Mobile Roaming Fraud Detection Based on User Behaviour: In case of ethio telecom

BY: SAMUEL MEKASA

SUPERVISOR: 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

(Telecommunication Engineering)

Addis Ababa University

Addis Ababa, Ethiopia

January 24, 2022

Declaration of Originality

Declaring that, this MSc thesis is my original work in accordance with it has not been submitted for a degree at any university before me, as well as that all sources and materials used in the thesis work have been fully acknowledged.

Name: Samuel Mekasa Signature: _____ Date: _____

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

Supervisor: Ephrem Teshale (PhD) Signature: _____ Date: _____

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Samuel Mekasa**, entitled *Mobile Roaming Fraud Detection Based on User Behaviour: In 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:

Examiner 1 Fitsum Assamnew (PhD): Signature: _____ Date: _____

Examiner 2 Sosina Mengistu (PhD): Signature _____ Date: _____

Supervisor Ephrem Teshale (PhD): Signature: _____ Date: _____

_____

Dean, School of Electrical and Computer Engineering

ABSTRACT

Mobile roaming data-internet fraud, committed on visitor networks is a continued challenge and significant source of revenue losses for telecommunications societies including customers. The actually introduced prevention and detection mechanism have limitations in protection of the service.

In this study, we used different data-sets and build roaming mobile data fraud detection model. Three supervised machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM) and J48 decision tree (J48 DT) where used to build model from each data-set. The model performance was computed based on different metrics. The model with merged data-set (roaming in and roaming out) achieved better performance and J48 DT is resulted greater in accuracy of 99.50, average F1_Score 99.00 and ROC 99.30.
For compiled usage behavior exceeds the detection of such fraud, organization better to periodically analysis of data rather than waiting for TAP file-user usage from visited network in addition to revising roaming agreement.

Keywords: *User behavior, Mobile data roaming fraud detection, Mobile data usage, Machine learning algorithms, Machine learning tools, Home network, Visited network*

ACKNOWLEDGMENTS

First and foremost, I like to praise and thank the one God almighty, who is, who was, and who is to come, who will be, and who is alpha-the beginning and omega-the end, for his boundless blessings, wisdom, and care in overflowing compassion, as well as the ability to successfully complete this study.
Apart from my efforts, the success of this thesis deeply dependent on the support and guidance of many people. I take this opportunity to thank everyone who contributed to the complication of this thesis.

I would like to express my heartfelt gratitude to Dr. Ephrem Teshale as the thesis supervisor. I cannot say thank you enough for his grateful, continuous assistance and support. I feel motivated and sharpen ideas and inspired every time I attend the advising session with him. Without his assistance and guidance, this thesis would not have been realized. I like to take this opportunity to thank my examiners for their constructive criticism and ideas during my progressive defense.

Next, I extremely thankful my beloved family for their unfailing love and encouragement through out academic year and the completion of this thesis, especially to my lovely wife Shitaye Gelmesa and Kidist Gelmesa and my humble sisters for their kindly helpful in carrying my home and best wish. In this, I would also like to convey my gratitude to my firstborn exquisite and lovely son Nathan Samuel, whom I perceive as a son, a brother, and also fried for his patience and matured understanding beyond his age. I like to express my exceeded thankfulness to my parents, who, although no longer with us, their unconditional love like a candle burning here self to light someone's life, and non stopped praying is continued to inspire throughout my career. Also I like to thank all of my relatives, church congress, and friends who have encouraged and prayed for me to succeed during this academic term.

Finally, and also most importantly, I would like to recognize telecom staff members, particularly those from ISD and ISec, who have shown tremendous dedication and support in many ways in my study.

Dedication

This MSc thesis is dedicated in loving memory to my parents, Mekasa Michael and Hawe Tolcha, who always believed in my ability to succeed academically.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

3G Third Generation Network

4G Fourth Generation Network

ANN Artificial Neural Network

AUC Area Under the Curve

CDR Call Detail Record

CFCA Communications Fraud Control Association

CIBER Cellular Intercarrier Billing Exchange Roamer

CSV Comma Separated Values

DCH Data Clearing House

DT Decision Tree

FBNN Feed-backward Neural Network

FFNN Feed-forward Neural Network

FMS Fraud Management System

GGSN Gateway GPRS Support Node

GPRS General Packet Radio Service support

GSMA Global System for Mobile Communications Association

HLR Home Location Register

HPMN Home Public Mobile Network

IRSF International Revenue Share Fraud

IDS Intrusion Detection System

IG Information Gain

IQR Inter-Quartile Range

ISD Information System Division

ISec Information Security Division

ITU International Telecommunication Union

LTE Long-Term Evolution

ML Machine Learning

MLP Multi Layer Perceptron

MMS Multi-Media Service

MSC Mobile Switching Center

NRTRDE Near Real-Time Roaming Data Exchange Node

RNN Recurrent Neural Network

ROC Receiver Characteristic Operator

S9 Signaling System for 4G (LTE)

SIM Subscriber Identity Module

SGSN Serving GPRS Support Node

SLP Single Layer Perceptron

SPIN Security/Police Information Network

SS7 Signaling System 7

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

TAP Transferred Account Procedure

USD United State Dollar

VPMN Visited Public Mobile Network

WEKA Waikato Environment for Knowledge Analysis

1

INTRODUCTION

The increased communication service demand pushes global telecommunication networks to encompass many different technologies and has experienced fundamental changes in the past several decades [1], [2]. The introduction of new communication technologies and merging with the internet, inflation its complexity which introduced security challenges such as fraud in telecommunication [1], [3], [4]. ethio telecom who is currently the only telecommunication operator and service provider in Ethiopia is in part of this challenge.

There are many different definitions of telecommunication frauds and the International Telecommunication Union (ITU) [5], specialized agency concerned with information communication technology defines fraud in telecommunication: it is the use of telecommunications network to avoid payment-with incorrect payment, no payment at all or someone else pays. According to Global Telecommunications Fraud Trend Analysis [6], several researchers were overlooks as: telecommunication fraud is defined as the stealing of telecommunication services or the use of telecommunication service to commit other forms of fraud. Such definition has aspect that, the use of the telecommunications service to perpetrate fraud in which the loss is tends to third party rather than telecommunication.

For the difficult understanding of the telecommunication complex ecosystem and rapid flexibility of fraudsters, a comprehensive understanding and avoiding fraud in the phenomena is a challenging task. To manage the problem by individuals in the sector, one needs to have a balanced knowledge of telecommunication systems such as expected threats and their detection and prevention techniques. Nevertheless, industry experts working in fraud management have a partial view because they usually specialize in the fraud type most likely encountered or detected in their businesses [1].

Fraud in telecommunication is a major source of revenue loss for telecommunication service provider and their customers [2], [7], [6]. The experts feedback-from phenomena states that, most telecommunication providers are losing three (3) to ten (10) percents of their income doe to frauds challenges in telecommunication [8].

Nearby (2019) Communications Fraud Control Association (CFCA) [7], is the non-profitable organization who conducting surveys on annual global fraud loss in every two years periods shows that, global fraud loss estimated $28.3 billion (USD) and billions of dollar losses associated with roaming fraud. [6], [7], [9].

The recent three years (2015, 20117, and 2019) of global telecommunication operators and losses from ethio telecom annual report due to frauds are discussed in table 1.1.

Years	Global Telecom Revenue Losses in case of Fraud ($)	Annual Income Lost in Percentage ($)	ethio telecom Revenue Losses in case of Fraud ($)
2015	38.1 Billion	1.69	33 Million
2017	29.2 Billion	1.27	89 Million
2019	28.3 Billion	1.74	48 Million

Table 1.1: Telecom Operators Revenue Loss in case of Fraud [7], [6].

Penetrative technology, employee disappointment, organizational inefficiencies, weakness of business models, financial crimes (money laundering), geopolitical and socioeconomic influences are some indication for the motivation of fraudsters [1], [3].
International mobile data roaming is one of the highly affected areas by such frauds.
It is possible to decrease the dimension of financial loss generated by impacts of fraud in the telecommunications industry by building models that notice fraudsters' acts.

1.1 MOTIVATION OF THE STUDY

In addition to resulting in revenue losses, the ongoing challenge of international roaming data fraud decreases user satisfaction and experience, as well as an organization's reputation.
To fight such fraud, telecommunications organizations like ethio telecom rely on using antiquated system-rule-based systems like the Fraud Management System (FMS). However, it is not advanced because it generates a high number of false positives.
Delay of a TAP file for further analysis from a visited network provides fraudsters with an extended possibility to execute similar frauds.

1.2 STATEMENT OF THE PROBLEM

In order prevent and detect roaming fraud, several mechanisms, guidelines, and suggestions have been declared by many studies.

The recommended Near Real-Time Roaming Data Exchange (NRTRDE) or CDR back to the home network for quick analysis as wel as exploring the weakness of technology and poor business process in the area are not advanced detection and prevention mechanism of fraud as it did not consider the subscriber behavior can changed frequently [1], [3], [10]. On other way, NRTRDE takes bout four hours-based on GSMA [11], roaming agreement which is long enough for fraudster to introduce benefit.

Similarly, [12], [13], [14], [15], have been built rule-based algorithms and applications to notice this fraud. However, a rule-based system, in which thresholds are set to produce alarm signals based on predefined rules, is ineffective in detecting such fraud. It can introduce significant proportion of false positives sometimes exceeding 50%, limited coverage and delay in response [16],[17] and the analysis needs a significant amount of human resources because it must constantly regulated to stay successfully. All these results in fraud losses.

