Near-Real Time SIM-box Fraud Detection Using Machine Learning in the case of ethio telecom

BY: FITSUM TESFAYE

ADVISER: EPHREM TESHALE (PHD)

A Thesis submitted to

School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Engineering

Addis Ababa University

Addis Ababa, Ethiopia

February 28, 2020

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices; I have fully acknowledged and referred all materials used in this thesis work.

Fitsum Tesfaye

_____ _____

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Fitsum Tesfaye**, entitled *Near-Real Time SIM-box Fraud Detection Using Machine Learning in the case of ethio telecom* and submitted in partial fulfillment of the requirements for the degree of Master of Science Telecommunication Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Internal Examiner	_____	Signature	_____	Date	____
External Examiner	_____	Signature	_____	Date	____
Adviser	Ephrem Teshale (PhD)	Signature	_____	Date	____
Co-Adviser	_____	Signature	_____	Date	____

_____

Dean, School of Electrical and Computer Engineering

ABSTRACT

The advancement of telecommunication era is rapidly growing, however, telecom fraudsters encouraged by the emerging of these new technologies. Interconnect bypass fraud is one of the most sever threats to telecom operators. Subscriber Identity Module Box (SIM-box) fraud is one of an interconnect bypass telecom fraud type and uses Voice over IP (VoIP) technology. In addition, it's difficult to detect such fraud types with Test Call Generation (TCG) and a traditional types of Fraud Management System (FMS). Both TCG and FMS easily bypassed by the fraudsters, telecom companies impacted by losing billions of dollars.

In this study, Sliding Window (SW) aggregation mode is applied to provide a relevant dataset instance and reduce detection delay to one hour by using supervised Machine Learning (ML) algorithm. Three supervised ML classifier algorithms were used, namely Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM) with the two validation techniques 10-fold cross-validation and supplied test. Call Detail Record (CDR) data were collected, relevant attributes were selected and preprocessing such as data cleaning, integrating and aggregating tasks were performed.

The experimental results depict that RF classifier using cross-validation on SW aggregation mode achieves a better classification accuracy (96.2%). ANN is placed on second with its overall performance accuracy and its detection delay, SVM algorithm using cross-validation exceeds the desired detection delay (49,965 second) with poor performance accuracy. RF classifier algorithm using SW aggregation mode overcomes the trade-off detection accuracy and detection delay.

KEYWORDS

Bypass Fraud, Machine Learning, SIM-box Fraud, Sliding window

ACKNOWLEDGMENTS

First I would like to thank God for giving me the strength to pass all the steps. Next, I would like to give special gratitude to my advisor Ephrem Teshale (PhD) for his constructive, and valuable comments and support. I would also like to thank my evaluators Yalemzewd Negash (PhD) and Murad Ridwan (PhD) for their feedbacks during the thesis progress presentations. I also want to thank my company ethio telecom for giving me this opportunity.

I would also like to give my special thanks to ethio telecom staffs' for their support on giving Data and resource, and also special thanks to my firends Gebremeskel G/medhin, Surafel G/Mariam and Tamirat Teshome for being supportive of this research work.

Lastly, I would like to give my special thanks to my beloved wife Liya Abiyu for her unforgettable support and patience, to my sweet kids too.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
CDR	Call Detail Record
CFCA	Communications Fraud Control Survey
DT	Decision Tree
DRSF	Domestic Revenue Share Fraud
F4H	Fixed 4-Hour
FMS	Fraud Management System
FN	False Negative
FP	False Positive
IQR	Inter Quartile Range
IRSF	International Revenue Share Fraud
ML	Machine Learning
PRS	Premium Rate Service
QoS	Quality of Service
RF	Random Forest
RMS	Root Mean Squared
ROC	Receiver Operating Characteristic
SIM-box	Subscriber Identity Module Box
SMS	Short Message Service

SVM	Support Vector Machine
SW	Sliding Window
TCG	Test Call Generation
TN	True Negative
TP	True Positive
UIS	United Intelligent Scoring
VoIP	Voice over IP
WEKA	Waikato Environment for Knowledge Analysis

1

INTRODUCTION

Telecommunication services' quality increases using advanced technologies. Telecom operators adopt new technologies to increase their revenues by improving their service quality to keep their customer's satisfaction. Similarly, the new emerging technologies avail a wide room for the fraudsters.

Telecommunication fraud is an activity or an intention to use any telecommunication infrastructure or services without willing to pay for the service. Globally there are different types of telecommunication frauds which brought a huge impact on the telecommunication operators. The most common telecom fraud impacts are; Revenue loss, Security and Quality of Service (QoS) reduction. Both telecommunication service provider and their customs can be impacted by those fraudulent activities. Earlier researchers categorize telecom frauds into two categories as Subscription fraud and Superimposed fraud [1], but others categorize more than five categories. Among the well-known telecom frauds categories: International Revenue Share Fraud (IRSF), Interconnected Bypass Fraud, Premium Rate Service (PRS), Domestic Revenue Share Fraud (DRSF), Device/Hardware Reselling and Wholesale Fraud are some of the top telecommunication fraud categories [2, 3].

According to the data provided by the Communications Fraud Control Survey (CFCA) [2] in 2017, telecom service providers lose around $29.2 Billion. Among the top listed telecom fraud categories, Interconnect bypass fraud places on second next to IRSF [4]. International call interconnection or call termination fees are the main interest point of interconnect bypass frauds. Fraudsters keep changing their behavior and being a challenge for telecom operators. On the other side telecom operators also work hard to overcome the impacts of telecom frauds. Recently they have been adopting data mining techniques to extract knowledge [5–7] and depict the fraudulent patterns to detect fraudulent activities. Detection techniques

need to be continuously assessed and improved so as telecom operators can cope up with the changing behavior of telecom frauds.

Telecom operators deploy different telecom fraud detection technique which is relevant to the problem.

In Ethiopia, the sole telecom company, ethio telecom, is not only losing their revenue, the country as well losses a lot in foreign currency. Telecom companies build their own FMS to protect themselves from fraudulent activities. SIM-box fraud is one of the major interconnect bypass fraud type which affects telecom operators. Telecom companies mainly use three approaches to overcome SIM-box fraud. TCG telecom operators make an international call to their own network through international get way, FMS a rule-based fraud detection technique and Controlling SIM card distribution is limiting the number of SIM card to be given for the customers.

A report made by ethio telecom's CEO on October 15, 2018, indicates that ethio telecom has lost a huge amount of foreign currency which is about 2.5 Billion Birr on 2017/18 [**Fanabc2018**][8]. Ethio telecom has given high attention and working hard to overcome these fraudsters using different ways of controlling techniques.

1.1 STATEMENT OF THE PROBLEM

An interconnect bypass fraud has a huge impact on telecom operators, especially SIM-box fraud. Interconnect bypass fraud is a reason for telecom operators to lose a huge amount of revenues. A report [2] indicates that telecom operators' estimated revenue loss is about $4.27 billion because of interconnect bypass fraud. Telecom operators must have a well-organized telecommunication fraud detection system to overcome various fraudsters' activities. The common telecom operators' detection approaches are FMS and TCG. Recently a classical ML is used to detect telecom frauds. A rule-based FMS requires an extensive list of rules and it is difficult to list out all fraudsters' behavior. Whereas TCG is an international call made by telecom operators to their own network, which is also expensive. Both FMS and TCG detection techniques are easily bypassed by fraudsters with the help of new technologies [5, 9, 10].

ML was found to be effective in the detection of telecom frauds, but there is a trade of between detection accuracy and delay in detection. Since SIM-box fraudsters are affecting telecom operators they should be detected before making huge damage on telecom operators. Overall, exploring SIM-box fraud detection technique is used easily to identify the fraudulent patterns and reduce detection delay. There is a need to investigate possibilities of quick and accurate detection techniques.

1.2 OBJECTIVE

1.2.1 *General Objective*

The main objective of the research is to detect SIM-box fraud near-real time using machine learning algorithms by analyzing users CDR data.

1.2.2 *Specific Objectives*

The specific objectives of this thesis are:

- Explore the best machine learning algorithm in the process of SIM-box fraud detection.
- Select the relevant attribute to build the SIM-box fraud detection model.
- Building models with the selected relevant algorithms for the detection of SIM-box fraud.
- Detecting SIM-box fraudsters in near-real time and explore fraudulent behavior using usage data.
- Evaluate the performance of the models.

1.3 CONTRIBUTIONS OF THE RESEARCH

To the best of the author's knowledge, there is no specific work done on SIM-box fraud detection using SW data aggregation mode by applying machine learning algorithms. The output of this research will give a better understanding of SIM-box

fraud detection near-real time and being the initial idea for further research work related to near-real time or real-time SIM-box fraud detection.

1.4 LITERATURE REVIEW

Different researches have made their studies on telecom fraud on the detection and prevention of telecom fraudsters' impact on telecom companies. SIM-box fraud is one of the top interconnect bypass telecom fraud brought a huge impact on telecom companies, a number of researches have been conducted in implementing machine learning algorithms for the detection of SIM-box fraud.

A research conducted by I. Ighneiwa et al cooperated with Tier 1 mobile operator found in Libya [4]. This research basically focused on increasing awareness on SIM-box fraud and prevent the company's revenue losses as well as denial of service, reduction of service quality and communications network congestion. The authors used CDR data for their experiment and two supervised algorithms used SVM and Decision trees (Random Forest), accuracy and precision are used as model's performance evaluation matrices. In [11] the author used SVM and ANN algorithms for the detection of SIM-box fraud using customer usage data from CDR, 234,324 call records made by 6415 subscribers. The collected CDR data is originated from one Cell-ID, 1/3 of the total number of subscribers are fraudulent subscribers and the rest are legitimate subscribers. Taking time and accuracy as performance metrics SVM achieve better performance than ANN did.

