and a discussion held with the domain experts. In addition, in this study some of the major behavior that we observed in fraudulent subscribers CDR are:

- Fraudulent customer has high volume of calls and prepaid recharge.
- Fraudulent customer has high distinct type of calls.
- Very low (absent) SMS and data traffic.
- The time difference between two consecutive calls, the end of first calls and the start of next calls are low.
- The traffic ratio for total incoming calls versus total outgoing calls is low. This might be SIM cards that are inserted in SIM-Box device have very few incoming calls.
- The fraudulent customer doesn't have valid profile (address). They use theft or fabricated address.

4.2.4 Data Integration

After data construction phase is finished the next relevant part is data integration, which is a process of bringing all the data records together into single instance records. However, integrating heterogenous tables using outer join operation have many challenges. In addition, different service types such as, outgoing calls, incoming calls, SMS sent, SMS received, data (GPRS) used and device identity (IMEI) history logs records for both normal and fraud samples are stored in different tables and then bringing all these records based on aggregation policy (4 hours, 8 hours, 12 hours, daily, and weekly) is labor intensive and time consuming task. The integration must be based on the unique value of subscriber service number (SERV_NO). As shown in Figure 4.1, all the tables have unique value (SERV_NO) except the voice incoming table which is (CALLED_NBR).

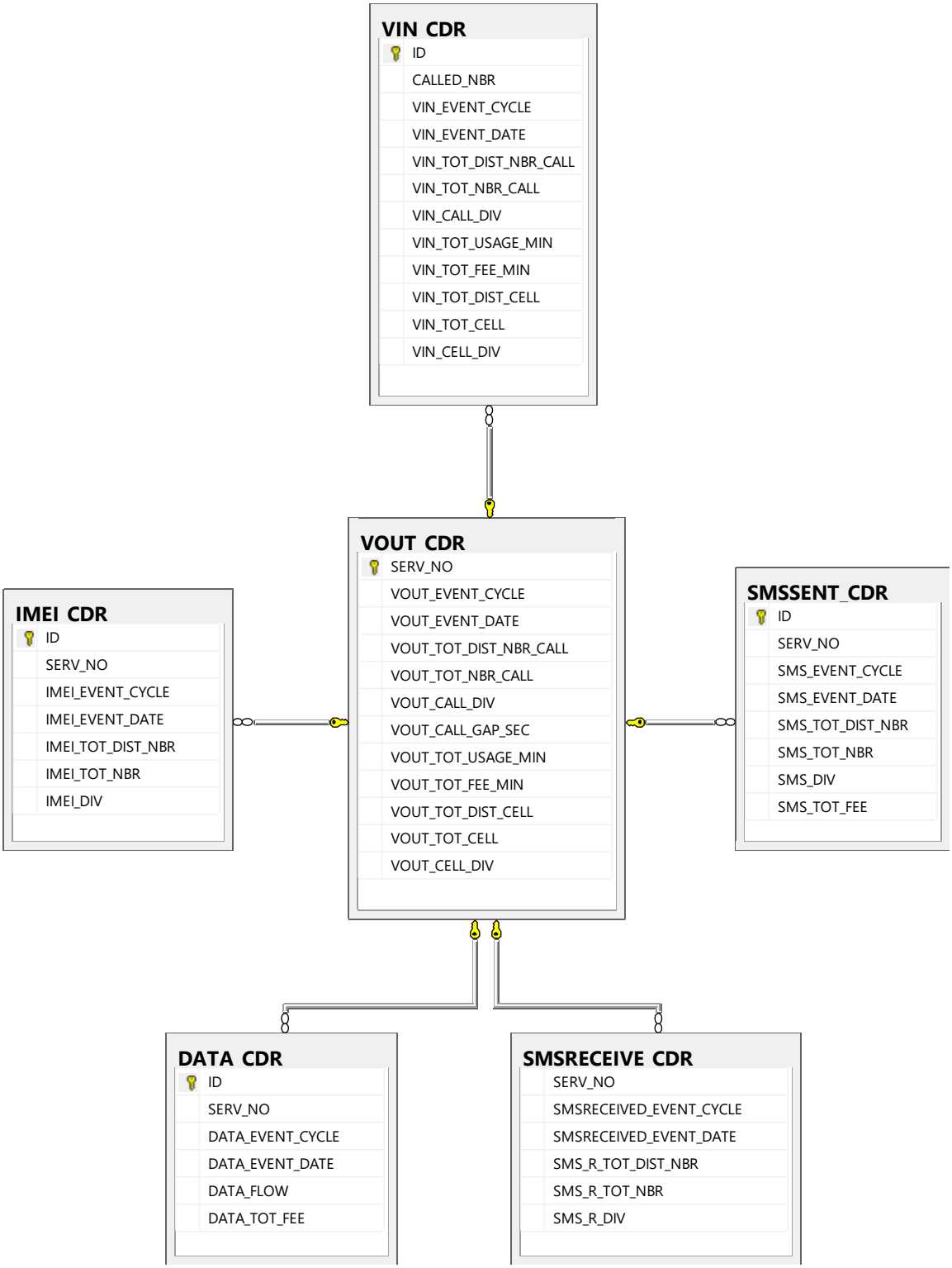

Figure 4.1: Aggregated CDR tables

4.2.5 Data Aggregation

CDR data aggregation is critical during data preprocessing which is called granularity level (time span) of data segregation. Interconnect bypass fraud can be detected based on usage behavior analysis, so that the right level of data granularity (segregation) is critical for detection success. Service usage time span is very important for subscriber's pattern analysis. Too narrow time span (granularity level) is difficult for fraud detection especially interconnect bypass fraud detection; too broad granularity level might potentially identify interconnect bypass fraud but too late to take preventive action and has high risk for the operators. So, average range of time span provides accumulated characteristics of a subscriber.

There is no fast rule for data aggregation. However, to gauge the proposed aggregation of this research work, we divided the data into five levels of granularity, these are 4 hours, 8 hours, 12 hours, daily, and weakly. In addition, each aggregation level has its own cycle as shown in Table 4.4. These techniques of aggregation give us a better potential identification of fraudulent subscriber behavior. The following steps have been taken on the sampled datasets to acquire the aggregated instances:

1) Each CDR record has associated subscriber service number, event date, and event time, based on these three attributes normalization has been done.
2) Each CDR record is aggregated based on service type like voice, SMS, data (GPRS), and device identity history logs with their respective subscriber service number.
3) Single instance can be formed per subscriber service number per granularity level (time and date).