Recently, a predictive model for detecting international roaming fraud is developed in [18]. This has substantial value for detecting roaming fraud. The model is confined to roaming out subscribers with a togethering of voice, SMS, and data-mobile internet service types. However, there are notable frauds with roaming in on the visited network-in this case ethio telecom. However, such considerations may limit from deep sight with a specific service type's fraud as usage behaviors are contributed from many service types as features. The study also utilized TAP file-out roamer's usage from the visited network forwarded through a third parties, which takes a long time and maybe enough time for a fraudster to commit fraud. The only evaluated long usage time as high usage, which is restricted in terms of depicting fraudsters' overall behavior.

By using mobile roaming user behaviour that can available in an organization's network, it is possible to detect roaming fraud from both roaming in and out, and also possible to minimized fraud detection time along with algorithm classification performance.

In light of the mentioned issues and the inadequacies of the state of the arts, this study would reflect and answer the following research questions.

1.2.1 *Research Questions*

1. What usage behavior or data features can be suitable for mobile data roaming fraud detection?
2. Which machine learning algorithms can significantly used in predicting the patterns of roaming fraudster's?

1.3 OBJECTIVE

1.3.1 *General Objective*

The main goal of this research is to detect mobile data roaming fraud-based on the user behaviour using machine learning algorithms.

1.3.2 *Specific Objectives*

The specific objectives of this study are:

- To insight the international roaming fraud nature specifically mobile data fraud with the tendency to ethio telecom.
- To select relevant usage fields or feature used in building the fraud detection model.
- To evaluate and compare selected ML classifiers based on their performance results.
- To build three alternative models and choose the optimal approach for holistic detection of mobile data roaming.

1.4 SCOPE AND LIMITATION OF THE STUDY

From many forms of frauds discussed under common types of fraud in telecommunication, this study confined to an in-depth understanding of what is roaming data fraud and its detection mechanism.

Because data beyond these months is inaccessible, we look for about 12 months of data during developing model_1. However, it does not appear less for the development of this model.
When compared to roaming out customers, roaming in customers have a much higher daily usage and subscriber base. As a result of storage constraints and limitation of device processing performance, we are confined to three months of data in the development of model_2, whereas we integrated common features from model_1 and model_2 used to build model_3.

1.5 CONTRIBUTIONS OF THE STUDY

In respective to the thesis work, the conceptual and technological significance of the study are provided as the following.

- Identifies undisclosed roaming fraudulent activity.
- Provides telecom operators awareness and allows them to identify and detect mobile data roaming fraud and related.t
- Provide practical approaches which can assist experts and analysts to understand roaming fraud effectively.
- Contribute by overcoming the existing fraud management system of the organization.
- Suggests potential machine learning algorithms for detecting the fraud.
- Future studies on topics related to the title could benefit from using it as a citation.

1.6 RELATED WORKS

Several studies were conducted on telecom fraud detection and prevention to explored fraudster's impact on telecom operators, service providers, and customers. The influences continued measured in financial loss and customer satisfaction and numerous detection methods are proposed.
To have insights into the research topics about fraud detection and prevention mechanisms specifically international roaming fraud detection, related

kinds of literature like journals, articles, magazines besides the internet are reviewed and discussed as follows.

According to Merve Sahin [1], the systematic exploration of fraud in telecommunication is studied. In the journey of the study the root causes of frauds, the vulnerabilities of industry, the exploitation techniques, the fraud types, and the way fraud benefits fraudulently is covered. To have an allencompassing understanding of the area many available the state of arts were conducted, domain experts in different telecom operates were interviewed, many telecom security detection and prevention techniques forums at global are attended and surveys were conducted for other community having knowledge in the phenomena.
The study contributed to the field of fraud detection a taxonomy that distinguishes the origin of fraud, the exposures, the fraud types the exploitation techniques, and finally the way fraud benefits fraudsters. For detecting international roaming high usage, the authors recommend that implementing Near Real-Time Roaming Data (CDR) Exchange (NRTRDE) systems back to the home network is important to detect on-time frauds in the area through quick analysis.
Following the fact of this conclusion, the transmission of CDRs from the visited network to the home network still takes about four hours [1], [3]. which is a long enough time gap for the fraudsters to make a profit.

In Gebriel.M [3], comprehensive ideas about how roaming service works, and the vulnerability of the service in the roaming scenario were discussed. Fraud enabled by technical factors and fraud enabled by errors in business areas is two main categories that exposure the service. Following the classification, they proposed different existing methods to challenge the problems. These proposed approaches are ranged from a statistical model to more complex methods such as data mining and machine learning such as neural networks [1],[3]. In the era of fraud, detection using either statistical or machine learning algorism are required data for analyzing, building a predictive model, and forecasting [19].
However, neither collected data nor tools from the machine were applied to detect fraud in the roaming scenario for high usage.

As Maciá-fernández, G.[10], roaming fraud attack and defense strategies which focused on the telecom service and their network security threats were discussed. The authors had been proposed fraud classification techniques for the expected attacks and highlighting the role of different players in joint actors to commit the fraud. The quantitative research method was used to quantify fraud techniques and their protection policies. As a

conclusion from assessment, less maturity of the technologies in the fiend was, sightsee as a reason for fraud in the phenomena.
On the other hand, in today's world, the behavior of fraud and the degree of impact is changing frequently. Therefore, limiting fraud detection to the high maturity of technology is not sufficient for fraud detection.
Moreover, using a mechanism such as a machine learning algorism that can learn the trends of subscribers or data is an advanced technique to detect such fraud.

Moudani and Walid [20] were built an algorithm that determines suspected fraud numbers. The purpose is to increase the income of the telecom service provider through the detection of international roaming fraud and to minimize the dependence on third parties in fraud detection. The algorism can also classify whether the fraud has been treated as local or international call fraud such as fraud during roaming.
To easy, understanding of store procedure-built algorism, in this case, a decision tree was used to illustrate the procedure. The algorism was verified with data from home networks and data from the cleaning house. From CDR data used, challenges such as timestamp or time zone and call duration were required adjustment in the process of classification.
Even though the built algorism used to classify the calls based on used CDR, it does not analyze the behavior or pattern of frauds. Also, the store-procedure was verified only with three days of CDR data.
However, verify using fewer data may not effective as much as very with large data [15]. Data is collected only from the billing system even with missed some data in time-bound and data from network elements was considered. Considering a larger pool of historical data and signaling data are better to speed up the cycle of fraud detection [12].

In H. M. Marah [13] detecting fraud based on user profiling and fuzzy logic was discussed. The user profiling approach was used which analyze the subscriber's (SIMs) activity and behavior on detection patterns. In this case, techniques used for fraud detection fall into two primary classes: statistical techniques such as data manipulation including computing user-profiles and artificial intelligence such as machine learning techniques.
Five features such as subscriber's mobility, incoming to outgoing calls ratio, suspicious cell activity, irregular calls, and service type were extracted and engaged as detection patterns in the proposed technique. Fuzzy logic was used in the decision-making process by using a fuzzy logic membership function. The degree of membership for all features was calculated from obtained each membership function. The result is compared with the threshold which the operator can set. Depending on the comparison deci-

sion is made either it's fraud or not.
Following the model, no data is verified to show the degree of detection. In today's world, the behavior of customers may not be limited to a specified number of detection patterns or features.
Moreover, weights for each selected pattern were not specified to prioritize frauds based on their risk of impact and the last step is a rule-based fraud management system in which is threshold has been set by the operator.

In Q. Zhao [14] detect telecommunication fraud by analyzing the contents of a call was discussed. In process of achieving the objective, to understand the fraud behavior or contents of the call, the telecom operator's reports or claims related to fraud were used as input data for the study. Machine learning algorithms were used to realize its relation with the definition of fraud stated in the internet or social media optimum prediction accuracy was achieved with the selected dataset. The Natural Language Processing (NLP) technique is another method used to extract features from the data and voice conversation to text to meet the requirement. Using the contents, an application that alerts the user by understanding the content of calls was developed on the android platform. The application uses speech recognition techniques to identify the content of the call and justifies fraud calls based on the features before warning the user.
On the other hand, identifying fraud late is earnings inefficient savings as most of the loss has already occurred [17], [21]. The developed application is used to perform online detection of telecommunication frauds. However, the fraudster does not require the internet or being online even for international calls to commit fraud when calling and unencrypted content stored on call receivers equipment.

Article [22], [23] and [24] aimed to show possible security vulnerabilities of signaling elements or SS7 and cyber attacks for roaming network and implementation machine learning against rule based filtering for detection of SS7 attacks. Specifically, article in [23] relies on inclusive review of the SS7 expected attacks and provides mitigation techniques such as machine learning based framework to detect anomalies in the SS7 network with comparative to rule based filtering. However, the real user data was not considered in the scenario which may makes fail to show the real behavior of roaming fraudster.

In parallel, in study of [22], a fake base-station is installed to establish a connection to a subscriber through the air interface. The IMSI (International Mobile Subscription Identity) is captured using this fake station. To explore the network-network communication an emulator based (jSS7 simulator)

LTE test bed is used. The author has investigated how Diameter messages can be manipulated over the S9 interface to perform a fraud or DoS attack using the IMSI number and in its result shows the difference between abnormal and legal usage. On the other hand, analyzing using labeled data from real network traffic is better realistic than using such fake base station.

Worku, Tarikua in [18] to build a predictive model used to prevent and detect international mobile roaming fraud through analyzing the international roaming traffic. In the process of developing the model, roamer's CDR data with different service types were considered and analyzed. The developed models experimented with selected three supervised machine learning namely Random Forest, ZeroR and J48 algorisms.
Finally, one algorism (Random Forest) with a better outcome was selected and recommended. We credit this works for detecting roaming fraud.
However, the measurement parameter or the evaluation of this model is not inclusive [16], [25] or evaluated with a limited number of parameters which may place the precision of the model under the question. In addition, all services with different data types such as Voice, Data, SMS and MMS were grouped under the same category and even in this case, less amount of attributes were considered. In accounting for this limitation and others, one needs to come up with a model that detects roaming fraud with selected service mean voice roaming service with both inbound and outbound roaming subscriber customer usage and CDR data.
In the process, multi-dimension measurement parameters, which include confusion matrix, accuracy, recall, precision, F-Measure, Receiver Operating Curve-ROC, and other required evaluation parameters will be considered.