Ilona Murynets et al. [12] made the analysis and detection of SIM-box fraud using high volume traffic CDR with the consideration of users' mobility. Fraudulent users have made a huge rate of call traffic than legitimate users as well as having a static location. Which means they have none or very low mobility as compare with the legitimate users. The authors set a classifier rule as a linear combination of the three classifiers algorithm by obtaining weight coefficients from the three classifiers by minimizing the model's error on the training dataset. The final result as shown in the new classifier rule obtains a better performance accuracy compared with the other three original decision tree algorithms.

Using data mining for the detection of SIM-box fraud is being common since the big-data analysis concept emerged, [13] is a research that is conducted to detect SIM-box frauds using data mining and compare data mining techniques and classification results of the algorithm. The author selects four supervised machine learning classifiers; Logistic Classifier, Boosted Trees Classifier, SVM and ANN algorithm. Five common machine learning model evaluation metrics have been used to analyze the results, Accuracy, Confusion matrix, Area Under Curve (AUC), Precision and Recall. Both Boosted Trees classifier and Logistic Classifier models performed with better results than the other two SVM and ANN algorithms.

Due to the impact of fraudsters telecom companies should try to detect fraudulent activities in real-time. So, [14] focused on fraud detection which is algorithm based namely United Intelligent Scoring (UIS) algorithm. Kun Niu et al. believes commonly used fraud detection approaches such as a rule-based, outlier detector and classifiers have a problem with high computational cost while processing mass data in terms of accuracy. So, telecom companies need to have a real-time solution to reduce fraudulent impacts. In order to achieve that , the authors propose a new algorithm which is called United Intelligent Scoring (UIS). UIS algorithm has less computational complexity in classification time and updates areal-time scores in addition to that UIS could have the chance to detect new fraud patterns effectively.

A recent research [10] studies SIM-box fraud detection using data mining techniques in ethio telecom's cases. The author collects a one-month CDR data for 20,000 customers from ethio telecom, 5,000 of them are fraudulent customers that are detected and blocked by ethio telecom security department. The research basically focused on data mining technique for SIM-box fraud detection. RF, SVM and ANN are the selected algorithm that is applied in the research. Each algorithm's model trained and tested with different granularity level as 4-hour, one day and one month. Each algorithm achieved different classification performances. Finally, RF algorithm model with a 4-hour granularity level achieved better accuracy than the other two algorithms SVM and ANN models on the consideration of daily and monthly granularity levels. As the granularity level becomes less classifier algorithm obtain better performance or classification accuracy. Another research [15]

conducted on SIM-box fraud detection a year before [10] research is done. [15] is also used ethio telecom's customers CDR data for its experimental processes using data mining technique. Those selected algorithms have the capability to depict user's patterns form their voice, Short Message Service (SMS) and DATA usage. 12,686 CDR record used for the experiment which is very fewer datasets as compare to the CDRs generated in the company. Decision tree (J48), Rule-based (PART) and neural network (Multilayer Perceptron) algorithms implemented for the training and testing model, the model evaluated using confusion matrices, Precision, Recall, F-measure and Accuracy. The decision tree algorithm performs better than other algorithms. Both researchers [10, 15] have used Waikato Environment for Knowledge Analysis (WEKA) data mining tool for the experiment.

1.5 METHODOLOGY

The main purpose of this research is building a near-real time SIM-box fraud detection model with **SW** using machine learning. In order to achieve the objectives the following steps are carried out.

- Conduct an extensive literature review on telecom fraud more specifically related to SIM-box fraud detection.
- Continuous discussion conducted with domain experts. .
- CDR and customer profile data collected, then the required preprocessing performed and apply SW aggregation mode to make ready final instance datasets.
- WEKA workbench tool is used to train the selected machine learning algorithms and analyze the classification algorithm performance using overall accuracy, Confusion matrix, F-measure and Receiver Operating Characteristic (ROC) curve as evaluation metrics.

1.6 THESIS ORGANIZATION

This thesis research is organized as follows. Chapter 2 discusses SIM-box fraud by demonstrating scenario and depict different prevention and detection techniques of SIM-box fraud. Chapter 3 is all about a discussion on concepts of machine learning algorithms. Chapter 4 deals with the experimental analysis of this research. Tasks performed under the process of building system model; data preprocessing and feature selection, sliding window technique, algorithm training, and evaluations discussed. In Chapter 5, discusses the results obtained by the performance evaluation of the algorithms' model. conclusion and recommendations are stated under the final chapter, Chapter 6

2

SIM-BOX FRAUD

Interconnect bypass frauds is basically interested in an international call interconnection or termination fees. Even though, telecom service providers are deploying different ways of fraud detection mechanisms, fraudsters are continuously look for a hole to take any advantages using new technologies as well. Due to that, telecom operators should perform a continuous assessment to overcome any fraud activities as soon as possible. Telecom fraudsters have the capabilities to adapt to the environment by changing their behaviors. So, telecom operators need to be ready to detect fraudulent behavioral change using an efficient and effective fraud management system [16, 17].

SIM-box fraud is one of an interconnect bypass fraud type, hijacks an international voice call with the help of smart SIM-box device and fraudulent international call transients. SIM-box is a hardware device with an Ethernet port, GSM antenna to connect the cellular network and lots of SIM cards slots.

Figure 2.1: SIM-box Device [**GoogleImage2019**].

SIM-box fraudulent hijack an international voice calls and transfer the call through the internet VoIP; in some countries like USA SIM-boxers even hijack the local

call, terminate it as local call using the local SIM cards inserted in a SIM-box device [4, 7]. SIM-boxers are very interested on countries with high international call termination costs and less local calls cost [12]. SIM-boxers are working with fraudulent transient operators, these fraudulent transient operators offer a list call routing cost and hijack the calls [10]. SIM-box fraud scenario is depicted on Figure 2.2. The green line indicates a normal or legitimate way of international call route and the broken red line indicates a hijacked route. Ethio telecom is the sole telecom service provider in Ethiopia, one of the telecom service providers impacted by interconnect bypass fraud mainly SIM-box fraud.

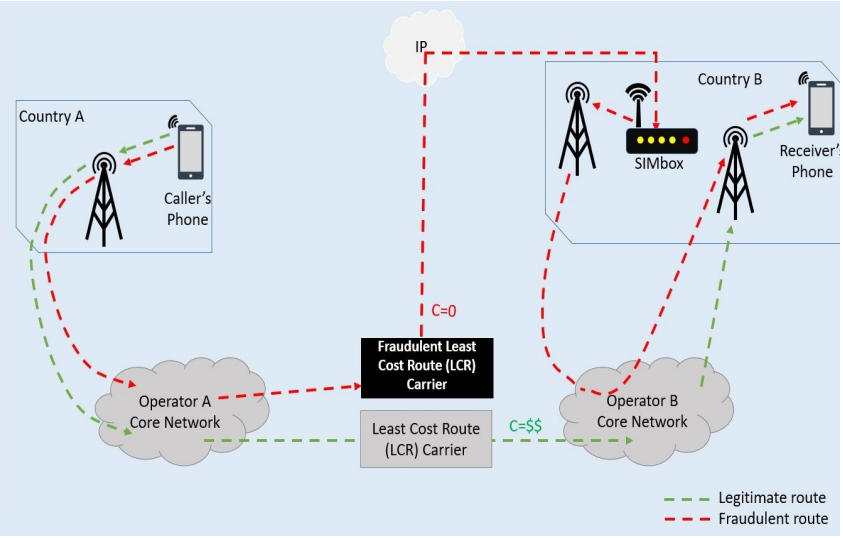


Figure 2.2: SIM-box bypass fraud hijacking of an international call [12].

Telecom operators are still battling with telecom fraudsters since the presence of telecommunication frauds. Telecom operators use different SIM-box fraud detection approaches, Test Call Generation, Fraud Management System, SIM card Distribution Control (SDC) is some of the detection techniques [4]. When telecom operators apply TCG, they make a huge number of international calls to their own network and check the calls terminate through legitimate route or SIM-box routes. In TCG there is no false positive on the test result and depends on the probabilistic nature, it is also costly making several international calls. FMS is a user profiling mated using CDR, it tries to detect the behavior of the fraudulent activity by providing a long list of rules to identify between the legitimate and fraudulent users. SDC is one way of controlling mechanism fraudsters not to get an exces-

sive number of SIM cards, like limiting the number of SIM-card per customer and demanding different customer identification information for SIM provisioning.

SIM-boxers are fighting back telecom operator's detection techniques using improved technologies not to get blocked. They analyze the incoming voice call pattern to determine whether the calls are from the real subscribers or from TCG, once they identify the call is coming from TCG either they block the test call or reroute to the legitimate route [4, 10]. SIM-box fraudsters act like legitimate users' behaviors, they imitate legitimate behaviors using special software Human Behavior Simulation (HBS) installed in the SIM-box device. Using the HBS software they do SIM migration and rotation acting like they are in mobility, they try to use other network services like SMS and GPRS to be more like a legitimate customer, and they prepare their own family lists and make a call one another not to look suspicious.