Table 4.4: Data aggregation policy (Data scan policy)

	4 hours	8 hours	12 hours	Daily	Weekly
Cycle 1	00:00 – 03:59	00:00 – 07:59	00:00 – 11:59	00:00 – 23:59	1st week
Cycle 2	04:00 – 07:59	08:00 – 15:59	11:00 – 23:59		2nd week
Cycle 3	08:00 – 11:59	16:00 – 23:59			3rd week
Cycle 4	12:00 – 15:59				4th week
Cycle 5	16:00 – 19:59				
Cycle 6	20:00 – 23:59				

4.2.6 Feature Selection

Attribute selection from dataset is a process of selecting a set of attributes (features) to build a predictive model (classifier). This phase is conducted in the model building processes to reduce training times, to reduce algorithm computational cost, to reduce the risk of over-fitting and under-fitting, to boost model generalization and to increase features interpretability.

A collection of selected attributes containing information that are highly correlated with the class, yet uncorrelated to each other, i.e., increase class relevancy with attribute subsets and decrease attribute to attribute correlation, can enable us to have a better predictive model. For this research work, we applied CFS technique. It is a metric to evaluate the efficiency of features subset. The objective is to reduce FFC, and FCC. The criterion is defined using Pearson coefficient, which is essentially a ratio of FCC to FFC (FCC/FFC), a higher ratio indicates a better subset. It measures the worthy of a subset of features by inspection of the individual predictive capability of the classifier and along with the degree of redundancy between them.

Therefore, for this study a subset of features which is highly correlated with the class but having lower feature to feature correlation are preferred. For instance, outgoing calls diversity, outgoing calls cell diversity, incoming calls diversity, incoming calls cell diversity, SMS diversity (sent and received), data flow, and IMEI are highly correlated to the output class but less correlated to each other. Table 4.5, shows the subset containing the features and their corresponding weight taken for the experiment after applying the CFS technique. However, in Appendix A.1 shows attribute raking based on information gain technique.

Table 4.5: Attribute ranking using CFS

No.	FCC/FFC	Feature	Description
1	0.6608	VOUT_CELL_DIV	Cell diversity, it shows mobility
2	0.59	SERV_NO	Subscriber service number
3	0.5489	VOUT_TOT_DIST_NBR_CALL	Total number of distinct outgoing calls
4	0.541	VOUT_TOT_CELL	Total number of cells used for outgoing calls

No.	FCC/FFC	Feature	Description
5	0.541	VOUT_TOT_NBR_CALL	Total number of outgoing calls
6	0.2783	VOUT_CALL_DIV	Ratio of distinct calls to total calls
7	0.1423	VOUT_TOT_DIST_CELL	Total number of distinct cells
8	0.1305	SMS_DIV	Ratio of distinct to total SMS sent
9	0.1032	DATA_FLOW	Total size of data used (MB)
10	0.0741	VIN_CELL_DIV	Ratio of distinct to total cells for calls received
11	0.074	VIN_TOT_DIST_NBR_CALL	Total number of distinct calls received
12	0.0717	VIN_CALL_DIV	Ratio of distinct to total calls received
13	0.0675	SMS_TOT_DIST_NBR	Total number of distinct SMS sent
14	0.0651	VIN_TOT_DIST_CELL	Total # of distinct cells used for calls received
15	0.0556	SMS_TOT_NBR	Total number of SMS sent
16	0.0543	VIN_TOT_NBR_CALL	Total number of calls received
17	0.0543	VIN_TOT_CELL	Total number of cells used for calls received
18	0.0493	VOUT_EVENT_CYCLE	Calls event time
19	0.0477	VOUT_CALL_GAP_SEC	Two consecutive calls gap
20	0.0288	CALL_RATE	Ratio of incoming calls to outgoing calls
21	0.0273	SMS_R_TOT_DIST_NBR	Total number of distinct SMS received
22	0.0241	SMS_R_DIV	Ratio of distinct to total SMS received
23	0.0241	IMEI_DIV	Ratio of distinct to total IMEI
24	0.0153	SMS_R_TOT_NBR	Total number of SMS received
25	0.0151	IMEI_DIST_TOT_NBR	Total number of distinct IMEI used
26	0.0141	IMEI_TOT_NBR	Total number of IMEI used
27	0.0121	EVENT_CYCLE	Service event time

However, telecom service subscribers are more legitimate subscribers than fraudulent subscribers in the operational environment. Therefore, appropriate dataset sampling is needed to avoid the unbalanced high class of dataset. For our study, we agreed to take 66.6% of the dataset instances as normal and 33.3% of the dataset instances as fraud. The sample dataset size is shown in Table 4.6.

Table 4.6: Data size for ML algorithms

Aggregation	Dataset composition		
	Fraud (33.3%)	Normal (66.6%)	Total
4 hours	192,431	384,862	577,293
8 hours	114,990	229,980	344,970
12 hours	86,802	173,604	260,406
Daily	64,362	128,724	193,086
Weekly	16,586	33,172	49,758

According to Table 4.6, as we go through from 4 hours' dataset to weekly dataset the number of instances (rows) decreases. This is due to data grouping and data fetching criteria used. The grouping is done using event time, event date, and service number.
The major challenge here is determining the threshold values for fetching instances from the database for models training. Based on discussion held with domain experts and deep investigation of one-month fraudulent subscribers CDR, we carefully decided the threshold values as indicated in Figure 4.2, Figure 4.3 and A.1 - A.4 in Appendix 3 for fetching fraud subscriber dataset from the database for each of aggregation level.

These thresholds are optimum values that we decided to fetch instances for training from fraud data. If we vary the threshold values, the total number of fetching instances also varies. In addition, when we are taking those threshold values for fraud behavioral representation, we also highly consider the normal subscriber usage behavior. Before training the algorithm, the training examples must be clearly demarcated for both fraud and normal dataset. Otherwise, it will lead us to high false positive rate (blockage of legitimate subscriber). Figure 4.2, depicts a 4 hours' fraud dataset fetching code snippet and the total number of instances fetched using this rule is 192,431.

```
Algorithm 1: Four hours' fraud dataset fetching rule
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells
to total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 2
and VOUT_TOTi greater than 2
and VOUT_CALL_DIVi greater than 0.5
and VOUT_CELL_DIVi less than 0.5
and VIN_DISi less than 2
and VIN_CELL_DIVi less than 0.5
and SMS_TOT_NBRi less than 1
and DATA_FLOWi less than 1 then
 Generate a feature for a period of T
 end if
end for
```