Tekeste, Derebe in [16], to detect subscription fraud through analyzing subscriber usage behavior or CDR data. They build and train the model with classified dataset into training and test dataset with three to one (3:1) ratio. Three selected supervised machine learning (ANN, SVM, and J48) algorism have been used for the experiment. Deferent evaluation parameters such as confusion matrix, accuracy, recall, precision, F-Measure, and others were computed to select algorism with the best outcome. From experimented algorisms, one with optimum evaluation parameter (J48) was recommended as the best algorism for detecting subscription fraud with parameter set and data used.
The objectives and methodology to overcome the roaming fraud problem are recognized, but it is better to consider roamers CDR data since such fraud is committed on different operator's network and the traffic pattern of the fraudulent is different from those of fraud.

1.7 RESEARCH METHODOLOGY

This study combines quantitative as well as qualitative (data science) approaches. Numerous indexes including the number of data collected and months taken into account, the ratio of fraudulent to non-fraudulent indicators, the number of training and test data (the ratio of training data to test data), the number of models and machine learning algorithms compared, and the number of tools used, lists of quality factors and their associated attainable values are determined using the quantitative system.
The qualitative approach is used in the expression of expert views, data preparation procedures, and narrative reasoning to determine which classifier and modeling approaches are recommended for feature use.

Concisely, this thesis used the following methodology to achieve the objectives and stated research questions.

- **Phase 1:** Insight the telecommunication related frauds specifically mobile data roaming fraud through reviewing related the state of the arts and domain experts interview.
- **Phase 2:** Collect legitimate and fraudulent mobile roamer usage(roaming out and roaming in) CDR data from ethio telecom business relation management and fraud management system and formulating domain expert's view.
- **Phase 3:** Applying data prepossessing (cleaning, Integration, Aggregation and Transformation).
- **Phase 4:** Building three models i.e detection for roaming out (model_1), for roaming in (model_2) and detection with merged data set (model_3).
- **Phase 5:** Evaluating models using identified metrics (Accuracy, F1 score, and ROC) to select one with higher classification performance. To do this, the python tool packages and libraries described in the tools for machine learning section are employed.
- **Phase 6:** For potential development, an option model in association with a better performance classifier is recommended.

1.7.1 *System Model*

In order to achieve goal of the study a compressed system model is illustrated in figure 1.1

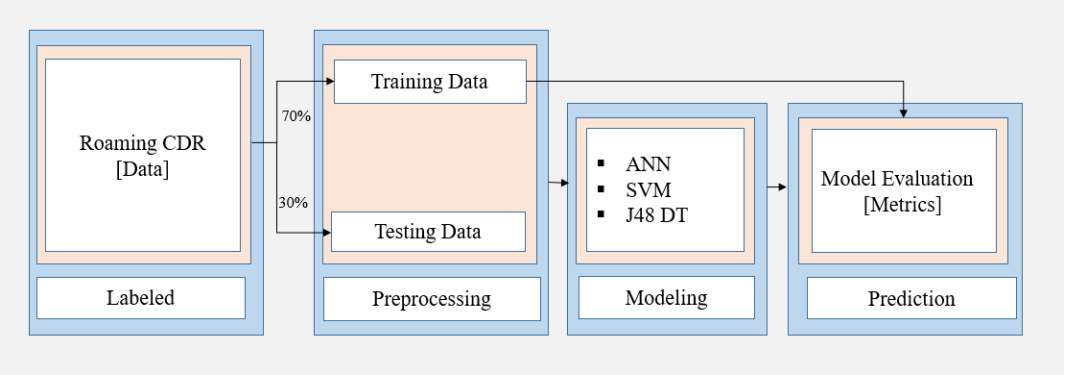


Figure 1.1: System Model of the Study

1.8 THESIS ORGANIZATION

This thesis paper categorized into six chapters and the focus of each chapter is discussed as follow.

- **Chapter_1:** Address the study's background, the problems it attempts to address, the goals and scopes, related works, and methods of study.
- **Chapter_2:** Describe major telecommunication services, such as service subscriptions , roaming services, as well as common frauds.
- **Chapter_3:** Deal with the machine learning classifiers algorithms [J48 DT, ANN, and SVM] and tools (python libraries).
- **Chapter_4:** Takes into account the data preparation techniques, including preprocessing, transformation, and train and test data split.
- **Chapter_5:** Explains compares, and contrasts the results of various algorithms and modeling approaches.
- **Chapter_6:** Concludes and makes suggestions for further studies connected to the claim.

2

COMMON TELECOMMUNICATION SERVICES AND FRAUDS

2.1 SERVICE SUBSCRIPTION TYPES

Telecommunication service providers offer types of services to their subscribers to reach the mission and to win the heart of their subscribers because of the market competition when numerous telecom operators have focused on the same market area as well as when telecommunication service providers are sharing the same infrastructure or others like ethio telecom also provide various services to join its mission and intensification their customer comfort. Nevertheless, most subscription services are common to most the operators, which includes prepaid and postpaid services [26]. The difference is defined by their commitment and ways of payment for their usage. With prepaid, you will always pay/recharge for your plan upfront; whereas postpaid is bases on monthly billing for the usage. In addition to the interest of the customer, segmentation may be based on the subscriber's capability of paying their bills after investigating customers recharging history [27].

The current total number of subscribers of ethio telecom is about 51 million with 85% geographical network coverage [28]. Some common service types provided by telecommunication service providers such as ethio telecom and its characterization are discussed under this section.

Prepaid Service: This is the most popular service provided by mobile operators; users of a mobile apparatus can use a prepaid Subscriber Identity Module (SIM) card to access particular or more services made available by the prepaid SIM card like internet or data, voice, and short message service (SMS) services. All transaction in this service is pay-as-you-go and it is easy for the mobile operator to maintain in the event of fraud and also less susceptible to fraud as compared to postpaid services as of the subscriber also supervise its timely usage [26], [29]. Subscription and SIM-Box fraud are examples of conman frauds that affect this service. [16].

Postpaid Service: A postpaid or pay-monthly subscription in mobile communications refers to a service contract where a user is billed towards the end of every month for the mobile services they have consumed in a given period [27]. It is the most convenient service offered by telecommunication service providers [26]. ethio telecom launched postpaid services with it's available technology. One of its advantages is that it provides clients with more alternatives and allows users to stay connected at any time and any where[30]. The services such as roaming use are currently limited to such categories.

2.2 ROAMING SERVICE

According to GSMA [11]- is an industry organization that represents the interests of mobile network operators globally, international mobile roaming is a service that allows mobile users to continue to use their mobile device to get the services remotely through accessing different networks referred to as the visited mobile network.
The seamless extension of coverage is enabled by roaming agreement between home mobile network operator and the visited mobile operator networks. This agreement addresses the technical and commercial components required to enable the services for roamers.
The service was globally introduced in 1990 and signed in 1992 between telecom of Finland and Vodafone UK [18] . Similarly, ethio telecom brings together its postpaid customers in 2011 and is currently active with 442 operators globally [31].

Common international roaming services are discussed below:
Voice Roaming: Making and receiving calls to or from home country, visited country or a third country, while abroad.
SMS Roaming: Sending and receiving text messages to or from home country, visited country or a third country, while abroad.
Data Roaming: It refers to the use of mobile data services in time traveling through visited networks. The most common mobile data roaming services are:

- **E-mail Service:** Reading and replying to e-mails using a visited network, automatically delivered to mobile devices or other recipient devices.

- **Multi-Media Service(MMS):** Exchanging high multimedia information at the moment a broad with other customers on networks at home or elsewhere.
- **Handset Internet:** Accessing internet services such as interesting web page applications, music upload or download, and video streaming through mobile devices.
- **Mobile Broadband:** Connect devices by data cards or USB dongles to the internet to gain access to applications such as e-mail, web browsers, and an organizational network.
- **Applications:** Using mobile applications while abroad that require mobile data, such as location-based services and language translators.

The roaming service also includes hotel services, such as Wi-Fi, national SIMs, and visited operator SIMs [11].

2.2.1 *Roaming Scenario*

There are three main players in the roaming scenario: the subscriber- who makes use of telecom services, the home mobile network-management subscriber profile, and the visited mobile network-who provides users with access to services on the owned network through a contract with the home mobile network of the user [3], [10].

When roamers beyond the home network switch on their cellular telephone, the device attempts to communicate with a visited mobile network. This network takes care of the connection from the user's device and verifies if it is registered with home system by checking the user's home network. For devices that are permitted to actually employ roaming agreements, a temporary subscriber record is created in the visiting country's network resource and the home network's subscriber record is updated with the device's current location. In this manner, the request is made using this mobile, it may be suitably routed by the visited network towards an international transit network [11]. When it comes to getting service requests to their destinations, carriers are in charge of delivering them over international transit networks. Once this is completed, the destination network will establish a connection and begin using. The overview of international roaming scenario is illustrated in figure 2.1.