3

MACHINE LEARNING ALGORITHMS

3.1 INTRODUCTION

Machine learning is a computer algorithm that uses certain instructions and rules to understand important concepts of information and services from enormous amounts of input data, those rules are not created by computer programmers [18]. Machines initially intended to do their tasks much faster with a higher level of precision as compared to humans; and made human life easy and smooth. Machine learning algorithms learn from experience sample data to extract knowledge or information without step-by-step instruction. Machine learning is a part of Artificial Intelligence (AI) which helps to extract knowledge patterns from the input huge data. Machine learning learns and improves a given problem based on their experiences [19, 20]. The basic process of ML is train and test the model to generate a new set of rules based on inference from source data [18]. ML is more related to Knowledge Discovery from Data (KDD), data mining and pattern recognition. ML uses different mathematical formulation to extract information from the prior or history data which are called machine learning algorithms. Machine learning algorithms are organized based on the desired output of the algorithm. There are four common machine learning algorithms types [20, 21]; Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning.

Machine learning in the telecommunication sector has a huge contribution, prediction on either business losses or profits, telecom fraud detection, prediction on customer churn status to make sure the satisfaction of their customers [3, 19, 22]. Machine learning is helping to classify a set of instances using labeled source data as training or learning, in addition to classification, MLs also work on clustering and numerical prediction. This research focuses on machine learning algorithms for classification tasks. Algorithms that are used for classification task are super-

vised machine learning algorithms, these algorithms use labeled dataset at the time of training then build a classification model.

3.2 SUPERVISED LEARNING

Supervised machine learning used labeled data as a training dataset inferring a function or a model and models can classify or predict class labels for new datasets/instances. Supervised learning algorithm has two major groups which are Classification and Regression. Algorithms that are listed under supervised learning are; Decision Tree, Rule-Based Classifier, Naive Bayesian Classifier, The K-Nearest Neighbors Classifier, Neural Network, Linear Discriminant Analysis and Support Vector Machine [20]. Three selected algorithms for this research are discussed in detail in the coming sections.

3.2.1 Decision Tree

Decision Tree is a suppervised ML algorithm. Decision Tree (DT) uses a hierarchical and statistical models to classify the datasets into a different classes and represent the flowchart. DT takes the experiment data to select a root node attribute using a statistical measure, DT will go through to the remaining attributes recursively until no attributes remain the end nodes are called leaf-node. While constructing the algorithm's tree structure, DT uses entropy and information to gain statistical measures in building the tree. ID3, C4.5, and CART are the common decision tree algorithms. Algorithms use Equation (3.1) to calculate the entropy of the attributes after splitting datasets. The final goal of using a decision tree is to create a model that predicts the values of the target variable by learning simple decision rules inferred from the data features.

$$H(s) = \sum_{c \in C} -P(c) \log_2 P(c) \tag{3.1}$$

Where,	S	The current dataset for which entropy is being calculated
	c	Set of classification in S C = {Yes, No}
	p(c)	The probability of c

The other basic concept to select a root node and decision node is an information gain. Attributes with the highest information gain value is set as a root node and the process continuously computes for the remaining attributes to develop the hierarchy of the tree. Information gain is computed taking entropy results as one component using the next Equation (3.2).

$$IG(A, S) = H(S) - \sum_{t \in T} -P(t) \bullet H(t) \tag{3.2}$$

Where,	H(s)	Entropy of the attributes
	T	Subset created from splitting set S by attribute A

3.2.1.1 Random Forest (RF)

Random Forest is a collection of a decision tree that develops to overcome DT's overfitting problem. As the name indicates the trees are built by randomly selected m number of attributes in each node of the tree [23, 24]. Take K number of trees to build the forest, select N number of attributes randomly from the total M number of attributes/features then build the first tree, this step repeated K times to get the desire K number of forests. Each tree will have an equal size of randomly selected instances. Each individual tree works with the same fashion that any decision tree work, root node and decision node selection all computation to constructing the tree. A random forest machine learning result depends on the strength of the individual tree of the forest and correlation of any tree in the forest, each individual tree of the forest gave their own decision and the most voted is taking as the final result of the algorithm.

3.2.2 *Support Vector Machine (SVM)*

SVM is a supervised machine learning algorithm and working for classification and regression. SVM is capable to analyzes and recognizes patterns of a given input dataset. SVM is trying to maximize the margin between the two different classes to create the possible largest distance between two class. SVM is using a hyperplane to isolate the given instances (which is called a vector) into two different classes [25]. Although, SVM uses a hyperplane to separate two-dimensional features, it also uses a kernel trick for dataset with high-dimensional features. Kernel trick is used to separate a high-dimensional features input and mapped in to a high-dimensional feature space [20, 22]. Linearly discriminant function Equation (3.3) computes the maximum distance between those support vectors and hyperplane [26].

$$S(x) = w^T x + b \tag{3.3}$$

Where, S(x) Linearly discriminant function

x Feature vector (input vector)

w Adjustable weight vector to control direction of the hyperplane

b Bias which control the hyperplane position.

SVM varies input's weight and combine the bias values on the training stage to separate each class, putting instances of class one (C_1) to one-side of the hyperplane and instance of class 2 (C_2) to the other side of the hyperplane. using Equation (3.4) and Equation (3.5) set the class of the new dataset instances, according to the result of Y(x) the class is identified as C_1 if $Y(x) > 0$ and C_2 if $Y(x) < 0$ [27].

$$Y(x) = w^T x + b > 0 \tag{3.4}$$

$$Y(x) = w^T x + b < 0 \tag{3.5}$$

The above scenario is not working for nonlinear separable datasets, in order to overcome the case SVM apply a technique called kernel trick as shown in Equation (3.6) generating a smooth separating nonlinear decision boundary. Using training datasets SVM built a model that assigns new example into one category or the other [20].

$$K\left(x_i, x_j\right) = e^{-\left(\|x_i - x_j\|^2/\sigma^2\right)} \tag{3.6}$$

3.2.3 *Artificial Neural Network (ANN)*

ANN is a mathematical model that used simulated interconnected neurons generating a pattern from input dataset to predict the output, artificial neural network idea is inspired by how biological neural system work [20]. There are three types of Neural Networks; Feed Forward Neural Network, Recurrent Neural Network, and Self-Organizing Map. Each NN has three parts in their network; input layer, hidden layer, and output layer. Input layer can take more than one inputs and a bias, each input has its own weight values that can be adjusted at the time of training in order to fit expected output. Each layer's node is directly connected to the next layer nodes.

The simplest kind of ANNs is a Perceptron that is used to classify linearly separable classes of any m-dimensional data, weighted sum is computed using Equation (3.7) and provided to an acctivation function like sigmoid function callculated in Equation (3.8).

$$y = \sum_{j=1}^{d} w_j x_j + w_0 \tag{3.7}$$

Where, w_0 Bias value

w_j Inputs' weight

x_j Adjustable weight vector to control direction of the hyperplane

$$sigmoid\,(a) = \frac{1}{1 + exp\,[-w^T x]} \tag{3.8}$$

The hyperplane is assigned $w^T x$ taking as the threshold function, classes labeled as class C_1 if $s(w^T x) > 0$ otherwise labeled class C_2. Most of the time Multilayer perceptron (MLP) implement for a nonlinear discriminant datasets, MLP is an ANN algorithm that has more than one hidden layer between the input and output layer. To minimize the prediction error backpropagation algorithm is used for Feedforward Neural Network. The error is computed at the output layer and the error is propagated back to adjust the weights of each inputs.

3.3 UNSUPERVISED LEARNING

Unlike supervised algorithm, an unsupervised machine learning algorithm is not using labeled datasets for model training, the result is difficult to evaluate [10]. The main purpose of the unsupervised machine learning algorithm is performed statistics called density estimation; One of the most popular density estimation methods is clustering, which tries to find clusters or groups of input [27]. K-means, Gaussian mixture model, Hidden Markov model, and PCA in the context of dimensionality reductions are the algorithms found under unsupervised machine earning [20].

3.4 SEMI-SUPERVISED LEARNING

The collection of data that combines both labeled data and unlabeled data. Most of the cases in semi-supervised machine learning labeled data are scarce, the target is

to train the model to predict classes of test data better than that of model generate using labeled data [20].

3.5 REINFORCEMENT LEARNING

One way of machine learning algorithm to gather an observation from environment interaction to take action for maximizing rewards and minimalize risk [20]. Steps that reinforcement learning goes through is

- Input state is observed by the agent
- Decision-making function is used to make the agent perform and action
- After the action is performed, the agent receiver reward or reinforcement from the environment
- The state-action pair information about the reward is stored

4

EXPERIMENTAL ANALYSIS

This chapter discusses the overall experimental process conducted through this research. Figure 4.1 shows the expermental process of the model; Data collection, Data Pre-processing, and Classification are the main tasks that have been done in order to detect SIM-box fraud. Details of the tasks done under these modules are described on the comming sections.