Figure 4.2: 4 hours fraud dataset fetching rule

Similarly, the below code snippet the total number of instances fetched using this rule is 114,990 and based on the rules as presented in Figures A.1, A.2, A.3 and A.4 in the Appendix the total number of instances fetched is 86,802, 64,362 and 16,586 respectively. As show in Figure 4.3, the normal subscriber instances were fetched from another database based on the ratio of fraudulent instances fetching rule, but additional modifications were applied.

```
Algorithm 1: Four hours' normal dataset fetching rule
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells to
total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 1
and VOUT_TOTi greater than 1
and VOUT_CALL_DIVi greater than 0.6
and VOUT_CELL_DIVi less than 0.5
Generate a feature for a period of T
 end if
end for
```

Figure 4.3: 4 hours normal dataset fetching rule

4.2.7 Data Formatting

Data formatting is a process of formatting of input dataset into the formats that are suitable for the intended ML algorithm. Each attribute may have a different data type, for instance, real for decimal numeric values, integer for numeric values without a fractional part, nominal for categorical data like “yes” and “no”, string for lists of words, like “interconnect fraud”. However, in a classification problem the output variable must be nominal. Table 4.7 shows this research data format.

Table 4.7: Attributes data format

No.	Data type	Attributes
1	Integer	SERV_NO, OUT_EVENT_CYCLE, VOUT_TOT_DIST_NBR_CALL, VOUT_TOT_NBR_CALL, VOUT_TOT_DIST_CELL, VOUT_TOT_CELL, VIN_TOT_DIST_NBR_CALL, VIN_TOT_NBR_CALL, VIN_TOT_DIST_CELL, VIN_TOT_CELL, SMS_TOT_DIST_NBR, SMS_TOT_NBR, SMS_R_TOT_DIST_NBR, SMS_R_TOT_NBR, IMEI_TOT_DIST_NBR, IMEI_TOT_NBR.
2	Real	VOUT_CALL_DIV, VOUT_CALL_GAP_SEC, VOUT_CELL_DIV, VIN_CALL_DIV, VIN_CELL_DIV, CALL_RATE, SMS_DIV, SMS_R_DIV, DATA_FLOW, IMEI_DIV
3	Nominal	Fraud or Not fraud

4.2.8 Removing Outlier

The ML algorithms are sensitive to the distribution and range of attributes values in the input examples of dataset instances. Outliers in input data can twist and mislead the training process of the ML algorithms resulting in longer training time, less accurate models and ultimately poor results. Even before predictive models are prepared on training data, outliers can twist the summary distribution of attributes values.

The most common data dispersion measures are range, quartiles, interquartile range, box plots, variance and standard deviation of the data, which are useful for identifying outliers. For this study based on the recommendation of different literatures we applied interquartile range (IQR) outlier identification and removal only for our normal customers' datasets. As shown in Table 4.8, the number of removed outliers and extreme instances of normal subscribers' behavior from the aggregated data which are labeled as 'Yes' and only 'No' labeled instances where included in the training dataset.

Table 4.8: IQR statistical measure of outliers values

Dataset	Outliers	
Label	No	Yes
4 hours	874,618	201,652
8 hours	377,612	730,30
12 hours	388,746	805,36
Daily	366,410	753,98
Weekly	204,384	642,00

Chapter 5 : Interconnect Bypass Fraud Detection Model

This chapter focuses on the model building processes using the selected algorithms. It also presents the final datasets prepared as input to the chosen algorithms and evaluation of the developed models through various performance measures. The outcomes of the experiments and selection of best models corresponding to the proposed algorithms and data aggregation time are also discussed.

5.1 Model Building Process

The core tasks performed in this phase are selecting modeling techniques, building models using the five aggregated datasets (4 hours, 8 hours, 12 hours, daily and weekly) and assessing the model's performance. The well-known approaches K-fold cross validation and percent split are used for training the selected ML algorithms.

5.1.1 Modeling Methods

Four ML algorithms ANN, SVM, RF, and J48 are selected based on their generalization capability and detection performance recommended by different literatures [3, 10, 20]. The dataset was partitioned into two parts for training and testing. The experimentation was conducted based on supervised learning technique.

5.1.2 Model Building

For each selected algorithm10 models were built using the two training modes and using the five aggregated datasets to propose the best classifiers.

A. Building ANN/MLP Models

This phase focuses on building a model using the MLP algorithm.

1) Model building using 4 hours aggregated dataset

The detail performance measures of the models are shown in Table 5.1.

Table 5.1: MLP model using 4 hours aggregated dataset

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross Val	1115.1	0.998	0.998	99.8348	0.998	1.000	0.0371
Percent Split	1140.78	0.998	0.996	99.8331	0.998	1.000	0.0368

The result shows that the highest classification accuracy attained in this experiment is 99.8% with a minimum classification error (RMSE) using percent split method. However, the time it took to build the model is higher than that of 10-fold cross validation.

2) Model building using 8 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.2.

Table 5.2: MLP model using 8 hours aggregated dataset

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	675.12	0.998	0.998	99.8404	0.998	1.000	0.0427
Percent-Split	681.23	0.998	0.998	99.7666	0.998	1.000	0.0435

The result shows that the highest classification accuracy attained in this experiment is 99.8% with a minimum classification error and build time using 10-fold cross validation method.

3) Model building using 12 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.3.

Table 5.3: MLP model using 12 hours aggregated dataset

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross-Val	497.86	0.998	0.998	99.8199	0.998	1.000	0.0379
Percent-Split	529.6	0.999	0.999	99.8825	0.999	1.000	0.031

The result shows that the highest classification accuracy attained in this experiment is 99.88% with a minimum classification error using percent-split method. However, the time it took to build the model is higher than that of 10-fold cross validation.

4) Model building using daily aggregated dataset

The detail performance measures of the model are shown in Table 5.4.

Table 5.4: MLP model using daily aggregated dataset

Training Modes		Result					
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val.	374.01	0.998	0.998	99.7954	0.998	1.000	0.0412
Percent-Split	365.43	0.998	0.998	99.8218	0.998	1.000	0.0395

The result shows that the highest classification accuracy attained in this experiment is 99.8% with a minimum classification error and build time using percent-split method.

5) Model building using weekly aggregated dataset

The detail performance measures of the model are shown in Table 5.5.

Table 5.5: MLP model using weekly aggregated dataset

Training Modes	Time(S)	Result					
	Build	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val.	93.91	0.999	0.999	99.8854	0.999	1.000	0.0315
Percent-Split	96.83	0.999	0.999	99.8759	0.999	1.000	0.0332

The result shows that the highest classification accuracy attained in this experiment is 99.8% with a minimum classification error and build time using 10-fold cross validation method.