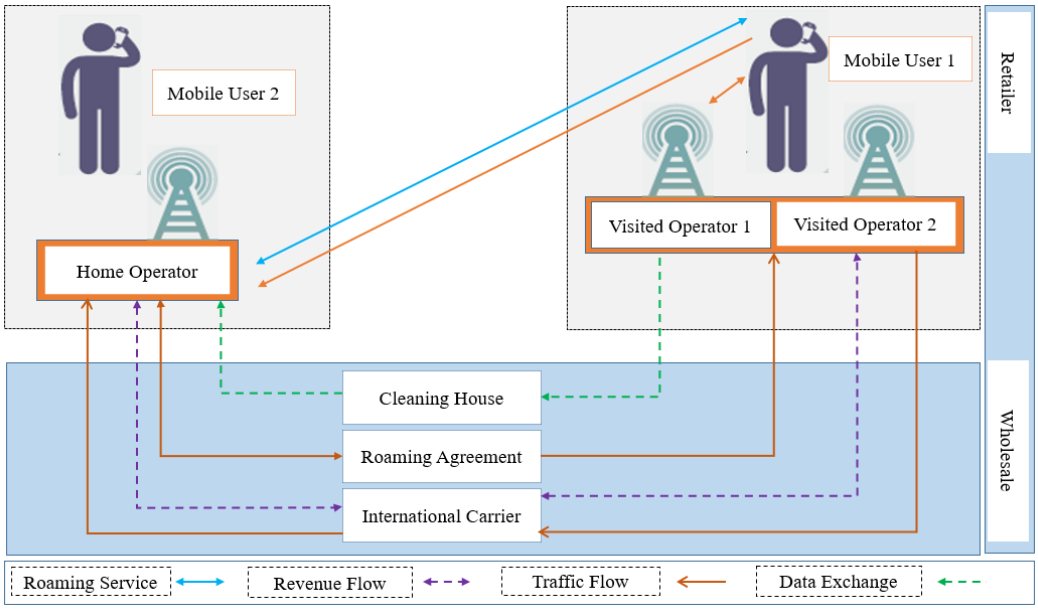


Figure 2.1: Roaming Scenario [3] , [10].

The roamer pays the home operator a regular fee for the roaming service and does not repay the visiting operator in either form. If the mobile user is not also roaming, there will be no additional costs for receiving or making calls to the roamer mobile user. TAP file is sent from the visited operator to a clearinghouse, which then delivers them to the home operators. The file is used to charge customers for their usage when roaming. The home operator can then pay the fees to visit the operator based on the usage volumes recorded in the and the international roaming contract's rates [3], [10].

There are certain criteria that separate mobile data roaming from voice usage while roaming. After querying the HLR, the subscriber is assigned to a node known as the serving General Packet Radio Service (GPRS) support node (SGSN). The roamer then specifies the data network to which a connection should be made, and a context is established between them through a node known as the gateway GPRS support node (GGSN) [3]. In a summary, figure 2.2 depicts the scenario for the data path in roaming.

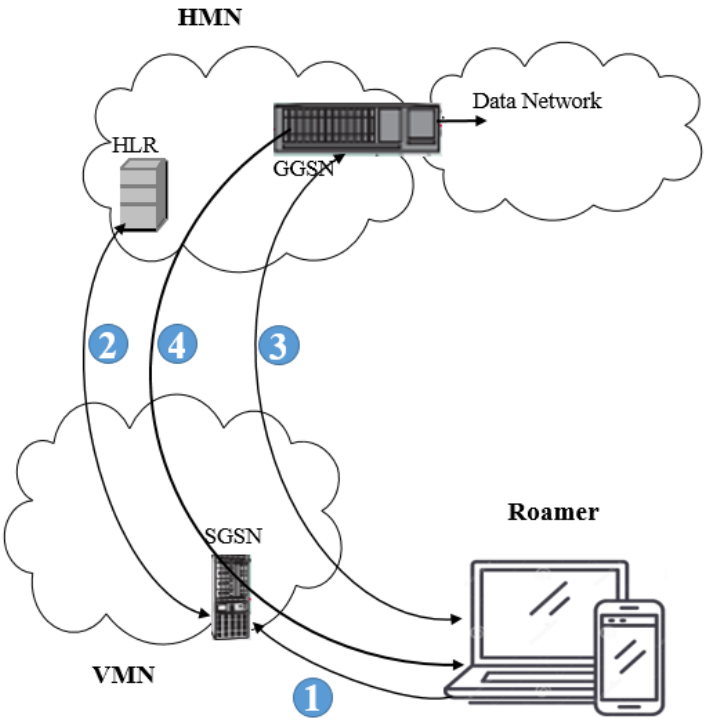


Figure 2.2: Scenario for Data Access in Roaming Based on [11], [10].

Referring to figure 2.3 numbers indicates:

- 1 Shows a network connection request to SGSN
- 2 Is SGSN inquiry to Home Public Mobile Network (HPMN) about the subscription status Home Location Register (HLR),
- 3 Designates a context establishment request to GGSN.
- 4 Indicates data connection setup.

Clear that the SGSN is part of the VPMN, whereas the GGSN is part of the HPMN. As a result, both the data sent and received by the mobile user must pass through the HPMN. However, in the case of voice traffic, this does not appear to be the case, as the VPMN-HPMN interaction is frequently reduced to simply the first inquiry to the HLR.

2.3 TELECOMMUNICATION FRAUDS

Authors and experts articulate the frauds from many perspectives. Nassau County-Security/Police Information Network (SPIN) Department[32], ITU

[5], GSMA [11] characterize as it is the theft of telecommunication services and other articles whereas others such as [26] , [33], [2], it is the use of telecommunication service to commit other forms of fraud.
Based on the fraudsters' targeting and technology behaviors, frauds telecommunication are grouped into the following categories.

Contractual fraud: fraudster uses telecom services with no intention to pay the service charge.
Hacking fraud: fraudsters breached the systems of business and take advantage of available resources illegally.
Technical fraud: fraudsters in this category capitalize on weaknesses that exist in mobile system technology.
Procedural fraud: frauds under this group involved attacks against the procedures implemented to reduce the risk of exposure to fraud, and often attack the weaknesses in the business procedures used to grant access to the system.
Internet Fraud: the use of internet technologies to display fraudulent solicitations to potential victims, to perform fraudulent transactions, or to move the proceeds of fraud to financial institutions or others linked with the scheme is referred to as a fraud scheme.
Identity Theft: the use of personal information to mislead others that the impersonator is that person, essentially passing oneself off as someone else.

2.4 COMMON TYPES OF TELECOMMUNICATION FRAUDS

There are numerous types of fraud that threaten the telecommunication sectors. It is estimated that more than 200 variants of frauds exist in the telecommunications industry according to [33]. The most commons are discussed below.

2.4.1 *Subscription Fraud*

Subscription fraud is in which the fraudsters tend to subscribe to telecom services by using false or fake personal information with no intention to pay for the service and usually it is known as an originator to other types of fraud meant for Premium Rate Fraud, International Revenue Share Fraud (IRSF), SIM-Box fraud and Roaming fraud[16]. The real impact of this type of fraud is difficult to measure because it does not stop with revenue loss

alone as it extended to poor customer experience and dissatisfaction support staff.

2.4.2 *SIM-BOX*

SIM-Box fraud is a technique by which local SIM cards are used for rerouting international calls away from mobile network operators and transfer them over the internet, and deliver them back through VoIP gateway device called SIM-Box as local calls to the operators cellular network [34], [33]. It can be committed as individuals or organizations by using thousands of SIM cards offering free or low cost calls to mobile numbers.

2.4.3 *International Revenue Share Fraud*

According to [34], [33], International Revenue Share Fraud (IRSF) is a telecommunication fraud that occurs when an operator agrees with another party that will make calls, especially to a premium rate number to generate revenue. This usually involves a combination of multiple fraud schemes such as misusing roaming service, call divert, or call forwarding techniques. Fraudsters generate high traffic calls to high-cost destinations and gets revenue from the sharing agreements [33]. This is the most challenging fraud for mobile network operators and service providers.

2.4.4 *Wangiri Fraud*

According to definition from [35], [36] Wangiri is a Japanese term states stated to as "one ring and cut fraud". It relies on this single ring method for a quick way to make money whereas either receiving missed calls from international numbers do not recognize on a mobile or a fixed-line phone is normal call or fraudulent calls. The fraudsters generating the missed calls and then immediately disconnect the calls to them hope that their expensive international numbers will be called back so that they can profit.
As [1] state, these techniques taking advantage several weaknesses in telecom system, which could be related to the basic technologies such as lack of caller ID authentication, third party services like offensive premium rate number which is expensive than normal tariff. The presence of legacy proto-

cols, convergence of multiple technologies and variety of service providers are also considered as the basis causes that result in these weaknesses.

2.5 FRAUD IN ROAMING SCENARIO

The fraud in the roaming case begin internally when subscription in the home network and ranged to international revenue-sharing fraud, in which fraudsters typically use illegal resources to gain access to an operator's network [10], [33].
According to the roaming synopsis the corresponding CDRs-usage detail is generated and transmitted from the visited network operates to the subscriber network-based or home network based on a global roaming service agreement for billing [37], [3].
Receiving roamer usage data passes various stages and takes several hours due to roaming behavior, which is a considerable amount of time to commit fraud.
The transfer of TAF file from the visited mobile network to the home mobile network for billing purposes is visualized in fig 2.3.