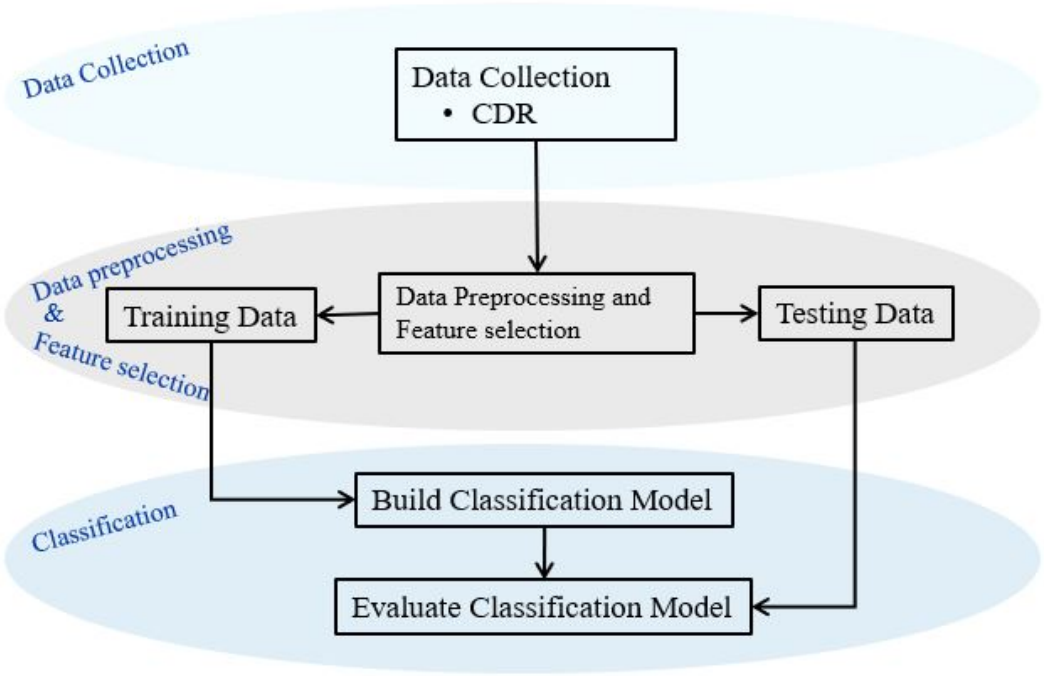


Figure 4.1: Steps of overall experimental process [3].

4.1 DATA COLLECTION

The current FMS at ethio telecom uses customers' CDR to analyse and detect telecom frauds, this research as well uses the same CDR source for the experiment. Raw CDR data is stored to our database server every five minutes, on average about 26 million CDR records are dump to the database server every day. Since the

CDR data size is huge, a separate storage place is required. Information System Division (ISD) prepares windows server 2012R2 with 8GB ram and 4TB storage capacity.

The raw CDR is stored on the storage server as a flat-file. So, it needs to be imported to the database which is installed on the same server. To do so, an automatic data loader script is used to fully import it to the database. Importing one day's CDR data flat-file takes more than 20 hours on average, so it takes much time. While collecting the CDR data, previously collected CDR data is imported to the database server. To speed up the data importing process automatic data loader scripts were running parallelly to import the CDR data to four different tables in parallel.

4.2 UNDERSTANDING THE DATA

The imported CDR data has 33 columns or attributes listed in Table 4.1 with their description. Some fields like CALLING_IMEI, CALLING_CARRIER, CALLED_CARRIER, CALLED_DISTRICT, HOTLINEINDICATOR, CALLING_TRUNK_ID, and CALLED_TRUNK_ID have no values. There are some fields that are generated for billing purpose, like CDR_ID, RE_ID, CDR_TYPE, CALL_FEE, STATUS_DATE, CHARGE_1, CHARGE_2, RATE_ID, and ACCOUNT_ITEM_ID. The remaining fields also used for biling purpose but they got their own values. CDR_ID uniquely identifies each CDR and RE_ID used to differentiate the service types Voice, SMS and Internet Data usage records. CDR_TYPE included for distinguishing Mobile Originating, Terminating or Forwarding call types. CELL_A and CELL_B as well include the ID of Calling and Called district or Cell. Some sensitive fields such as called_number, calling_number or Billing numbers have been hashed for privacy reasons [10].

The RE_ID has a value between 1 and 6, each value represents deferent types of services' CDR record. Like, 1 represents the voice service, 2 represents SMS service and 5 also represents the DATA service records. To simplify the experiment, the three services which are Voice, SMS and DATA CDR data segregated into different tables named VOICE_SOURCE_TABLE, SMS_SOURCE_TABLE and DATA_SOURCE_TABLE. In addition, customer profile data (activation date) also

collected only for currently active customers. While doing the segregation, fields with no value are removed.

Table 4.1: Row CDR Attribute With Their Description

No	Attributes	Description
1	CDR_ID	CDR Sequence Number
2	RE_ID	Service Identifier
3	BILLING_NBR	Billing Number
4	CDR_TYPE	Call type Id
5	CALLING_NUMBER	Calling Number(call initiate number)
6	CALLED_NUMBER	call destination number
7	CALLING_IMEI	International mobile equipment identity
8	CALLING_IMSI	IMSI of the calling party
9	THE_THIRD_PARTY_NUMBER	Third Party Number
10	CALL_START_TIME	the time when call start
11	CALL_END_TIME	the time when call end
12	CALL_DURATION	Call duration
13	CALL_FEE	the actual money deducted
14	CALLED_COUNTRY	called number country code of
15	CALLING_CARRIER	Calling carrier
16	CALLED_CARRIER	Called carrier
17	CALLING_DISTRICT	Cell ID of the calling party
18	CALLED_DISTRICT	Cell ID of the called party
19	STATUS_DATE	Billing date
20	CALLING_SUB_ID	Calling subscriber ID
21	BILLING_CYCLE_ID	Billing cycle ID
22	CHARGE_1	Charge amount of you spend
23	CHARGE_2	Charge amount of you get disscount

Continued on next page

Table 4.1 – continued from previous page

No	Attributes	Description
24	RATE_ID1	Rate ID
25	ACCOUNT_ITEM_ID1	Account item ID
26	UPLOAD_TRAFFIC	Upload tra fficc
27	DOWNLOAD_TRAFFIC	Download traffic
28	BILLING_OFFERING_ID	Billing offering ID
29	ERROR_CDT_TYPE	Error CDR Indicator
30	CALLFORWARDINDICATOR	Call Forward Indicator
31	HOTLINEINDICATOR	Hot Line Indicator (voice mail)
32	CALLING_TRUNK_ID	Calling Trunk ID
33	CALLED_TRUNK_ID	Called Trunk ID

4.3 DATA PREPROCESSING AND FEATURE SELECTION

ML pattern creation using previously stored or history data is properly described in Chapter 3. According to different literature works, the behavior of SIM-box fraudsters can be defined using their usage. These usage data comes from CDR data sources. Even though CDR has 33 attributes, all of the fields are not useful for the detection of SIM-box fraud process, nine (9) basic fields from the total of 33 attribute listed in Table 4.1 and one field from customer profile is selected for this research experiment. CDR fields with better information to describe about fraudulent user are listed in many research areas [10], this research as well select suitable fields for SIM-box fraud problem domain. Due to that, fileds listed in Table 4.2 are selected.

Table 4.2: Initial Selected Fields

No	Field	Description	Data Source
1	Activation_date	Service starting date	Customer Profile DB
2	Call Duration	Total time from call start to end	Billing CDR
3	Call End Time	Date and time where call is ended	
4	Call Start Time	Date and time where call is started	
5	Called Number	Call receiving subscriber number	
6	Calling Number	Call originating subscriber number	
7	Cell_A	The caller cell ID	
8	SMS	Number of text message sent	
9	Traffic_down	Amount of Downloaded traffic	
10	Traffic_up	Amount of Uploaded traffic	

4.3.1 *Sample Selection*

Before proceeding to the data preprocessing stage, sampling is part of the prior mandatory task to be done. ML is applied/used to detect SIM-box fraud, ML uses a supervised data type which is a labeled data with two class. The data type is either nominal or numeric.

Table 4.3: Class Label

Subscriber Type	**Class Label**
Fraudulent Customer	Yes
Legitimate Custome	No

Fraudulent customer numbers are provided from ethio telecom security department, there are a lot of SIM-box fraudulent numbers detected every day. The company provides 5,000 SIM-box fraudulent numbers that are detected and suspended by FMS within the CDR data collection period. The ratio of fraudulent and legitimate numbers must be proportional. Most researches proposed the ratio to be 25% fraudulent numbers and 75% legitimate numbers. Currently, ethio telecom's active customers are 34 million, which is a huge number when it is compared with the sample fraudulent numbers. So, 15,000 legitimate numbers are chosen randomly from an active customer database. Each active customer has equal probabilities to be selected, then Simple Random Sampling (SRS) is applied to get the 75% legitimate sample numbers. The research is going to use a total

of 20,000 sample subscribers number. Table 4.4 shows the number of subscriber sample size of legitimate and fraudulent service numbers.

Table 4.4: Subscriber Sample Size

Subscriber Type	No of Sample	No of Total Records	Class Label
Fraudulent Customer	5,000	134,992	Yes
Legitimate Customer	15,000	1,778.36	No

SIM-box fraud detection using machine learning or states of the art broadly uses customer usage data or CDR , which is highly helpful to extract knowledge about customer behavior.

4.3.2 *Preprocessing Data*

Real-world databases are susceptible to noise, missing values, and data inconsistency due to huge data size. These data sources are multiple and heterogeneous which needs to be properly processed to improve the data quality [28]. There are a few data processing techniques that help to improve data quality.

- **Data cleaning** is applying to remove the noise and correct inconsistencies in data.
- **Data integration** as well, useful to merge data from multiple sources into coherent data.
- **Data reduction** is used to reduce the instance data size by aggregating, eliminating redundant features or clustering.
- **Data transformations** is normalizing the data which scaled to fall within a smaller range. These techniques are working together to improve their results.

4.3.2.1 *Data Cleaning*

Data cleaning is a step of preparing a relevant data source for ML. Remove or fill the vacant values of the data, maintain the consistency of the data, remove noisy data values and remove redundant values by keeping a single record, not to

bias the ML. The data cleaning process is time taking and requires high attention not to to avoid creation of irrelevant data at the end. Since the collected data are stored in different tables or places, all the data cleaning activities applied to all data sources. Making sure the same columns found in each table must have equal size, data taype, and the same format. Like Calling numbers, Called Number, Call Start Time and Call End Time found in more than two places or tables.