One of the challenges of designing ANN/MLP is determining the number of hidden layers and the number of neurons in these layers. As a starting point, it is to use one hidden layer, with the number of neurons equals to half the sum of the number of input and output units and then change accordingly. Based on this approach, and the model with one hidden layer and twelve nodes give the highest accuracy which is 99.80% using percent-split training mode in the 4 hours aggregated dataset and similar acceptable results are also obtained in other aggregated datasets. Thus, the proposed multilayer perceptron neural network model contains three layers. The first layer corresponds to the input values, one hidden layer having twelve nodes and output layer with two nodes which represent the two classes (fraud and normal). The details of the MLP models for different hidden layers and neurons using 4 hours aggregation datasets are show in Table 5.6.

Table 5.6: Performance comparison of different number of hidden layers using 4 hours dataset

No.	No. Hidden layer	No. Nodes	Accuracy	Build Time(s)	RMSE
1	1	12	99.80	307.41	0.0365
2	2	8	99.78	1298.06	0.0438
3	2	12	99.76	3603.89	0.0455
4	3	12	99.79	6925.02	0.0423

B. Building SVM Models

This phase focuses on model building using SVM algorithm.

1) Model building using 4 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.7.

Table 5.7: SVM model using 4 hours aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross Val.	4623	0.956	0.956	95.5832	0.956	0.947	0.2102
Percent-Split	4544.75	0.955	0.955	95.5414	0.948	0.946	0.2112

The result shows that the highest classification accuracy attained in this experiment is 95.58% with a minimum classification error using 10-fold cross validation mode. However, the time it took to build the model higher as compared to percent-split.

2) Model building using 8 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.8.

Table 5.8: SVM model using 8 hours aggregated datasets

Training Modes	Time(S)	Result					
	Build	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	528.68	0.956	0.956	95.61	0.956	0.946	0.2095
Percent-Split	597.28	0.955	0.955	95.4979	0.955	0.946	0.2122

The result shows that the highest classification accuracy attained in this experiment is 95.6% with a minimum classification error and build time using 10-fold cross validation.

3) Model building using 12 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.9.

Table 5.9: SVM model using 12 hours aggregated dataset

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	434.57	0.961	0.961	96.1049	0.961	0.952	0.1974
Percent-Split	367.19	0.960	0.960	95.9915	0.960	0.951	0.2002

The result shows that the highest classification accuracy attained in this experiment is 96.1% with a minimum classification error using 10-folder cross validation method. However, the time it took to build the model is higher as compared to percent-split.

4) Model building using daily aggregated dataset

The detail performance measures of the model are shown in Table 5.10.

Table 5.10: SVM model using daily aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	118.34	0.952	0.952	95.1772	0.951	0.937	0.2196
Percent-Split	148.88	0.951	0.950	95.0433	0.950	0.936	0.2226

The highest classification accuracy attained in this experiment is 95.17% with a minimum classification error using 10-fold cross validation method. However, the time it took to build the model is higher as compared to percent-split.

5) Model building using weekly aggregated dataset

The detail performance measures of the model are shown in Table 5.11.

Table 5.11: SVM model using weekly aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	1.97	0.982	0.982	98.151	0.981	0.974	0.136
Percent-Split	1.64	0.981	0.981	98.1261	0.981	0.974	0.1369

The result shows that, the highest classification accuracy attained in this experiment is 98.15% with a minimum classification error using 10-fold cross validation method. However, the time it took to build the model is higher as compared to percent-split.

C. Building RF Models

This phase focuses on models building using RF algorithm.

1) Model building using 4 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.12.

Table 5.12: RF model using 4 hours aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross Val.	630.92	1.000	1.000	99.9527	1.000	1.000	0.022
Percent Split	633.47	1.000	1.000	99.9536	1.000	1.000	0.0223

The result shows that the highest classification accuracy attained in this experiment is 99.95% with a minimum classification error using percent-split mode. However, the time it took to build the model is higher as compared to 10-fold cross validation.

2) Model building using 8 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.13.

Table 5.13: RF model using 8 hours aggregated dataset

Training Modes	Time(S)	Result					
	Build	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	143.69	1.000	1.000	99.9894	1.000	1.000	0.011
Percent-Split	149.08	1.000	1.000	99.9922	1.000	1.000	0.0106

The highest classification accuracy attained in this experiment is 99.99% with a minimum classification error and build time using percent-split method.

3) Model building using 12 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.14.

Table 5.14: RF model using 12 hours aggregated dataset

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	96.19	1.000	1.000	99.9881	1.000	1.000	0.0123
Percent-Split	94.99	1.000	1.000	99.9887	1.000	1.000	0.0128

The highest classification accuracy attained in this experiment is 99.98% with a minimum classification error and build time using percent-split method.

4) Model building using daily aggregated dataset

The detail performance measures of the model are shown in Table 5.15.

Table 5.15: RF model using daily aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	118.34	0.952	0.952	95.1772	0.951	0.937	0.2196
Percent-Split	148.88	0.951	0.950	95.0433	0.950	0.936	0.2226

The highest classification accuracy attained in this experiment is 95.18% with a minimum classification error and build time using 10-fold cross validation.

5) Model building using weekly aggregated dataset

The detail performance measures of the model are shown in

Table 5.16.

Table 5.16: RF model using weekly aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	8.11	1.000	1.000	99.9759	1.000	1.000	0.0188
Percent-Split	8.13	1.000	1.000	99.9764	1.000	1.000	0.02

The highest classification accuracy attained in this experiment is 99.98% with a minimum classification error using percent-split method. However, the time it took to build the model is higher as compare to 10-fold cross validation.

D. Building J48 Models

This phase focuses on models building using J48 algorithm.

1) Model building using 4 hours aggregated dataset

The detail performance measures of the model are shown in Table 5.17.

Table 5.17: J48 model using 4 hours aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	28.47	1.000	1.000	99.9525	1.000	1.000	0.0217
Percent-Split	29.75	1.000	1.000	99.9536	1.000	1.000	0.0215

The highest classification accuracy attained in this experiment is 99.95% with a minimum classification error using percent-split mode. However, the time it took to build the model is higher as compare to 10-fold cross validation.

2) Model building using 8 hours aggregated dataset

The detail performance measures of the models are shown in Table 5.18.

Table 5.18: J48 models using 8 hours aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	14.24	1.000	1.000	99.99	1.000	1.000	0.01
Percent-Split	13.54	1.000	1.000	99.9914	1.000	1.000	0.0093

The highest classification accuracy attained in this experiment is 99.99% with a minimum classification error and build time using percent-split mode.