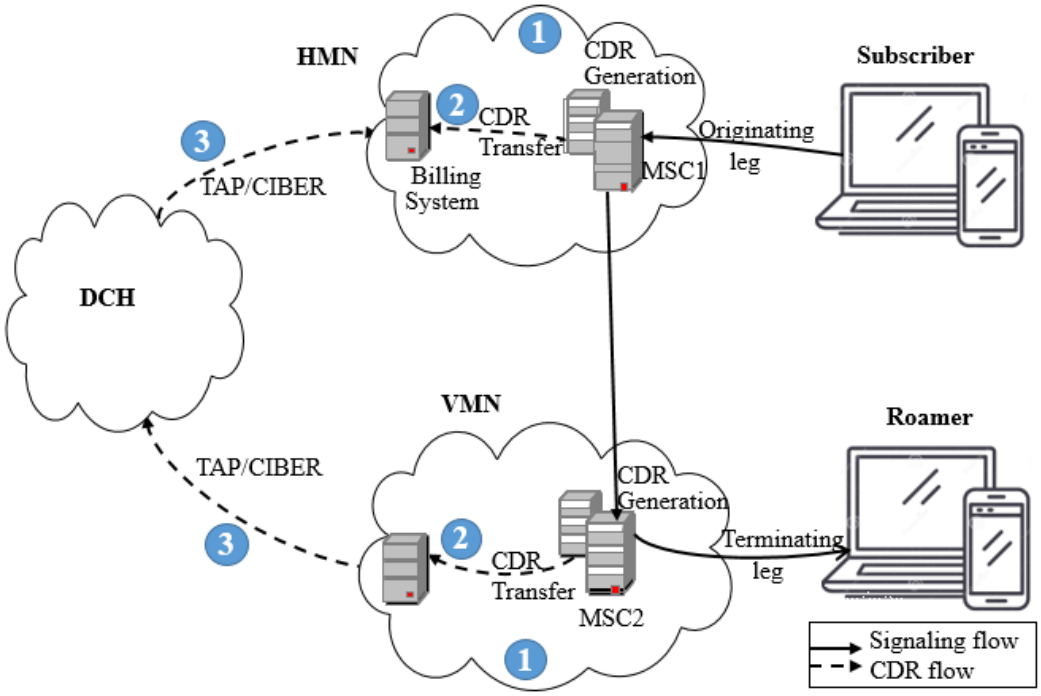


Figure 2.3: Billing information flow from VMN to a HMN: [3], [10].

- Step 1: CDRs are generated from mobile switching centers (MSCs).

- Step 2: CDR data is transferred from MSCs to billing systems.
- Step 3: TAP/CIBER is used to transmit CDRs from visited mobile network operators to home mobile network's through data clearing house (DCH). A DCH is a single interface for an operator that is responsible for all elements of transmitting, receiving, and transforming TAP CIBER data on behalf of the service provider.

As part of a roaming deal, the home network should pay the amount owed to the visited mobile network only by subscriber. However, the user can receive the service by using fraudulent and exaggerated CDRs data, and the home network is unable to charge the customer the equivalent price. As a result, the operator suffers loss [3], prompting the company to pursue. Moreover, SIM snatching, device theft, SIM cloning, identity theft, and renting a phone are a few ways to actively perpetrate such fraud.

3

BASICS OF MACHINE LEARNING

Arthur Samuel [38] a pioneer in the fields of computer games and artificial intelligence, characterized machine learning as a area of research that provides computers the ability to learn without being explicitly taught.
According to experts from [39] and [40], Machine learning, is the capacity of the machine to generalize knowledge from data by utilizing algorithms to collect data, learn from it, and then forecast future patterns for that topic. The goal of machine learning algorithms is to learn how to do specific tasks, such as making correct predictions or locating particular patterns in data.

The machines are fed processed data, and various techniques are employed to create models through train the machines using data. In fact, the algorithm used is determined by the nature of data and the sort of activity that has to be handled. However, in order to work with machine learning techniques, we should be familiar with prevalent terms.

3.1 COMMON TERMINOLOGIES OF MACHINE LEARNING

A working knowledge of several terminology is required while dealing with Machine Learning (ML). The following are examples of machine learning terms, according to [41] and others.
Feature: It is a measurable property or parameter of the data set.
Feature Vector: A set of multiple numeric features used as an input to the machine-learning model for training and prediction purposes.
Model: It is the mathematical representation of a real world process. The machine-learning algorithm along with the training data builds a machine-learning model.
Training: Machine-learning algorithm such as supervised learning takes a set of data known as training data as input and the learning algorithm finds patterns in the input data and trains the model for expected results and the output of the training process is the machine-learning model.
Training Dataset: A subset of a more complete data set used to train a model whose practical performance will be tested on a test data set.

Test Dataset: A subset of a more complete data set used to test the experimental performance of an algorithm trained on a training data set.
Prediction: Once the machine-learning model is ready, it can be fed with input data to provide a predicted output.
Target (Label): The value that the machine-learning model has to predict is known as the target or label.
Overfitting: When a massive amount of data trains a machine-learning model, it tends to learn from inaccurate data entries such as noise and the model fails to characterize the data correctly.
Under-fitting: It is the scenario when the model or the algorithm does not fit the data well enough. The model fails to decipher the underlying trend in the input data and destroys the accuracy of the machine-learning model.
Accuracy: Proportion of results correctly classified against to total number of results predicted.

3.2 TYPES OF MACHINE LEARNING

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

3.2.1 *Supervised Learning*

Supervised machine learning is a category of machine learning defined by its use of labeled datasets to train and classify data or predict outcomes accurately. The training dataset includes inputs and correct outputs, which allow the model to learn over time [16] [42].
The typical supervised machine learning algorithm consists three components to identify historical trends to inform future models [40].
A decision process: a method of calculations or other steps that takes in the data and returns a conclusion at the kind of pattern in the data your algorithm is looking to win.
An error function: by comparing the estimate to known cases, you can determine how good it was? Is it safe to say that the decision-making process was successful? If that's the case, how can you determine "how awful" the error was?
Optimization process: where the algorithm looks at the error and then up-

dates how the decision process comes to the final decision so that the next time the miss won't be as great.
Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying frauds (financial, spam, subscription, SimBox, Roaming etc.), intrusion detection, network traffic prediction and others. Supervised learning problems can be grouped to two:

Regression Problems: it is as supervised learning that used to understand the relationship between dependent and independent variables. The aim is to model the relationship between a certain number of features from the historical data used to build model that predict future values.
Linear regression, logistical regression, and polynomial regression are popular regression algorithms [42].

Classification Problems: classification models are used to predict new outputs based on classification rules by train the algorithm to identify items within a specific category. Detecting fraud or classifying fraud calls from the non-fraud call is an example of the classification category. Common classification algorithms are, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest [42], [39].

3.2.2 *Semi-supervised Learning*

In this type of learning, the given data are a mixture of classified and unclassified data [43]. Such types of machine learning lies between supervised and unsupervised learning since it involves a small number of labeled samples and a large number of unlabeled samples.
The main goal of this approach is to train a classifier from both labeled and unlabeled data. Such learning has more advantage compared to supervised learning because it achieves better performance by utilizing both types of data (labeled and unlabeled).
Semi-supervised learning are applied in studies such as removing noise images happened dues to noise corruption in digital images through developing model that uses a devised cost function on a vast amount of corrupted image [44] and others areas like, real-time network traffic classification, text classification and others field of studies [45].

3.2.3 *Unsupervised Learning*

Unsupervised machine learning, is a method of analyzing and clustering unlabeled datasets using machine learning algorithms. Without any need for human participation, these algorithms find hidden patterns or data groupings. Because of its capacity to find similarities and contrasts in data, it's perfect for exploratory data analysis, cross-selling techniques, consumer segmentation, and picture identification [39, 46]. The approach in which the algorithm learns by itself and discovers an impressive structure in the data. In case, the output is unknown and only the input variable at hand. The purpose of this technique is to translate the underlying distribution in the data to gain more knowledge about the data. The example of these categories are:

Clustering: In which the input variables with similar aspects are belonging together. For example grouping customers based on their call or data usage history and group fraudster based on committing pattern.
Association: Computations are rule-based in determining the relationship among input or variables and to make predictions. For example, fraudsters, who do roaming call fraud will also commit mobile data fraud.[39], [46].
Pattern Recognition: Computations are used to provide a description or label to input data, such as in classification. Each input is evaluated against a pattern identified. This can also be used for supervised learning [16], [39]. This types of machine learning were applied in many flied of study of fields like [15] to identify the user model that best identifies fraud cases and network traffic classification using Kmeans and expectation maximization (EM) algorithm to cluster the network traffic application based on similarity between them overcome the drawback of port based classification and K-means have resulted higher accuracy for single traffic class and then EM for all class of application [47]. Such learning also adapted in the field of optical communication networks [48] and applied in many others field of studies.

3.2.4 *Reinforcement Learning*

Reinforcement learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment [49]. The agent receives positive feedback for each good action, and negative feedback or a penalty for each wrong

action. Unlike supervised learning, the agent learns naturally utilizing feedbacks and no labeled data in Reinforcement Learning. Reinforcement Learning is made up of four basic components: policy, reward signal, value function, and environment model [50].
An agent is set up for performing an individual task without any instruction. To achieve this final goal, that agent performs a trial and error through interaction with that specific environment. That agent then receives a reward value which indicates the quality of the performance. Consequently, the agent updates its strategy immediately based on the rewards and focus on maximizing the reward value [51].
Reinforcement based on a model learning optimal behavior is accomplished indirectly by constructing a model of the world by doing actions and seeing the results, which include the next state and the immediate reward. Because there is no response provided, such learning differs from supervised learning in that the reinforcement agent determines the steps to complete a task [50], [16].
Some articles such as "Analysis of Network Intrusion Detection System (IDS) with Machine Learning Algorithms (Deep Reinforcement Learning Algorithm)" case in which deep Q network intrusions detection model were built and its accuracy is evaluated . The goal is to detect different types of attacks at its first attempt with new ways of study. As stated in [51], they were achieved improved accuracy and intrusion detection system. Similarly, adaptive neural networks that is capable of autonomously learning new attacks rapidly through the use of a modified reinforcement learning method that uses feedback from the protected system is applied in [52] to detection of next generation network intrusion. It is also applied in areas such as financial fraud detection and others.