4.3.2.2 *Data Aggregation*

Data aggregation is one type of data preprocessing task performed on the collected CDR data, which helps to give the full information of the specific users. A single record found under the raw CDR shows activities of a specific user; but, it is difficult to understand users' behavior with a single CDR record. A collective CDR record needs to be aggregate together in order to give a full picture of users' behaviour. An aggregation is the cumulative result of each individual user within a given time span. The time span of the aggregation is depending on the behavior of the research.This research is all about SIM-box fraud detection near-real time using usage data, and understand users' usage behavior patterns within the given time span. This research uses the SW aggregation technique to merge out the instances for the experiment, detail description about SW stated in Section 4.3.3. Finding the minimum granularity level of instance needs to consider some points fraudulent activities. Points that are listed bellow tries to depict customer behavior.

- The number of VOICE call made by the user within the time span
- The number of SMS sent within the time span
- The amount of DATA usage within the time span
- How long the user last using the service
- How frequent uses all the services

When aggregation time span is less, an unsuitable pattern may not be able to detect the SIM-box fraud. On the other hand, taking a high aggregation time span

could not be convenient to detect SIM-box fraud with near-real time. A research [10] is conducted using ethio telecom's CDR data to detect SIM-box fraud. It uses three different granularity levels which are 4-hour, 1-day and 1-month. As the cumulative results of the research shows that, the minimum granularity level (4-hour) achieves better performance compared with the other granularity levels. In addition to that the current ethio telecom's FMS also uses 4-hour granularity level as a minimum aggregation time span. So, since this research is near-real time, it would be more persuasive setting the minimum aggregation time span or granularity level to 4-hour.

4.3.2.3 *Data Integration*

Applying machine learning algorithm would be the final process. But,shown in Figure 4.2, the following list of tables created to manipulate the collected CDR.

- Aggregated_VOICE table
- Aggregated_SMS table
- Aggregated_DATA table
- Aggregated_IN_VOICE and
- Service_AGE

In order to identify and collect the aggregated values from each table, a unique identifier is required for each aggregation time range and FLAGE label is used to do so. If we take one service number's instance record, it is the integration of all the above-listed tables' records with the same FLAGE value or aggregation hour. While collecting the values of each service number's record from those tables each service number's record must much the calling time, this means the FLAGE and date of calling time must be the same. For those records which has no value at that given time, zero value will be assigned to indicate that specific service not used by the user on that specific time range. All instances are prepared using the same fashion.

Service_AGE table that has the information about service activation date of all service numbers taking as a sample for this research, which helps to get to know how

long the customer uses the service number. Keep in mind most of the time fraudulent numbers service life is too short compared with the legitimate customers. The other four tables provide the total number of a given service type used by customers, which provides lots of information about customers' behavior. Figure 4.2 shows the sample screenshots of final aggregated instances implemented for both SW and F4H.

MSISDN	TOTAL_OUT	TOTAL_DIS_OUT	TOTAL_CELL	TOTAL_DIS_CELL	TOTAL_DURATION	CALL_GAP	RATIO_DIS_OUT	RATIO_DISCELL_OUT
25198	1	1	1	1	90	239	1	1
25199	5	1	5	2	240	47	0.2	0.4
25195	2	2	2	1	360	117	1	0.5
25192	1	1	1	1	30	240	1	1
25190	2	2	2	2	60	120	1	1
25196	2	2	2	1	570	115	1	0.5
25197	1	1	1	1	180	237	1	1
25190	4	2	4	1	180	59	0.5	0.25
25194	6	4	6	1	1620	36	0.67	0.17
25194	1	1	1	1	30	240	1	1
25196	1	1	1	1	30	240	1	1
25196	3	1	3	3	90	80	0.33	1
25198	1	1	1	1	600	230	1	1
25194	1	1	1	1	1080	222	1	1
25196	1	1	1	1	60	239	1	1
25194	2	1	2	2	180	119	0.5	1
25195	5	2	3	1	510	46	0.4	0.2
25195	1	1	0	0	60	239	1	0
25191	2	1	2	2	420	117	0.5	1

Figure 4.2: Final Integrated Instance Sample.

4.3.3 *Data Aggregation Mode*

This research applies two aggregation technique.

- Sliding Window (SW)
- Fixed four-hours (F4H)

4.3.3.1 *Sliding Window (SW)*

As Section 4.3.2.2 describes the minimum time span or granularity level for aggregation is four hours. The window size equals to the minimum time span, then slides one hour to the next timeline. It will continue until the end of the collected CDR data.

As the aim of this research is to detect SIM-box fraudulent in near-real time, SW is more supportive to make it possible. The minimum granularity level of current detection system is 4-hour. So, near-real time is relative to the current detection

system granularity level. This research mainly focused to minimize the detection delay to 1-hour from 4-hour with a better detection accuracy. SW is the basic idea to achieve this research's main target. The scenario for the SW will be explained next. As shown in the Figure 4.3; let us say, the current day is the initial date for the detection process is started and goes to the next window.

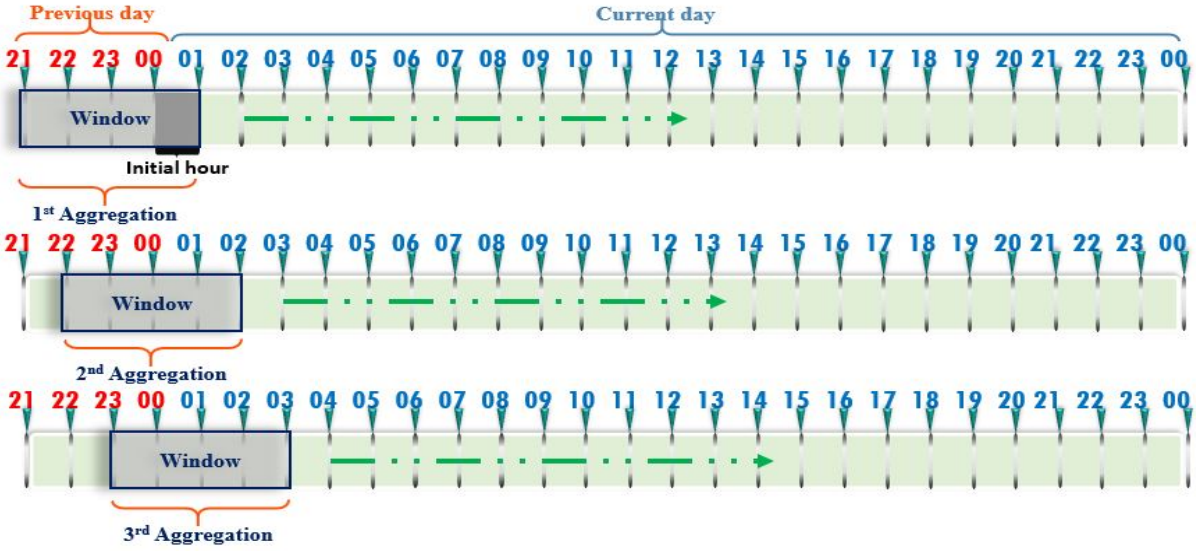


Figure 4.3: Sliding Window Scenario.

The collected CDR data within the initial hour aggregated with the previous three hours of collected CDR to get the cumulative instance result of windows time span. On the next slide, the same scenario is processed and get the instance. Keep doing the process till the collected CDR data is fully covered with the SW. Every

time window slides aggregated result will be stored into the database and the final aggregated result collected. It would be possible to get customers' behavior within one hour by considering the previous three hours using the SW technique without waiting for an additional four hours.

4.3.3.2 *Fixed Four-Hour (F4H)*

Other than SW, F4H is applied in this research as a comparison of perivious research made by Kahsu Hagos in [10] and the current FMS as well taking a CDR of every four hour. The aggregation time range is minimum of 4-hours as the name indicates. Every four hour's collected CDR is aggregated for those service numbers who uses the telecom services within that 4 hours. The daily 24 hours chunked into six parts with a time frame of 4-hours. These way of CDR data aggregation as well is applied in [10].

4.3.4 *Outliers detection and removal*

An outlier is a value that is not consistent with the remaining dataset and, also considered as noisy data. These values in a dataset need to be detected and removed to enhance the classification performance of algorithms [3, 29]. Inter Quartile Range (IQR) method is applied for the detection of outliers in this research. Outliers are individual values that fall outside of the overall pattern of the rest of the datasets. The first thing that the IQR does is sorting all the dataset and divided into four equal parts. Then, find out the three quartile values which are Q1, Q2, and Q3. The Q2 value is almost the same as the median. Since the outliers that are found somewhere outside on a specific boundary, IQR tries to find the upper and lower boundary which is basically called the fence. IQR value is calculated as shown on Equation 4.1.

$$IQR = Q_3 - Q_1 \tag{4.1}$$

The boundaries are calculated using an outlier's factors which is basically set to '1.5', Equation 4.2 and Equation 4.3 shows the upper and lower boundary respectively.

$$Upper\_Limit = Q1 + (1.5 * IQR) \tag{4.2}$$

$$Lower\_Limit = Q1 - (1.5 * IQR) \tag{4.3}$$

Pictorial description about IQR is shown in Figure 4.4

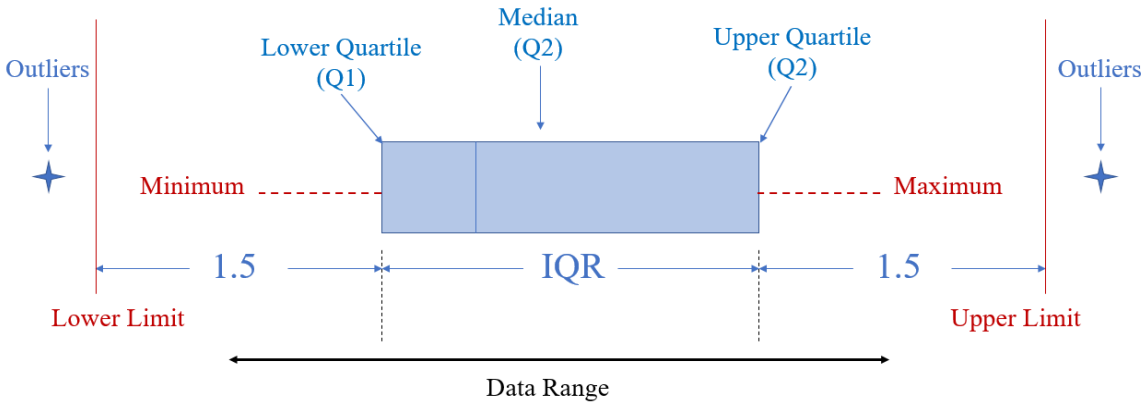


Figure 4.4: Pictorial Presentation of IQR [3].