3) Model building using 12 hours aggregated dataset

The detail performance measures of the models are shown in Table 5.19.

Table 5.19: J48 models using 12 hours aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	10.08	1.000	1.000	99.9873	1.000	1.000	0.0112
Percent-Split	9.91	1.000	1.000	99.9887	1.000	1.000	0.0106

The highest classification accuracy attained in this experiment is 99.98% with a minimum classification error and build time using percent split mode.

4) Model building using daily aggregated dataset

The detail performance measures of the models are shown in Table 5.20.

Table 5.20: J48 models using daily aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	6.42	1.000	1.000	99.9777	1.000	1.000	0.0149
Percent-Split	7.08	1.000	1.000	99.9848	1.000	1.000	0.0123

The highest classification accuracy attained in this experiment is 99.98% with a minimum classification error using percent-split mode. However, the time it took to build the model is higher as compared to 10-fold cross validation.

5) Model building using weekly aggregated dataset

The detail performance measures of the model are shown in Table 5.21.

Table 5.21: J48 model using weekly aggregated datasets

Training Modes	Result						
	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
10-Fold Cross- Val	1.14	1.000	1.000	99.9759	1.000	1.000	0.0155
Percent-Split	1.09	1.000	1.000	99.9645	1.000	1.000	0.0188

The highest classification accuracy attained in this experiment is 99.97% with a minimum classification error using 10-fold cross validation mode. However, the time it took to build the model is higher as compared to percent-split.

5.1.3 Model Evaluation and Discussion

In this Section, we present the evaluation of the results of the experiment and the statistics of the performance of the models developed using different ML algorithms and aggregated datasets. Hence, from the experimentation, the models developed using different ML algorithms have a comparison performance. However, J48 and RF algorithms attain some much better results than all the rest of the algorithms. The details of the performance of the algorithms are summarized Table 5.22.

Table 5.22: Performance comparison of algorithms based on accuracy

Algorithm	Training Methods	Training Datasets				
		4 hours	8 hours	12 hours	Daily	Weekly
MLP	10-Fold	99.83	99.84	99.82	99.8	99.89
	Percent split	99.83	99.77	99.88	99.82	99.88
SVM	10-Fold	95.58	95.61	96.10	95.17	98.15
	Percent split	95.54	95.5	95.99	95.4	98.13
RF	10-Fold	99.95	99.99	95.18	95.18	99.98
	Percent split	99.95	99.99	99.98	95.04	99.97
J48	10-Fold	99.95	99.99	99.99	99.98	99.98
	Percent split	99.95	99.99	99.98	99.98	99.96

According to Table 5.22, J48 and RF attain the highest classification accuracy level using 8 hours dataset. For telecommunication fraud detection, the algorithm that produces the highest accuracy or less false positive (FP) is preferable because it minimizes the risk of blocking legitimate subscribers that might bring customer dissatisfaction on service providers.

Table 5.23, shows the performance of RF algorithm with different dataset aggregation. We can observe that whenever the dataset aggregation level from 4 hours to weekly increases which clearly indicates subscriber's usage statistics and enable us to determine the behavior of the customer. Thus, an increasing in data aggregation level will give us a better potential identification of fraudulent customer. However, too late to act on the fraudsters will have high damage on the operator's revenue, so longer period of aggregation is not recommended rather an average range of data aggregation (time span)

is preferable. Therefore, among the models that we built using RF algorithm and different dataset, 8 hours dataset model attains the highest accuracy level than all others dataset models, whereas 12 hours dataset model also attains the second highest classification accuracy.

Table 5.23: RF models performance comparison by different datasets

No.	Dataset	Accuracy	RMSE
1	4 hours	99.9536	0.0215
2	8 hours	99.9922	0.0106
3	12 hours	99.9887	0.0106
4	Daily	99.9848	0.0123
5	weekly	99.9764	0.02

5.2 Deployment of Machine Learning Model

This phase focused on the core tasks such as plan deployment, design integration and selecting implementation programming language (C++, java, or Python). Hence, the knowledgebase we developed in the form of the model needs to be implemented and integrated with existing distributed systems across the required enterprise applications using API.

The detail design and integration with existing customer relationship management (CRM), and convergent billing system (CBS) can be show in Figure 5.1.