3.3 MACHINE LEARNING ALGORITHMS OR CLASSIFIERS

Machine-learning algorithms can effectively classify complex data-sets of two-biclass or more multi-class data-sets through the classifier or algorithm that maps the input data to a specific category. For this study, three machine learning algorithms-classifiers are selected to compare and recommend one with better resulted in classification parallel to selecting a better modeling approach in detection of mobile data roaming fraud. Artificial Neural Network (ANN), Support Vector Machine (SVM), and J48 decision tree (J48 DT) are the selected classifiers. Further details are discussed as the following based on [13], [16], [39].

3.3.1 *Support Vector Machine (SVM))*

Support Vector Machine is a discriminative supervised machine learning approach with given labeled data to categorize the classes [53],[54]. SVM is a new technique suitable for binary and multi-class classification tasks in addition to a new promising non-linear, non-parametric classification technique [16]. It is the state of the art classification and regression algorithm and optimization procedure to maximize predictive accuracy while automatically avoiding over-fitting the training data. It draws hyperplane in multidimensional space, which segregates data based on their classes [53].

One advantage SVM classifier uses only a small subset of the total training set for classification, thus reducing the computational complexities through the use of kernel trick and over-fitting of data is avoided by classifying with a maximum margin [55]. On other side SVM often involves higher time to train the model. SVM successful in many areas of fraud detection in such as subscription fraud in telecommunication based on the subscriber's usage behavior [16], and intrusion detection [55]. Figure 3.1 shows the simple architecture of SVM.

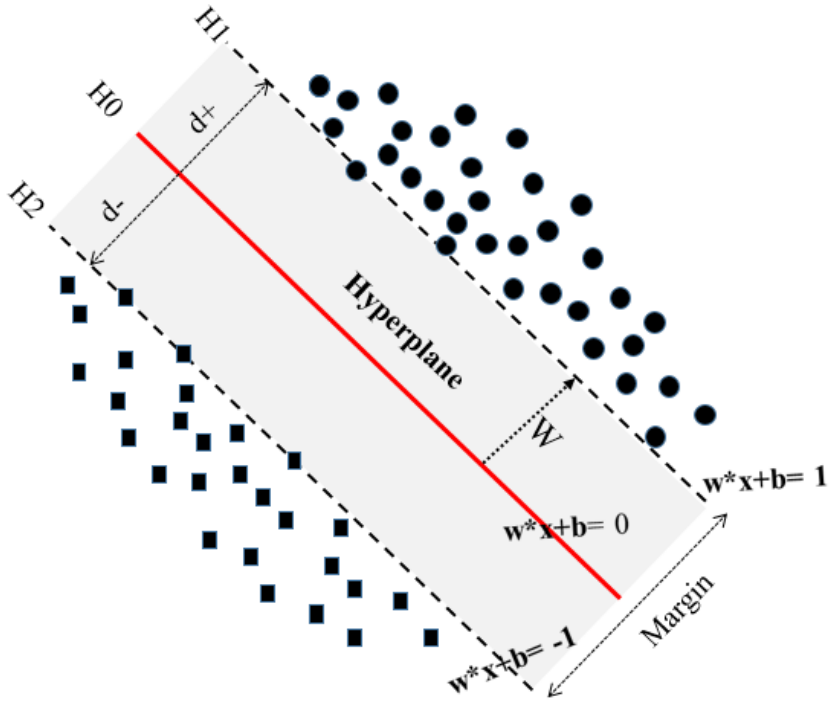


Figure 3.1: Support Vector Machine Architecture Based on [55]

.

Line or hyperplane, classifies the classes at random. Support vectors are input vectors that simply hit the margins H1 and H2. The training stage is used to define the region (border) where data is classified as normal in this approach. The occurrences are compared to that region during the testing step, and if they fall within the delimited region, they are classified as normal; otherwise, they are categorized as fraud. The main goal is to keep points in the margin as small as possible [16], [55] . Equation 3.1 defines the linear kernel of SVM function-decision surface separating the classes.

$$y(x) = w_1.x_{j1} + w2.x_{j2} + w3.x_{j3} + ... + w_n.x_{jn} + b \tag{3.1}$$

Where y(x) indicates a function that is linearly discriminant, x represented the feature vector chosen for classification, w is hyperplane's place space and b denotes the space bias that controls hyperplane location.

Maximizing the Margin: The hyperplane and nearest data point are separated by the maximum perpendicular distance. The function (hyperplane) that gives the most least distance to the data points is found by using margin SVM techniques. The data closest to the margins is referred to as support vectors, and the distance between them is called a margin. Referring to figure 3.4, the support vectors are the points set on the margin lines, and the distance between these margin lines is the margin width. The remaining variables are not significant in creating the model because the solution is only dependent on the support vectors. The tips of the support vectors are the places on the planes H1 and H2. The plane Ho is the midpoint plane, with $wx_1 + b = 0$. The margin of hyperplane is discussed in equation 3.2.

$$d + d - or(\frac{(1-b)+(1+b)}{||w||} = \frac{2}{|w||} \tag{3.2}$$

Where d+ is the shortest distance to the closest positive point and d- is the shortest distance to the closest negative point. As a result, the margin is maximized while some measure of loss on the training data is minimized and the optimization problem for computing w and b like equation 3.3.

$$min_w \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} E_i \tag{3.3}$$

Ei is errors, capacity (C) is a tuning parameter. As stated in equation (3.12), C is a tuning parameter that regulates the generalization ability of an SVM

by weighting classification errors. The higher C, the more weight is given to in-sample abnormal classification and the worse the machine's generalization. Such generalization means that while the machine looked impressive on the training set, it would fail poorly on a new sample.

3.3.2 *Artificial Neural Network*

Artificial Neural network (ANN) stands as a computing system, which consists of highly organized elements called nodes, which is known as neurons. The neurons are organized in multiple layers with each layer receiving inputs from previous layers, and passing outputs to further layer[39]. The weight assigned to a certain link, which is determined by the cost function and the optimizer, determines how each layer output becomes the input for the next layer. The epochs are the number of times a neuron iterates. The cost function is evaluated after each epoch to determine where the model can be improved. Based on the information provided by the cost function, the optimizing function then changes the internal mechanics of the network, such as weights and biases, until the cost function is minimized [39], [56]. Such a machine learning (ML) model becomes a famous and applicable model for classification, clustering, pattern recognition, and prediction by learning.
One good advantage of ANNs application is that it can make models easy to use and more accurate from complex natural systems with large inputs. In another way, the complexity to design is comparatively slow due to it's number of hidden layers [39], [57].
There are types of neural networks. The categories depends on their data flows, layers, and depth activation filters. Based on the data flow, it can be categorized as discussed under figure 3.2.

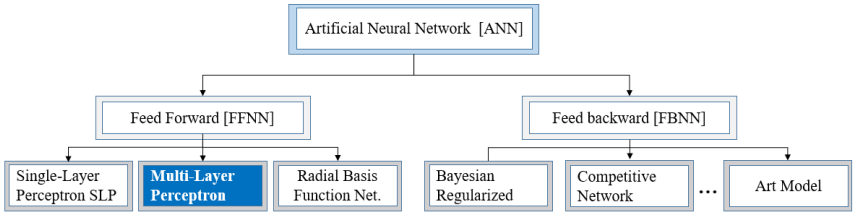


Figure 3.2: Classification of Artificial Neural Network [57]

.

A feed-backward neural network (FBNN): it uses internal state mean memory or store information to process a sequence of data inputs. The Feed-backward neural network application areas include mathematical proofs, data fitting, engineering, time-series prediction, classification such as fraud detection and prediction problems.
Recurrent neural network (RNN) that feeds into the next time step rather than feeding into the next layer concurrent time of step is an example of FBNN. It is dynamical networks with the recurring path of synaptic connections help as monitoring time-dependent problems [16], [39], [57].

A feed-forward neural network (FFNN): A types of ANN classification algorithm that contain organized layers and each unit in a layer relates to all the other units in the layers. The layers' connections with units are not all equal because each connection can have a different weight or strength. The weights of the network connections measure the potential amount of knowledge of the network. Single Layer Perception (SLP), which is a simple form of the neural network, and Multi-layer Perception (MLP) that has dense fully connected layers used for deep learning are the two examples of FFNN.

Single Layer Perceptron (SLP): The smallest unit of a neural network performs specific functions to detect characteristics or business analysis in the incoming data. Nodes in the input layer are fully linked to a node or a group of nodes in the output layer. Hence, Input layers and output layers are the only layers in a single perceptron. Activation functions are applied to weighted inputs, and the output is obtained as a result. The following layer's node determines a weighted average of all of its inputs [58], [16]. The simple diagram of SLP is illustrated in figure 3.3.

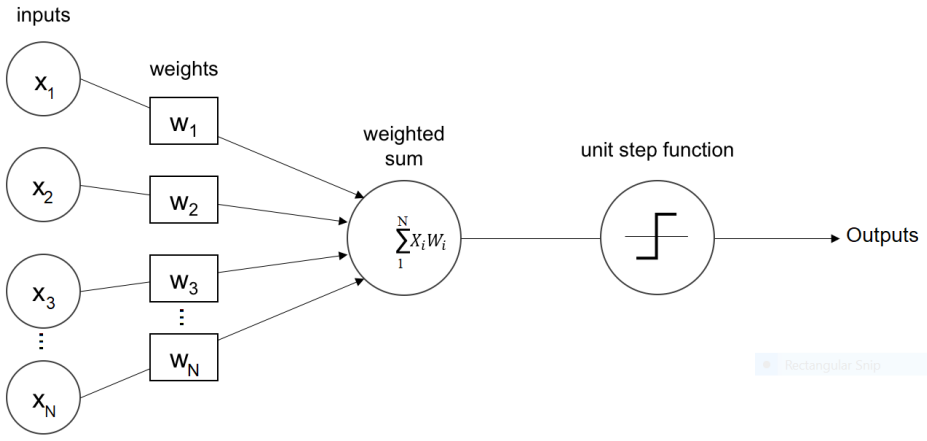


Figure 3.3: Simple Architecture of ANN [39], [16].