Once the process is finalized, the outliers are detected from both SW and Fixed Four-hours then removed. Table 4.5 shows the number of outliers that are detected and removed.

Table 4.5: Number of outliers per aggregation mode

Aggregation Mode	Initial Datasets	Outliers	Extreme Values	Cleaned Datasets
SW	700,816	41,347	55,167	604,302
F4H	174,568	10,399	13,901	150,268

4.4 ALGORITHM TRAINING

Once all the data preprocessing, feature selection, aggregation, and integration is completed, the next step is performing training and building classification model using the selected algorithm. Machine learning classification technique is discussed in detail on Chapter 3. The selected three supervised machine learning algorithms Random Forest, Artificial Neural Network and Support Vector Machine are to create the models.

For the training of proposed ML algorithms, two separate training techniques are used to train the models. Explicitly, K-fold cross-validation and Separate test data.

K-fold cross validation is the most used training method. What it does is, chunk the instance dataset into k-equal parts or folds, then the classifier algorithm trained using k-1 folds and tested by the remaining fold. This process is repeated iteratively by changing the test fold starting the first up to K^{th} fold. Finally, the cumulative average error of each training and testing result is provided [10, 28]. 10-fold cross-validation technique is used for medium size datasets, so this research as well uses the technique.

Separate test data is also another way of model training technique. The datasets divided into two parts, one for training and the other is for testing purposes. There is no a specific labeled dataset ratio of testing and traning data, but, the divided datasets should be enough for both training and testing. If there are enough datasets, it is also possible to split the dataset 50% to 50%. Unless the training and testing process biased by insufficient datasets [3]. Table 4.6 depicts that separate test data uses 40% of the total dataset for algorithm testing purpose and Table 4.7 containes 60% of the total dataset for algorithm training.

Table 4.6: Testing Dataset (40% of Total Dataset)

Aggregation Mode	Normal	Fraud	Total Dataset
SW	186,225	55,497	241,722
F4H	46,390	13,718	60,108

Table 4.7: Training Dataset (60% of Total Dataset)

Aggregation Mode	Normal	Fraud	Total Dataset
SW	279,336	83,244	362,580
F4H	69,585	20,575	90,160

4.5 MODEL BUILDING

Once the experiment environment is ready, relevant algorithms selected (RF, ANN, and SVM), training mode as well selected which fits for this research (Cross-validation and Separate Test data) and the aggregated dataset mode (SW and F4H). With the possible combinations, a total of 12 models built to detect SIM-box fraud. Table 4.8 shows the possible number of building models with a combination of all the three selected modes (Algorithms, Training, and Datasets)

Table 4.8: List of Built Model

Aggregation Mode	Algorithm	Training Mode
SW	RF	10-Fold Cross Validation
		Separate Test Data
	ANN	10-Fold Cross Validation
		Separate Test Data
	SVM	10-Fold Cross Validation
		Separate Test Data
F4H	RF	10-Fold Cross Validation
		Separate Test Data
	ANN	10-Fold Cross Validation
		Separate Test Data
	SVM	10-Fold Cross Validation
		Separate Test Data

Each model building explained in detail on the coming subsections. Model built with RF, ANN and SVM algorithms addressed in detail on Section 4.5.1, Section 4.5.2 and Section 4.5.3 respectively. Their classification performance as well collected and evaluated.

4.5.1 *RF Model Building*

RF is the top-ranked algorithm for classification problems, detail description is stated on Section 3.2. Among the 12 built algorithm models, four of them are RF's model. RF algorithm model is suitable for instances with less granularity level. The experiment is conducted to build a model of an RF algorithm. The two training modes Cross-validation and Separate test data applied by taking a consideration of data aggregation modes SW and F4H. Each SW and F4H aggregation mode used to build two RF models, finally we get a total of 4 independent built models. Table 4.9 show the developed RF building models.

Table 4.9: RF Build model

Algorithm	Aggregation Mode	Training Mode	Time(s)		Result	
			Build	Evaluate	ROC	Accuracy
RF	SW	10-Fold Cross Validation	1,062.95	1,134	0.99	96.20
		Separate Test Data	798.69	780	0.99	94.90
	4H	10-Fold Cross Validation	217.19	205.09	0.96	91.38
		Separate Test Data	110.98	132	0.95	90.56

4.5.2 *ANN Model Building*

Building model using ANN algorithm follows the same fashion as RF model building explained on Section 4.5.2, as a result of ANN as well built four models.

Table 4.10: ANN Built Model

Algorithm	Aggregation Mode	Training Mode	Time(s)		Result	
			Build	Evaluate	ROC	Accuracy
ANN	SW	10-Fold Cross Validation	1,240.3	1,320	0.8	84.87
		Separate Test Data	1,103.80	1,003.12	0.8	84.52
	4H	10-Fold Cross Validation	365.65	370.52	0.8	84.87
		Separate Test Data	236.87	255.05	0.81	85.09

4.5.3 *SVM Model Building*

The remaining four models are build using SVM algorithm. The training and data aggregation mode is the same building model as indicated on Section 4.5.1 and Section 4.5.2 used. The building model using SVM explained in this Section 4.5.3.

Table 4.11: SVM Built Model

Algorithm	Aggregation Mode	Training Mode	Time(s)		Result	
			Build	Evaluate	ROC	Accuracy
SVM	SW	10-Fold Cross Validation	49,965	52,989.78	0.59	68.90
		Separate Test Data	6,449.47	8,367.69	0.58	68.50
	4H	10-Fold Cross Validation	4.895.28	5,543	0.59	68.34
		Separate Test Data	1,452.83	2.376.25	0.59	68.42

4.6 ALGORITHM EVALUATION

The main objective of this research is to evaluate and compare the classification performance of machine learning algorithms with the desired time span. Once the data collection is completed, preprocessing task is handled and model training and testing are continued. Within one-hour SIM-box frauds detected in near-real time. So, validation performed using 10-fold cross-validation and separate test data to evaluate performance algorithms. All ML models evaluate their performance using different evaluation metrics. The common evaluation matrices are confusion matrix, classification accuracy, F-measure, Recall, Root Mean Squared (RMS) and ROC curve. These evaluation metrics discussed in the coming pages.

4.6.1 Confusion Matrix

Confusion matrix is one way of measuring the performance of supervised machine learning algorithms, as the name indicates the fact that the model gets confused on the two classes. It is a 2X2 matrix which contains True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values of classification class labels.

Table 4.12: Conceptual Confusion Matrix

	Class 'N'	Class 'Y'
Class 'N'	**TP**	**FN**
Class 'Y'	**FP**	**TN**

Where, TP Number of instances correctly classified as Normal (class 'N').

FP Number of instances that belongs to SIM-box fraudulent (class 'Y') but classified as normal (class 'N').

FN Number of instances that belongs to Normal (class 'N') but classified as Fraudulent (class 'Y').

TN Number of instances correctly classified as SIM-box Fraudulent (class 'Y').

Correctly classified instances for both fraudulent and Normal (legitimate) are indicated by TP and TN respectively. On the other ways around, incorrectly classified instances of fraudulent and normal (legitimate) identified by FP and FN respectively. Several researchers [3, 10, 30–32] uses confusion matrix and classification accuracy as a common classification metrics.

4.6.2 *Classification Accuracy*

Classification accuracy measures the ratio of correctly classified (both normal class 'N' and Fraudulent class 'Y') with respect to the overall dataset. the below mathematical Equation 4.4 shows the percentage of correctly classified instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

4.6.3 *F-Measure*

A harmonic mean of precision and recall is F-measure and is calculated as shown in Equation 4.5.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.5}$$

Recall measures the ratio of correctly classified instances as Normal (class 'N') divide by the sum of correctly and incorrectly classified instances of normal (class 'N') as (class 'N') and (class 'Y'). Recall for Fraudulent (class 'Y') computed the same way using Equation 4.6.

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

Precision measures the ratio of correctly classified instances as Normal (class 'N') divide by the sum of correctly classified instances of Normal (class 'N') as (class 'N') and incorrectly classified instances of Normal (class 'Y') as (class 'N'). Precision for Fraudulent (class 'Y') computed the same way using Equation 4.7. The

higher precision indicates the number of miss-classified instances of the model is less (small FP).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.7}$$

4.6.4 *ROC curve*

ROC is a graphical representation of True Positive Rate (TPR) and False Positive Rate (FPR). FPR and TPR displayed in the X and Y-axis respectively. When ROC curves are too close to the top-left corner of the area, the algorithm is considered as a perfect classifier. On the contrary, if ROC curves lie under the linear line (X=Y), the algorithm is considered as low-level classifier.