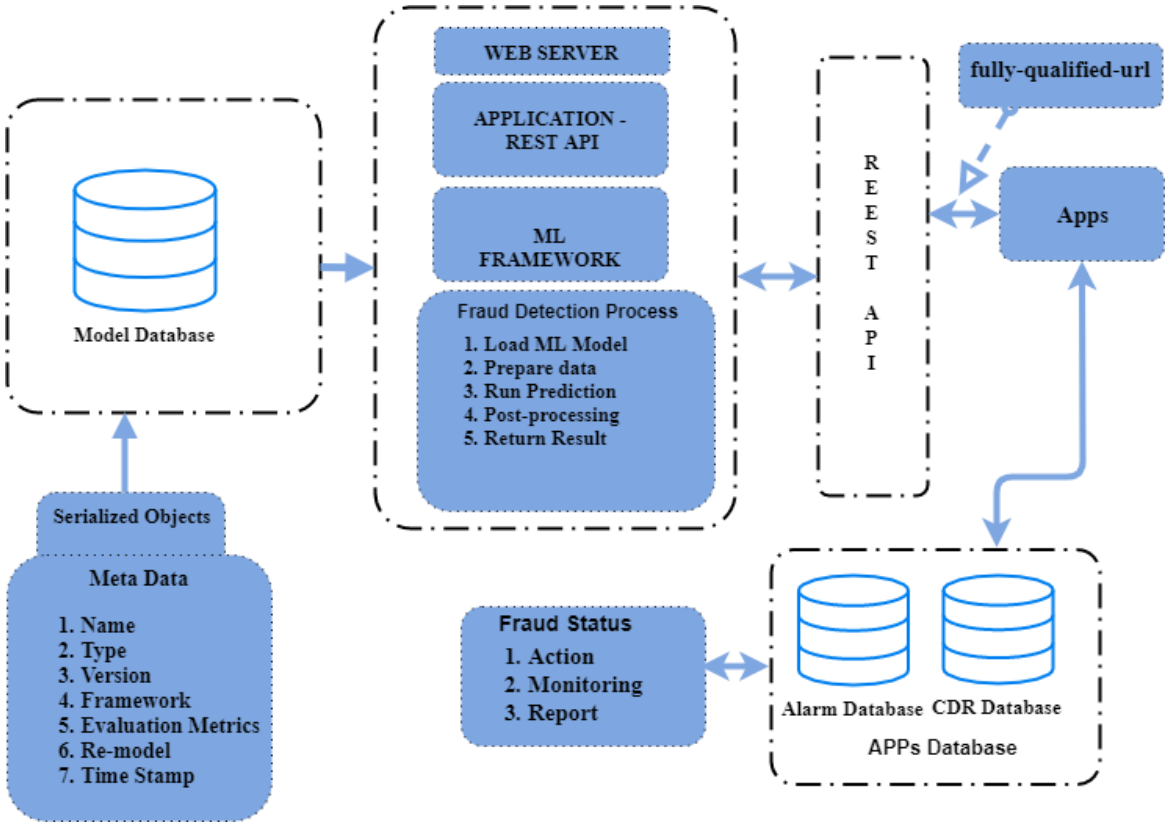

Figure 5.1:Interconnect Bypass Fraud Detection Model

Chapter 6 : Conclusion and Future Works

6.1 Conclusion

Telecommunication fraud detection has been enhanced through time. However, this improvement seems to be a continuous process, because an advancement of the technology opens a door with a loophole for fraudsters every time. In this thesis, we have tried to develop a model for interconnect bypass fraud detection that enables the operators to solve the limitation of FMS and TCG.

A sample of 30,000 bypass fraudulent service numbers were provided by the fraud management section under ethio telecom security department. These service numbers were blocked from accessing any of the communication network services. In addition, those service numbers CDRs (voice, data (GPRS), SMS and IMEI) are collected from the distributed database (oracle).

We have analyzed the subscribers CDR and identified 20 fundamental attributes that we think are best for representing the characteristics of interconnect bypass fraudsters. However, the attributes by themselves are difficult to distinguish fraudsters from legitimate subscribers, so that we derived 26 features from the existing attributes. Based on the selected attributes, we preprocessed and formatted the CDR data. We prepared five group of aggregated datasets (4 hours, 8 hours, 12 hours, daily, and weekly). We took data sampling for legitimate subscribers and fraud subscribers with a ratio of 66.6% and 33.3% respectively for training and testing.

Based on generalization capability and classification performance we selected four ML algorithms (MLP, SVM, RF and J48). Using two training modes (10-fold and percent split) and five datasets, we developed 10 models for each algorithm. Among the models, we obtained those with better detection performance and minimum RMSE classifier were selected.

From the experiments, RF and J48 scored the highest classification performance than all the other ML algorithms on 8 hours aggregated dataset. The performance accuracies of the models are 99.9922% & 99.9914. Therefore, either RF or J48 is a promising algorithm for interconnect bypass fraud detection.

6.2 Contribution

Some of the major contribution of this thesis work are: -

- It maximizes correct prediction and minimizes incorrect prediction of fraudulent subscriber at a satisfactory level.
- It shows the limitation of existing rule-based FMS and TCG.
- The thesis clearly identified which GSM identifiers (features) potentially important to be concentrated on changing values to tackle bypass fraud.
- It gives the operators a better qualitative understanding of subscriber fraudulent behavior.

6.3 Recommendation

There are a lot of fraud types in the telecom industry, doing similar study on the other types of fraud is recommended because some are enablers for interconnect bypass fraud.

Some relevant works that should also be done by the operators are as follows:

- Subscription fraud (identity theft).
- Roaming fraud.
- Mobile subscribers' complaints on service quality (QoS) and call delays should be investigated.
- Base stations (cell) with high congested traffic, high volume of prepaid recharges and customers with large number of contracts (SIM cards) should also be studied.

6.4 Future Work

The outcomes of this study have produced promising results. However, it is necessary to improve the model further to detect interconnect bypass fraud in nearly real time and decrease false alarm rate.

Some important future works that could be a continuation of our work are as follows:

- More refinement of normal subscriber data features may improve performance and accuracy of the technique, for instance including the customer demography, if the data is available. In our case, ethio telecom is not willing to provides us the data to verify whether those normal service numbers have valid customers or not. Otherwise, there might be a possibility of fraudsters within the legitimate subscribers CDR data.

- Interconnect bypass fraud can also be committed while roaming CDRs exchange between operators. Which is a collection of CDRs used in GSM to send billing records of roaming subscribers from the visited public mobile network to their respective home public network operator. Hence, those CDRs features should also be included to enhance the characterization of customer base and behavior.

References

[1] Communications Fraud Control Association and others, "2017 Global Fraud Loss Survey," Press Release, June, 2017.

[2] Sahin Merve, Francillon Aurelien, Gupta Payas and Ahamad Mustaque, "Fraud in telephony networks," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2017, pp. 235--250.

[3] Abdikarim Hussein Elmi and Roselina Sallehuddin, "Classification of SIM Box Fraud Detection Using Support Vector Machine and Artificial Neural Network," *International Journal of Innovative Computing,* vol. 4, no. 2, pp. 19-27, 2015.

[4] Zou Kaiqi, Sun Wenming, Yu Hongzhi and Liu Fengxin, "ID3 decision tree in fraud detection application," in *International Conference on Computer Science and Electronics Engineering*, Hangzhou, China, 2012.

[5] Lee Jea-Young, Lee Jong-Hyeong, Yeo Jung-Sou and Kim Jong-Joo, "A SNP harvester analysis to better detect SNPs of CCDC158 gene that are associated with carcass quality traits in Hanwoo," *Asian-Australasian journal of animal sciences,* vol. 26, no. 6, p. 766, 2013.

[6] Kou Yufeng, Lu Chang-Tien, Sirwongwattana Sirirat and Huang Yo-Ping, "Survey of fraud detection techniques," *Networking, sensing and control, 2004 IEEE international conference, IEEE,* vol. 2, p. 749–754, 2004.

[7] Akhter Mohammad Iquebal and Ahamad Mohammad Gulam, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *International Journal of Scientific & Engineering Research,* vol. 3, no. 3, 2012.

[8] Sarno Riyanarto, Dewandono Rahadian Dustrial, Ahmad Tohari, Naufal Mohammad Farid and Sinaga Fernandes, "Hybrid Association Rule Learning and Process Mining for Fraud Detection," *IAENG International Jounal of Computer Science,* no. 42 (2), pp. 59-72, 2015.

[9] Becker Richard A, Volinsky Chris and Wilks Allan R, "Fraud detection in telecommunications: History and lessons learned," *Technometrics,* vol. 52, no. 1, p. 20–33, 2010.

[10] Ighneiwa Ibrahim and Mohamed Hussamedin, "Bypass Fraud Detection:Artificial Intelligence Approach," in *arXiv e-print,* New York City, 2017.