X_N, Reflects the network layer's feature inputs.
W_N, The weight assigned to each input and is multiplied by a connection weight create the output.
The architecture of SLP contains:
Input layer: The input layer represents the dimensions of the input vector and takes the values.
Output layer: Represents the output of the neural network.
Weights: Weights are numeric values which are multiplied with inputs or machine learnt values from ANN. It's self-adjust depending on the difference between predicted outputs vs training inputs. In back-propagation, they are modified to reduce the loss [58], [16].
Single perceptron is a supervised learning model that is used to undertake binary classification. It divides the input space into two categories using an hyperplane equation 3.4.

$$Y = f(\sum_{N=1}^{N}(X_1W_1 + X_2W_2 + X_3W_3...X_NW_N)) \tag{3.4}$$

Y Represents the output, f denotes function (sigmoid), X_N input (features) and W_N reflects weight of bias.

Multi Layer Perceptron (MLP): Unlike SLP, the multi layer perceptron has one or more hidden layers. An introduction to sophisticated neural networks, in which input data is routed through many layers of artificial neurons. It is a fully linked neural network since every node is connected to all neurons in the next layer. Figure 3.4 shows a basic MLP architecture.

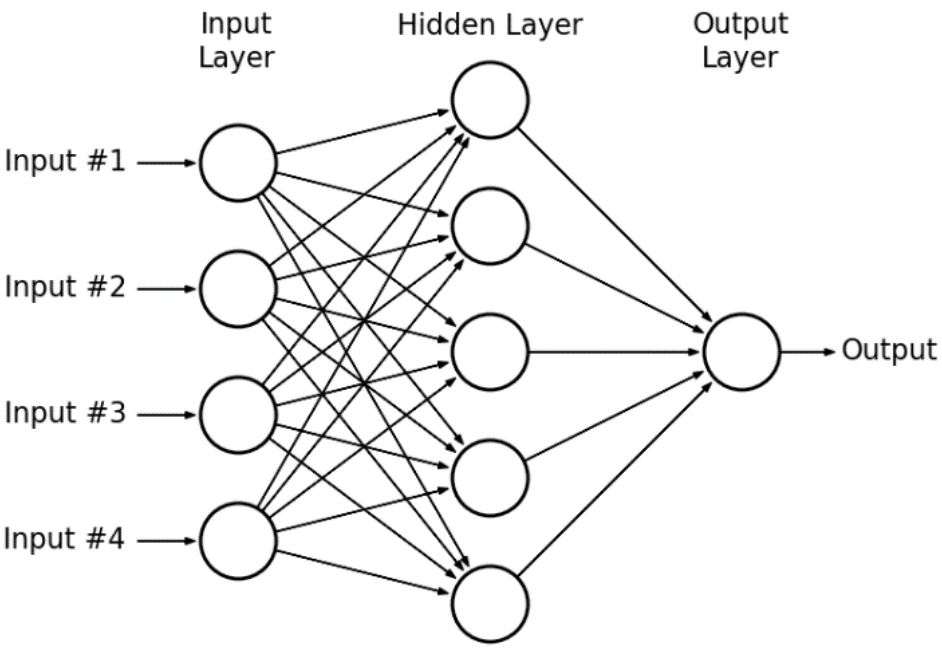


Figure 3.4: MPL Architecture of ANN [39], [16].

MPL architecture includes hidden layer, in addition to the SLP shown in figure 3.2:
Hidden layer: An intermediary nodes that divide the input space into regions with boundaries.It contain a set of weighted used to identify the relationship between successive inputs, and produces output through an activation function. The layers make the network faster and efficient by identifying only the important information from the inputs and leaving out the redundant information [59], [16].

MLP has the capability to solve problems that are not linearly separable. It is widely utilized to solve supervised learning challenges. The FFNN is a system that transforms a set of inputs into a set of output is, a directed graph that connects multiple layers in a single-direction neural network. The MLP is commonly used for pattern recognition, classification, prediction, optimization, control, time series modeling, and data mining. The presence of dense completely linked layers and back-propagation is another advantage of MLS, which is commonly employed in deep learning. Increasing the number of hidden layers, on the other hand, increased the design and maintenance complexity.

When training a neural network, the optimal weight for the edges between all of the network's units is taken into account. Its advantages over other statistical methods include high accuracy, noise tolerance, independence

from prior information, and ease of maintenance, while the ability to be implemented in parallel hardware, minimal human intervention, and suitability to be implemented in non-conservative domains are some benefits of classifying with ANN features [58], [16]. Although it is not robust due to factors such as low transparency, trial-and-error design, data hungriness and dependency on data quality, overfitting, a lack of a clear set of principles for picking an appropriate neural network, and a lack of classical statistical features [55]

3.3.3 *J48 Decision Tree*

J48 decision three is a non-parametric supervised learning method used for classification and regression. It follows a top-down approach in which data recursively splitting into smaller mutually exclusive subsets. Root node, intermediate branches, and leaf nodes are parts the tree. A decision tree before starting usually considers the entire data as a root. It starts splitting by means of branches or intermediate nodes and makes a decision until it produces the outcome as a leaf and reduces impurity present in the attributes of data and simultaneously gains information to achieve the proper outcomes while building a tree.

Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. One of the advantages of a J48 decision tree is that it produces unambiguous findings when the data is largely categorical and conditional. However, if complications occur in the computation, outcomes are connected and training a model can take longer.
There are certain parameters help to decide how well a decision tree performs during the final building of a model [58], [16], [60].

Entropy(E): It's a measure for the amount of impurity in the data set. When the sample approaches homogeneity, the value is close to zero, but when it divided in balance, the value tends to one. Because it segregates the classes better, entropy with the lowest value makes a model better in terms of prediction. Entropy has a greatest value in the middle (up to 1) and a minimum value at the ends (up to 0). Entropy is computed as equation 3.5.

$$E(s) = \sum_{i=1}^{n} P(c_i) \log_2 P(c_i) : \tag{3.5}$$

*W*here Pci is probability of class ci in node

Information Gain(IG): It is the measure of change in the entropy after the dataset is split based on an attribute. The decision tree always tries to maximize the information gain and the node means that the attribute having the highest information gain is split first. An event with low probabilities to occur gives lower entropy and high information knowledge and vise versa. The idea of information gain as computes as equation 3.6.

$$IG = E(s)_{parent} - E(s)_{Avg(childs)} \tag{3.6}$$

Gini Impurity(GI): It is a measure of the impurity/purity or misclassification used while building a decision tree in the algorithm with multi-class labels data. The Gini Impurity is equated as equation 3.7.

$$GI = 1 - \sum_{i=1}^{n} (P_i)^2 \tag{3.7}$$

The three steps of J48 DT follows to perform the classification are discussed below.
Step 1: The leaf is labeled with the same class if the instances belong to the same class.
Step 2: The potential information is computed by using entropy, which is a measure of the data disorder for each attribute and the gain in information taken from it.
Step 3: Lastly, attribute with high values will be assigned for root.
Nodes with an entropy of zero are thought of as leaf nodes whereas nodes with entropy higher than zero fragmented until entropy vaue getting zero.

Easy handling of missed attribute values of data, easily remove over fitting or tuning, as well as minimized error pruning and increasing precision through pruning are some advantages of J48 DT.

3.4 MACHINE LEARNING TOOLS

Machine learning tools are algorithmic applications of intelligence that enable systems to learn and develop without need for human interference. Experts in machine learning used a number of tools and approaches in a variety of contexts to create a high-quality model [39]. Prior to narrowing our emphasis to one tool utilized in this study, we discussed the top three machine learning tools, as well as their specialized applications and benefits based on various studies including predictive modeling, to have better understand of their fitting domains.

3.4.1 *R Programming Studio*

R Programming Studio is one of frequently used open-source and programming language tools for statistics, visualizations, and data analysis, and it's one of the most frequently used in data science. [61]. Classic statistical tests, linear and nonlinear modeling, time-series analysis, classification, and clustering are only a few function R programming tool [62].
The language enables for the creation of high-quality plots and is the ideal option for analysis or visualization because it allows for rapid prototyping and works with datasets to create machine learning models [63]. In a comparison of five data analysis tools namely, Python, Statistical Package for the Social Sciences (SPSS), R-language, Statistical Analysis System (SAS), and WEKA conducted in [64], based on the state that, R language is taking the best analysis tool among all and benefited for novel method prediction. However, it is a challenging language [63].
In the R language, the dplyr package provides tools for subsetting, summarizing, reordering, and merging data sets. The mlr platform, which includes classification and regression algorithms for data processing and analysis, is another set of popular R-based tools.The tidymodels-collection of packages for modeling, and the tidyverse principles-collections includes resample, parsnip, recipes and room are some libraries in R language used to model data [65].

3.4.2 *Waikato Environment for Knowledge Analysis (WEKA)*

WEKA is a Java-based, open-source Data Mining (DM) platform developed at the University of Waikato. It contain a collection of packages for machine learning which can widely adopted in academic and business problems [63]. It provides a user interface such as exploratory data analysis to support different data pre-processing, an experimental environment for evaluating machine learning algorithms, and the flow knowledge for new process model and suitable for developing new machine learning schemes. Several authors such as [16], [18] have used the tool for data analysis.
Similarly, for comparison conducted in [63], [66] it is better performed next to python. Moreover, it acts as the best data mining tools in classification of for defective software modules and non-defective modules conducted in [67] with Decision Tree and Logistic Regression algorithms classifier.

3.4.3 *Python Programming Language*

Python is an open-source programming language that provides large number of standard library for machine learning algorithms [68]. The built-in libraries assist in the development of useful models that perform effectively in solving business challenges. It is fast, powerful, easily extendable and simple [64]. Python is quite popular nowadays, and employed in a lot of machine learning-based data processing.