5

RESULTS AND DISCUSSION

This Chapter describes the experiment results of the research with ten cross-fold and separate test data validation techniques. Both SW and F4H aggregated datasets applied while using selected classifier algorithms RF, ANN and SVM.

5.1 MODEL EVALUATION

The main target of this research is detecting SIM-box fraud near-real time using ML algorithms, and compare each algorithm's performance. In order to overcome SIM-box fraud activities, near-real time SIM-box fraud detection experiments has been discussed in the above chapters, the SW data aggregation technique is applied to achieve the desired results of the research.

Experiments are conducted using the selected algorithms for the detection process of SIM-box fraud. The two aggregation modes (SW and F4H) as well applied to get the final dataset instances for each experiment. 10 cross-fold and Separate/Supplied test data validation techniques are applied to perform the experiment with a supervised ML algorithms. The final experiment results of the model are recorded, evaluated and compared each other.

While comparing the models, RF ML classifier algorithm has better accuracy than the other two classifier algorithms ANN and SVM models. Four independent models were build using those ML algorithms, on both training technique 10-cross fold and Supplied test data, all four models of RF algorithm achieves better performance than the other models build using ANN and SVM algorithms. Next, RF's model compared based on their aggregation mode, SW aggregation mode achieves the highest accuracy of **96.2%** and **94.6%** respectively with 10-fold cross-validation and Separate test data training technique.

10-cross fold validation with SW mode performs better than supplied test data, the model used the same data for training and testing recursively, due to similar data source is used for training and testing purpose.

The overall result of SW mode is better than F4H mode, due to the size of instances used by SW mode is much higher than F4H used. SW has provide huge data size to train and test the model than F4H and obtained better performance than F4H mode.

ANN classifier algorithm models as well achieve the highest accuracy while comparing with SVM classifier models. While comparing ANN's model with each other, each model achieves almost similar performance values which is about 85% accuracy. Each models' performance result stated in Table 4.10. The last but not the least classifier algorithm SVM's model performance, the result is very less as compared with the other two classifier algorithm models. SW with a 10-fold cross-validation training technique gets the result of 68.9% accuracy. The other three SVM classifier models performance presented in Table 4.11.

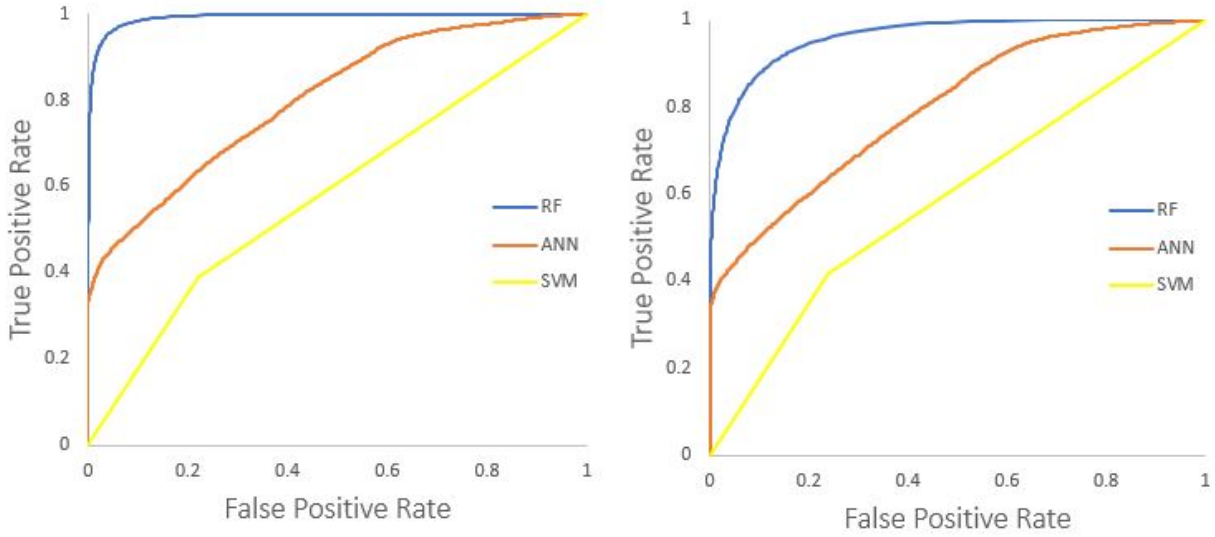
Due to the case explained earlier, using instances as training and test datasets in the case of cross-fold validation technique, experiments usually obtained better performance compared with separate test data validation. Since this research is near-real time, training and evaluation time is one of a major comparison for those selected algorithms. In this research, SVM takes much longer time on both cases (building and evaluation) compared with the other two algorithms while they are doing classification on cross-fold validation. SVM takes more than a day to build a model, similarly with separate test data. Due to that SVM is not recommended for this near-real time research case.

Supplied test data uses small test data size instance for testing compare with the training instance data; because of that, its evaluation time is very less compared with 10-cross fold technique. Using huge instance data increases model building and evaluation time. Windows operating system with 8Gb RAM laptop is used for the experiment. Detail experiment results of 10-Fold cross validation test depicted in Table 5.2. Table (5.1) shows the allover time consumption of each algorithm on model building.

Table 5.1: Modle Build Time

Validation Technique	Algorithm	Aggregation Mode	Build Time (Second)
SUPPLIED	RF	SW	793.69
		H4	110.98
	ANN	SW	1,103.8
		H4	236.87
	SVM	SW	6,449.47
		H4	1,452.83
10-FOLD	RF	SW	1,062.95
		H4	217.19
	ANN	SW	1,240.25
		H4	362.65
	SVM	SW	49,965
		H4	4,895.28

ML Algoritm's classification performance can be presented using graphical representation. To do so, ROC curve is one of the graphical representations technique easily to show the performances of the models and mainly shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) graphically. Figure 5.1 show performance evaluatioin of the model built on 10-cross fold for both SW and Fixed Four Hour aggregation mode.



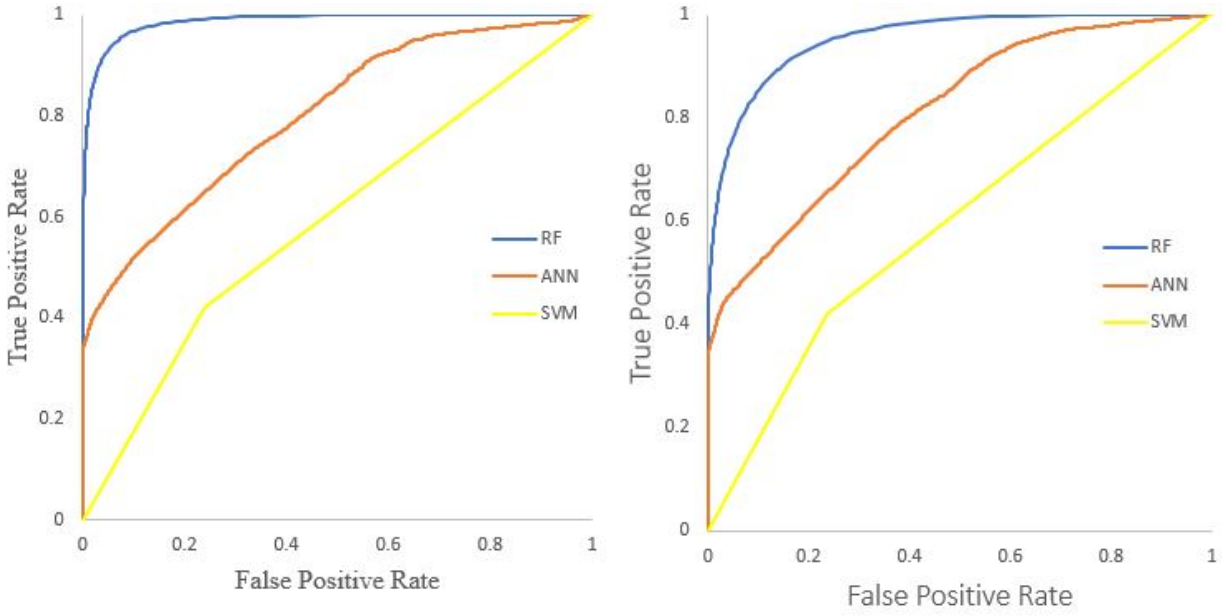
(a) SW (b) F4H

Figure 5.1: ROC curve for 10-Fold Cross Validation of SW and F4H

Table 5.2: Overall Performance of Classification algorithms with 10-cross Fold Validation

Validation Technique	Algorithm	Aggerigaion	Confusion Matrix			Accuracy	F-Measure	Time (Second)
10-cross fold	RF	SW		NO	YES	96.2%	0.961	1,062.95
			NO	460,570	4,991			
			YES	17,952	120,789			
		F4H		NO	YES	91.38%	0.91	217.19
			NO	113,127	2,848			
			YES	10,110	24,183			
	ANN	SW		NO	YES	84.87%	0.822	1,240.25
			NO	461,671	3,890			
			YES	87,563	51,178			
		F4H		NO	YES	84.87%	0.824	365.65
			NO	114,497	1,478			
			YES	21,255	13,038			
	SVM	SW		NO	YES	68.9%	0.694	49,965
			NO	362,466	103,095			
			YES	84,867	53,874			
		F4H		NO	YES	68.34%	0.694	4,895.28
			NO	88,426	27,549			
			YES	20,031	142,62			

Figure 5.1 depicts the performance values' of model built by suppleied test adata with SW and F4H aggregated data instances. A model built using RF algorithm is close to the top-left corner at (0,1) of the graph, which indicates that RF's model performance better than other two algorithm's model. Similarly, ROC curve Figure 5.2 show RF algorithm models achieve better performance over the other amodels. both Figure 5.1a and Figure 5.2a has simmilar looks as a result of having close results.