[11] Gosset Phil and Hyland Mark, "Classification, detection and prosecution of fraud in mobile networks," *Proceedings of ACTS mobile summit, Sorrento, Italy,* 1999.

[12] Murynets Ilona, Zabarankin Michael, Jover Roger Piqueras and Panagia Adam, "Analysis and Detection of SIMbox Fraud in Mobility Networks," in *IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2014.

[13] Akhter Mohammad Iquebal and Ahamad Mohammad Gulam, "Detecting telecommunication fraud using neural networks," vol. 3, no. 3, p. 601–6, 2012.

[14] Reaves Bradley, Shernan Ethan, Bates Adam, Carter Henry, Reaves Bradley, Shernan Ethan, Bates Adam, Carter Henry and Traynor Patrick, "Boxed out: Blocking cellular interconnect bypass fraud at the network edge," in *in USENIX Security Symposium*, 2015.

[15] Elmi Abdikarim Hussein, Ibrahim Subariah and Sallehuddin Roselina, "Detecting SIM Box Fraud Using Neural Network," vol. 2015, no. IT Converg. Secur. 2012, p. 575–582, 2013.

[16] AlBougha Mhd Redwan, "Comparing Data Mining Classification Algorithms in Detection of Simbox Fraud," *Unpublished Masters Thesis, Department of Information Systems, St. Cloud State University*, 2016.

[17] Witten Ian H and Frank Eibe, Data mining : practical machine learning tools and techniques, 3rd ed., Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[18] Han Jiawei, Kamber Micheline and Pei Jian, Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann is an imprint of Elsevier, 2011.

[19] Cios Krzysztof J, Swiniarski Roman W and Pedr, "The knowledge discovery process," in *Springer*, 2007.

[20] Kahsu Hagos, "SIM-Box Fraud Detection Using Data Mining Techniques: The Case of ethio telecom," *Unpublished Masters Thesis, School of Electrical and Computer Engineering, Addis Ababa University Institute of Technology*, 2018.

[21] Williams Graham, Data mining with Rattle and R: The art of excavating data for knowledge discovery, Springer Science & Business Media, 2011.

[22] Udwan Mohammed Suleiman Mohammed and others, A Practical Approach to Design, Implementation and Management, New Delhi: Dorling Kindersley (India), 2012.

[23] Hall Mark Andrew, "Correlation-based feature selection for machine learning," in *University of Waikato Hamilton, New Zealand*, 1999.

[24] Han Jiawei, Kamber Micheline and Pei Jian, Data mining: concepts and techniques, Elsevier, 2011.

[25] Veerasamy Ravichandran, Rajak Harish, Jain Abhishek, Sivadasan Shalini, Varghese Christapher P and Agrawal Ram Kishore, "Validation of QSAR models-strategies and importance," *International journal of drug design & discovery*, vol. 3, pp. 511--519, 2011.

[26] Dua Sumeet and Du Xian, Data mining and machine learning in cybersecurity, CRC press, 2016.

[27] Sawale Gaurav J and Gupta Sunil R, "Use of artificial neural network in data mining for weather forecasting," *International Journal Of Computer Science And Applications,* vol. 6, pp. 383--387, 2013.

[28] Gaur Priyanka, "Neural networks in data mining," *International Journal of Electronics and Computer Science Engineering,* vol. 3, 2013.

[29] Alraouji Yousef and Bramantoro Arif, "International Call Fraud Detection Systems and Techniques," in *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*, Buraidah, Al Qassim, Saudi Arabia, 2014.

[30] Verma Kritika and Singh Pradeep Kumar, "An insight to soft computing based defect prediction techniques," *International Journal of Modern Education and Computer Science,* vol. 7, no. 9, p. 52, 2015.

[31] Zhang Shichao, Zhang Chengqi and Yang Qiang, "Data preparation for data mining," *Applied artificial intelligence,* vol. 17, pp. 375--381, 2003.

[32] Breiman Leo, "Random forests," *Machine learning,* vol. 45, pp. 5--32, 2001.

[33] Mishra Arvind K., Ramteke Silao V. and Sen, "Random Forest Tree Based Approach for Blast Design in Surface Mine," *Geotechnical and Geological Engineering,* vol. 36, pp. 1647--1664, 2017.

[34] Bhargava Neeraj, Sharma Girja, Bhargava Ritu and Mathuria Manish, "Decision Tree Analysis on J48 Algorithm for Data Mining," *International Journal of Advanced Research inComputer Science and Software Engineering,* vol. 3, no. 6, 2013.

[35] Rokach Lior and Maimon Oded Z, Data mining with decision trees: theory and applications, World scientific, 2014.

[36] Han Jiawei, Pei Jian and Kamber Micheline, *Data Mining: Concepts and Techniques,* 2012.

[37] Cristianini Nello, Shawe-Taylor John and others, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[38] Mohammed Mohssen, Khan Muhammad Badruddin and Bashier Eihab Bashier Mohammed, Machine learning: algorithms and applications, CRC Press, 2016.

[39] Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O and Akinjobi J, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT),* vol. 48, no. 3, pp. 128--138, 2017.

[40] Hinton Geoffrey E, Osindero Simon and Teh Yee-Whye, "A fast learning algorithm for deep belief nets," *Neural computation,* vol. 18, pp. 1527--1554, 2006.

[41] Ayodele Taiwo Oladipupo, "Types of machine learning algorithms," InTech, 2010.

[42] Kotsiantis Sotiris B, Zaharakis I and Pintelas P, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160, pp. 3--24, 2007.

Appendix 1 Dataset feature selection

Table A.1:Attribute ranking based on information gain technique

S. N.	Weight	Feature
1	0.550332	VOUT_CELL_DIV
2	0.50609	VOUT_TOT_DIST_NBR_CALL
3	0.384054	VOUT_TOT_NBR_CALL
4	0.384054	VOUT_TOT_CELL
5	0.311125	VOUT_CALL_DIV
6	0.234377	DATA_FLOW
7	0.12621	VOUT_CALL_GAP_SEC
8	0.021015	VOUT_TOT_DIST_CELL
9	0.018185	SMS_TOT_NBR
10	0.018147	SMS_TOT_DIST_NBR
11	0.01711	SMS_DIV
12	0.006556	VIN_CELL_DIV
13	0.006367	VIN_CALL_DIV
14	0.006335	VIN_TOT_CELL
15	0.006335	VIN_TOT_NBR_CALL
16	0.005932	VIN_TOT_DIST_CELL
17	0.005932	VIN_TOT_DIST_NBR_CALL
18	0.00301	VOUT_EVENT_CYCLE
19	0.001107	CALL_RATE
20	0.001107	IMEI_DIST_TOT_NBR
21	0.000622	SMS_R_TOT_DIST_NBR
22	0.00061	SMS_R_DIV
23	0.000256	SMS_R_TOT_NBR
24	0.000246	IMEI_DIV
25	0.000216	IMEI_TOT_NBR