(a) SW (b) F4H

Figure 5.2: ROC curve for Supplied Test Data of SW and F4H

As shown on both Figure 5.1 and Figure 5.2 RF classifier algorithm remain being a better classifier algorithm with a high performance or accuracy. A model with SW is the top performer on both validation technique 10-cross validation and Supplied test case. unlike RF classifier algorithm SVM is the least performer ML algorithm in this research experiment. The detail experiment results of supplied test depicted in Table 5.3.

Table 5.3: Overall Performance of Classification algorithms with Supplied Test Data

Validation Technique	Algorithm	Aggerigaion	Confusion Matrix			Accuracy	F-Measure	Time (Minutes)
Supplied Test Data	RF	SW		NO	YES	94.9%	0.948	798.69
			NO	183,668	2,557			
			YES	9,777	45,720			
		F4H		NO	YES	90.56%	0.901	110.98
			NO	45,221	1,169			
			YES	4,504	9,214			
	ANN	SW		NO	YES	84.52%	0.824	1,103.80
			NO	181,615	4,610			
			YES	32,818	22,679			
		F4H		NO	YES	85.09%	0.823	236.87
			NO	35,376	11,014			
			YES	7,971	5,747			
	SVM	SW		NO	YES	68.5%	0.695	6,449.47
			NO	141,903	44,322			
			YES	32,071	23,426			
		F4H		NO	YES	68.42%	0.695	1,452.83
			NO	35,376	11,014			
			YES	7,971	5,747			

6

CONCLUSION AND RECOMMENDATION

6.1 CONCLUSION

Telecom operators have suffered by telecom fraudulent activities. A list of telecom frauds categorizes under the behavior of their fraudulent activities. SIM-box fraud is one of the interconnect bypass fraud category which brought a huge impact from customer dissatisfaction up to huge revenue loss and security issues on the telecom companies. Ethio telecom as well impacted by SIM-box fraud activities. SIM-boxers more interested in countries with high international call termination costs and low local call costs, they offer a very less international call transit cost take the money away. Even if telecom operators apply different detection techniques, telecom fraud still being a challenge to telecom operators.

The main focus of this research is detecting SIM-box fraud in near-real time using users' CDR data with the help of ML algorithms. SW aggregation mode is used with the minimum time span of 4-hour.

An aggregation mode with a minimum of 4-hour time span (window size) slides every one-hour to the next. In each windows slide aggregated instance delivered. SW within 4-hour window aggregation technique improve the trade off between detection accuracy and detection delay.

In algorithm performance evaluation, the RF algorithm perform better compared with the other two algorithms model in the detection of SIM-box fraud with ten cross-fold and separate test data. Mainly, RF algorithm with SW aggregation mode scores higher performance values on all evaluation metrics and validation techniques within an hour. SW concept works for near-real time on RF algorithm with better and better detection time. So, the overall accuracy for RF with SW aggregation mode using ten cross-fold validation achieved 96.20% accuracy and 94.90.% accuracy achieved while using separate test data validation techniques.

The amount of time taking by the RF algorithm with SW for classification is much higher than the classification time of RF with F4H aggregation mode. Evaluation time is increased due to SW dataset instance is much higher than F4H data instance.

6.2 RECOMMENDATIONS FOR FUTURE WORK

Ethio telecom reduces the local voice call price which could increase the interest of SIM-box fraudster to hijack international call termination. As future work or recommendations to improve near-real time SIM-box fraud detection, continuous research is required using CDR data analysis and incorporate additional CDR features like, International Mobile Equipment Identity (IMEI) and Mobil Termination ID (Receivers cell ID). In addition to that reducing detection time using state of the art and making more closer to real-time, and quality reduction investigation of a voice call in the detection of SIM-box fraud.

REFERENCES

[1] S. Rosset, U. Murad, E. Neumann, Y. Idan, and G. Pinkas, "Discovery of fraud rules for telecommunications—challenges and solutions," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1999, pp. 409–413.

[2] . G. F. L. Survey, "Cfca," 2018, p. 26.

[3] H. Tewodros, "Network traffic classification using machine learning: A step towards over-the-top bypass fraud detection," 2018.

[4] I. Ighneiwa and H. Mohamed, "Bypass fraud detection artificial intelligence approach," *arXiv preprint arXiv:1711.04627*, pp. 3–6, 2017.

[5] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *IEEE International Conference on Networking, Sensing and Control, 2004*, IEEE, vol. 2, 2004, pp. 749–754.

[6] M. I. Akhter and M. G. Ahamad, "Detecting telecommunication fraud using neural networks through data mining," *International Journal of Scientific & Engineering Research*, vol. 3, no. 3, pp. 1–5, 2012.

[7] R. Alves, P. Ferreira, O Belo, and J. Lopes, "Discovering telecom fraud situations through mining anomalous behavior patterns," *ACM Workshop on Data Mining for Business Applications (DMBA)*, 2006.

[8] Apanews. (2017). Ethiopia loses over $52m to telecom fraud-official. 2017-03-06, [Online]. Available: https://mobile.apanews.net/en/news/ethiopia-loses-over-52m-to-telecom-fraud-official (visited on 01/20/2019).

[9] J. Y. Lee, J. H. Lee, J. S. Yeo, and J. J. Kim, "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, pp. 766–771, 2013.

[10] H. Kahsu, "Sim-box fraud detection using data mining techniques : The case of ethio telecom," p. 84, 2018.

[11] R. Sallehuddin, S. Ibrahim, azlan Mohd zain, and A. Hussein Elmi, "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, 2014.

[12] I. Murynets, M. Zabarankin, R. P. Jover, and A. Panagia, "Analysis and detection of simbox fraud in mobility networks," pp. 1519–1526, 2014.

[13] M. R. Albougha, "Comparing data mining classification algorithms in detection of simbox fraud," 2016.

[14] K. Niu, H. Jiao, N. Deng, and Z. Gao, "A real-time fraud detection algorithm based on intelligent scoring for the telecom industry," *Proceedings - 2016 International Conference on Networking and Network Applications, NaNA 2016*, vol. 1, pp. 303–306, 2016.

[15] F. Mola, "Analysis and Detection Mechanisms of SIM Box Fraud in The Case of Ethio Telecom," *Journal of Chemical Information and Modeling*, p. 76, 2017. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

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

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

[18] I. Society, "Artificial intelligence and machine learning : Policy paper," no. April, 2017.

[19] S. Marsland, *Machine Learning An Algorithmic Perspective*. 2014. DOI: 10.1017/CBO9781107298019.

[20] M. B. K. Mohssen Mohammed and E. B. M. Bashier, *Machine learning: algorithms and applications*. Crc Press, 2016, p. 243.

[21] T. Oladipupo, "Types of machine learning algorithms," *New Advances in Machine Learning*, 2010. DOI: 10.5772/9385.

[22] N. T. Amanpreet Singh and A. Sharma, "A review of supervised machine learning algorithms," 2016.

[23] F. A. Borges, R. A. Fernandes, A. Lucas, and I. N. Silva, "Comparison between random forest algorithm and j48 decision trees applied to the classification of power quality disturbances," in *Proceedings of the International Conference on Data Mining (DMIN)*, The Steering Committee of The World Congress in Computer Science, Computer . . ., 2015, p. 146.

[24] A. K. Mishra, S. V. Ramteke, P. Sen, and A. K. Verma, "Random forest tree based approach for blast design in surface mine," *Geotechnical and Geological Engineering*, vol. 36, no. 3, pp. 1647–1664, 2018.

[25] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised machine learning algorithms: Classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017.

[26] S. Haykin, *Neural Networks and Learning Machines*, Third. McMaster University Hamilton Ontario Canada: Pearson Education, 2009.

[27] Y. Baştanlar and M. Ozuysal, *Introduction to Machine Learning Second Edition*, Second. 2014, vol. 1107, pp. 105–28.

[28] M. K. Jiawei Han and J. Pei, *Data Mining Concept and Technique*, third. Morgan Kaufmann Publishers.

[29] N. Schwertman, M. Owens, and R. Adnan, "A simple more general boxplot method for identifying outliers," *Computational statistics and data analysis*, vol. 47, no. 1, pp. 165–174, Aug 2004.

[30] G. Al-Naymat, M. Al-Kasassbeh, N. Abu-Samhadanh, and S. Sakr, "Classification of VoIP and non-VoIP traffic using machine learning approaches," *J. of Theoretical and Appl. Inform. Technol.*, vol. 92, no. 2, p. 403, Oct 2016.

[31] M. Rathore, A. Paul, A. Ahmad, M. Imran, and M. Guizani, "High-speed network traffic analysis: Detecting VoIP calls in secure big data streaming," in *41th Conf. on Local Comput. Netw. (LCN)*, IEEE, 2016, pp. 595–598.

[32] J. Datta, N. Kataria, and N. Hubballi, "Network traffic classification in encrypted environment: A case study of google hangout," in *21th Nat. Conf. on Commun. (NCC)*, IEEE, 2015, pp. 1–6.