Appendix 2 Performance measure of different layer of ANN architecture

Table A.2: Performance comparison of different number of hidden layers

No. Hidden layer	No. Nodes	Build Time(S)	Precision	Recall	Accuracy	F-Measure	ROC	RMSE
1	8	1162.07	0.998	0.998	99.8339	0.998	1.000	0.0402
1	12	307.41	0.998	0.998	99.8019	0.998	1.000	0.0365
2	8	1298.06	0.998	0.998	99.7779	0.998	1.000	0.0438
2	12	3603.89	0.998	0.998	99.7599	0.998	1.000	0.0455
3	12	6925.02	0.998	0.998	99.7959	0.998	0.999	0.0423

Appendix 3 Fraud instances fetching rules

Algorithm 1: eight hours' fraud dataset fetching rule

```
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells
to total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 4
and VOUT_TOTi greater than 4
and VOUT_CALL_DIVi greater than 0.6
and VOUT_CELL_DIVi less than 0.5
and VIN_DISi less than 2
and VIN_CELL_DIVi less than 0.5
and SMS_TOT_NBRi less than 2
and DATA_FLOWi less than 1 then
 Generate a feature for a period of T
 end if
end for
```

Figure A.1: 8 hours fraud instances fetching rule

```
Algorithm 1: twelve hours' fraud dataset fetching rule
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells
to total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 6
and VOUT_TOTi greater than 6
and VOUT_CALL_DIVi greater than 0.5
and VOUT_CELL_DIVi less than 0.4
and VIN_DISi less than 2
and VIN_CELL_DIVi less than 0.5
and SMS_TOT_NBRi less than 2
and DATA_FLOWi less than 1 then
 Generate a feature for a period of T
 end if
end for
```

Figure A.2: 12 hours fraud instances fetching rule

```
Algorithm 1: twenty-four hours' fraud dataset fetching rule
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells
to total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 8
and VOUT_TOTi greater than 8
and VOUT_CALL_DIVi greater than 0.6
and VOUT_CELL_DIVi less than 0.5
and VIN_DISi less than 2
and VIN_CELL_DIVi less than 0.5
and SMS_TOT_NBRi less than 2
and DATA_FLOWi less than 2 then
 Generate a feature for a period of T
 end if
end for
```

Figure A.3: Daily fraud instances fetching rule

```
Algorithm 1: weekly hours' fraud dataset fetching rule
T= Time interval for collecting CDRs
S= Set of instances of all know fraud in database
VOUT_DIS= Total distinct outgoing calls
VOUT_TOT= Total outgoing calls
VOUT_CALL_DIV= Ratio of total distinct outgoing calls to
total outgoing calls
VOUT_CELL_DIV= Ratio of total distinct outgoing cells to
total outgoing cells
VIN_DIS= Total distinct incoming calls
VIN_TOT= Total incoming calls
VIN_CELL_DIV= Ratio of total distinct incoming calls cells
to total incoming calls cells
for all i ∈ Si do
 if VOUT_DISi greater than 10
and VOUT_TOTi greater than 10
and VOUT_CALL_DIVi greater than 0.7
and VOUT_CELL_DIVi less than 0.6
and VIN_DISi less than 2
and VIN_CELL_DIVi less than 0.5
and SMS_TOT_NBRi less than 2
and DATA_FLOWi less than 2 then
 Generate a feature for a period of T
 end if
end for
```

Figure A.4: Weekly fraud instances fetching rule

Appendix 4 Some of the major discussion points with domain experts

Some of the major discussion points with domain experts to understand telecom fraud and gain knowledge.

1) Interviewer job function title are: -
   - ✓ Fraud expert and fraud analysts
   - ✓ Revenue assurance analysts
2) How the current fraud and revenue assurance departments interact in ethio telecom?
   - ✓ They are fully separated structure and function.
3) How many employees does ethio telecom fraud and revenue assurance departments have?
   - ✓ 20-50 people each of them department have.
4) What percentage of ethio telecom annual budget do you expect spent on fraud prevention and detection methods in the last fiscal year 2017/2018?

- ✓ Not exactly know but most estimated that 5-10%

5) What is/are the key priority in terms of fraud prevention and detection in ethio telecom now?
   - ✓ Subscription fraud (Identity theft)
   - ✓ Roaming fraud
   - ✓ Bypass fraud
6) What is/are the number one financial leakage for ethio telecom?
   - ✓ Different fraud type, the most damage fraud are subscription, bypass fraud, IRSF, and others
7) What is the most effective fraud detection and/or prevention method for ethio telecom?
   - ✓ Some staffs answered: Big data analytics
   - ✓ Some staffs answered: Machine learning
   - ✓ Some staffs answered: AI
   - ✓ Some staffs answered: training and a warning staffs on fraud and the consequences
   - ✓ Some staffs answered: Monitor PBX activities
   - ✓ Some staffs answered: Security software
8) What is the big challenge you are currently facing in your role?
   - ✓ Most staffs answered: Poor Customer acquisition
   - ✓ Most staffs answered: Technology complexity
   - ✓ Most staffs answered: New technology brings new fraud origins
   - ✓ Most staffs answered: Subscription fraud
   - ✓ Most staffs answered: Incorrect data entry
9) What is/are ethio telecom stopping from overcoming these challenges?
   - ✓ Most staffs answered: Regulation
   - ✓ Most staffs answered: Cost
   - ✓ Most staffs answered: Lack of effective solutions
   - ✓ Some staffs answered: Lack of knowledge
10) What has been done to tackle bypass fraud in ethio telecom so far?
   - ✓ Answer: - Traditional fraud management solutions like Fraud Management System (FMS) and Test call generator (TCG). However, thousands of alarms with a false positive rate will generated. Most of the fraud alerts from unusual high false

positives (i.e. legitimate high usage call alerts) so that this is one of critical drawback of traditional systems and waste of valuable operator time and resource in unnecessary reviews and undetected cases.

11) What should be done to tackle this ever-increasing telecom fraud in ethio telecom?

- ✓ Answer: - the rules-based systems serve as the primary defense firewall. However, fraud and revenue assurance departments, they are now looking for a new solution to ban those organized fraudsters like big data analysis, network link model and data mining technologies rather than relying on defined threshold rules.

Declaration

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

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____

Place and Date of Submission: Addis Ababa University. August, 2019