The experimented survey were conducted among MatLab, R, and Python to identify which language can be easily understood by students in [68]. The survey were addressed top thirty-nine (39) computer science departments in the United States. As a result, Python is simplest for students to understand and write for the projects they want to achieve, whereas MatLab is the runner-up.
Recently, researchers in [69] used python libraries such as Keras, which is built on TensorFlow and Scikit-Learn, to detect long short-term memory based classifiers for wireless intrusion a detection system (WIDS) and achieved an accuracy of more than 99% with ANN algorithm.

Similarly, python libraries such as sci-kit-learn in [35] were used to detect wangiri fraud in telecommunications using collaborative learning with selected random forests, AdaBoost, and XBoost algorism, with the performance of the XBoost algorism scoring more than 99 percent in accuracy, precision, recall, ROC, and F1 measures.
Similarly, in [70], from Jupyter Notebook (Anaconda)-python open source, a package of libraries was utilized for feature selection in the modeling of network intrusion detection with multiple optimization strategies using J48 and SVM machine learning classifiers.
According to a different rule, reducing the amount of features in an intrusion detection system improves accuracy, and the J48 DT classifier achieved over 90% accuracy.
The same used in the area of medical to predict cardiac disease [71]. A survey undertaken in this case is to compare the performance of machine learning algorithms such as Nave Bayes (NB), SVM, and Decision Trees such as J48 DT, Random Forest (RF), and K-Nearest Neighbor (KNN). As a result of their triumphs, the researchers prefers to use python programming code in cooperation with the machine learning techniques that have been proposed.

To be sure, we compare the described tools with data followed the same approach of this study, and as the result a python programming tool achieved

better performance comparatively based on the same parameters.
On the anaconda3 navigator Jupyter Notebook, the python programming language provide useful packages and libraries for data pre-processing and classification with function help to importing data, data processing and computing evaluation metrics as well as plotting and visualizing results. The mainly packages and libraries used in this thesis study are detailed further below [72–76].

Numpy: It is python library provides us fundamental packages for data preprocessing such such as scientific calculations like removing inconvenient data from the data-sets, merging data-sets for example concatenating data-sets, slicing fields, extracting samples of data such fetching fraudulent numbers from non-fraud data, and bug fix. The handling of files in any encoding is also possible using this library in Python. Importing Numpy as np is part of the library's import statement, and we were using np throughout the program.

Pandas: It a python library frequently used in data processing and analysis tools.In this study it looked for importing and handling data-sets using extendable functions like pandas import read_csv() function. One of the key benefits of the Pandas library is it's ability to translate complex operations of data using one or two instructions. It is introduced as import pandas as pd, and used pd as a simplification of library through this study.

Scikit-learn: It is mostly efficient and comprehensive machine learning package in Python.It providing a variety of efficient tools for modeling including classification.
In this study, the LabelEncoder class from the sklearn.preprocessing package is used to successfully encode the variables into digits, and the train test split() function from sklearn.model selection to split the dataset into the training and test sets. This library was also used to import the classifier algorithms MLPClassifier from sklearn.neural network, DecisionTreeClassifier from sklearn.tree for J48 DT, and from sklearn.svm for SVM. As part of the feature selection, ExtraTreesClassifier class from sklearn.ensemble and the mutual-info-classif, SelectKBest and chi2 functions classes from the sklearn.feature-selection libraries are imported to complete this part. For evaluation the models, sklearn library for model evaluation metrics like classification_report, confusion_matrix, and accuracy_score from sklearn.metrics used to evaluate models. Additionally, make_classification function from sklearn.datasets is used in solving of data imbalancing problem faced. **Matplotlib:** A python 2D charting library, used to create charts required expressions such as illustrating of feature's correlation and feature importance,

verifying outliers in data-set for example using quantile approaches with boxplot function as well as figuring of results in bar chart.

Others built-in libraries from the Python version mentioned included, for example, RepeatedStratifiedKFold for model testing and validation, VarianceThreshold to set the threshold used to remove fields with high duplication, pipeline for count weighted training data-sets between classes when balancing, and SMOTEoversampler for re-sampling data-sets.

4

DATA PREPARATION

The most valuable resource in every business is it's data. However, if we do not look deeper into the concept, it has the potential to neutralize itself.

Data from industries such as ethio telecom used for a variety of purposes example to make well-informed business decisions, for effective management of traffic flow, customers churn monitoring, fraud detection and so on. These, on the other hand, cannot be done just on the basis of raw data. The fraud detection model development process includes several steps, from data collecting to model building. Nevertheless, it is always a good idea to make a hypothesis prior diving into the data as well as trying to figure out the relationships between variables. As a result of these practices, better features can be built that not really influenced by rather data present in the data-set. This is a big determinant which enhances the accuracy of a model.

4.1 PROBLEM IDENTIFICATION

Annual reports for the previous three years that are related with revenue loss due to roaming fraud are revised in order to recognize the reality of the problem in the sector. Informal interviews are conducted with domain experts from various departments of an organization's ITsec-division that is concerned with security in the organization to validate the presence of a problem and it's potential to damage business. This is helpful contribution to follow better approaching in solving the problem.

4.2 LITERATURE REVIEW

The available of state-of-the-art and accessible supportive materials particularly literature related the problem with their proposed detection and preventive approaches are revised. This enable us to get unbiased picture

of the challenges and provide possible fraud detection methods. The domain experts' such as analysts, specialists, and supervisors opinions are also included to enhance the carrier of study.

4.3 DATA COLLECTION

To succeed in this study through overcoming the prevailing problems, the mobile data roaming subscriber's usages or CDR and fraud data are collected from ethio telecom systems for further insight into their behaviour. The data were collected in every two hours by fraud management experts using a familiar CDR exploration script.
All the collected data are Comma-Separated Values (CSV) format- a file format which allows us to save the tabular data, such as spreadsheet and particularity much useful for large data-sets.

First, we examine the data used to develop a model that used to detect roaming out fraud, which we refer to as model_1. We collected the data from the international system division (ISD) switch because this data can be accessed from a local switch—in this case, ethio telecom. The international business division is an interface of an organization that handles international usages including roaming warnings such as high usage while roaming.

In this case we consider 12 months of consumption from June first 2020 to May end 2021 with total of 295,446 records. Considering much month's usage in past may handle variety of user's behaviours increase the model's quality. In addition to the CDR data, the corresponding month's fraud number-single field contains service numbers stored after fraud analysts identified as fraud usage are collected. The total number of unique rows in this data is 1536, and we classified it as fraud for further labeling.
With these fraud records in hand, the corresponding CDR data for fraud numbers are extracted from the entire CDR data in mentioned months. In this process, pandas (discussed earlier in python libraries)-filter data frame method on anaconda3 navigator version 2.0.4 of jupyter notebook 6.0.1 version, and base root environment is used and possible to find 26,701 fraud-related CDR data for discussed months.Following this separation, the previous entire CDR is reduced to 268,745, which is tagged as normal or non-fraud. Finally, we have labeled data frames related to the prediction of roaming out fraud.

The second (model_2) is intended to detect roaming in fraud. In this case, the mobile data roaming CDR from roaming in users are looked in. This CDR can also be found on visiting mobile networks-in this case ethio telecom. The visited network operators can assess the usages of roaming in addition to sending TAP files to a clearing house which should passed to the home operator for billing purpose. The three months CDR data (January, February, and April) of 2021 is considered to develop this model. The CDR is generated from switches managed by the international system division (ISD).
The number of roaming-in subscribers and their monthly traffic of mobile data is very high compared to roaming out subscribers' usage. This may be for unmatured WiFi everywhere in the country. We hope that it will be balanced in the near coming years for the speedy growth of ethio telecom both in service quality and coverage.
Accordingly, 3,138,055 data CDR from three months in concern are collected.
A similar approach from the same system in roaming out fraud number collection is used to acquire related fraud numbers and can find 527 numbers over the three months mentioned.
From the entire CDR data, the corresponding usage behaviors of fraudster are extracted and the CDR are letting classified or labeled into fraud and non-fraud data frames. And accordingly, 32,068 fraud-related CDR are extracted. This usage behaviors are termed as fraud, while the rest are marked as normal or non-fraud usage. The number of normal data is reduced 3,105,987 due to the separation of fraud information.

In order to build third (model_3) from two data-sets, the shared fields appear common for both data-sets (roaming our and in) are discovered. This section is fully described in Chapter 5 of the experimental setup subsection.

4.3.1 *Data Understanding*

Before going through more data preprocessing steps, data is reviewed together with the fraud management specialists to incorporate their unique quality, which is then used to get insight into essential fields that important for attribution. This is a significant task to develop quality models.
In addition to the expert's feedback, the sample data before gathering data for model formulation were reviewed, and preliminary models were investigated.

The roaming out data contains 262 columns whereas roaming in usage holds 171 at all. As a result of the previewed sample data, we found data that is irrelevant to the study's goals. Some fields are exceptional-empty by nature whereas a few are hold homogeneous data-all values in the columns are similar. As a result, not all of these fields have well-suited contribution in the creation of models and these challenges us to focus on some of the important fields-columns contain variety values. After some investigations, 42 columns are included from the data fields discussed. It also does not mean that these data are good and clean enough to build a model from. Furthermore, it is required to adapt to the various levels of data preparation to verify the data completeness, redundancy checking, missing elements handling, and attribute values' sensibility, and others to meet the ML objectives.

4.4 DATA PREPROCESSING

When learning from data, algorithms use analytical equations that deal with values from the data-sets. "If garbage goes in, garbage comes out," as the general goes. According to this reality, data projects can only be effective if the data used in machine learning is at good quality. In the experience of this study, the concept from experts [77] and [78]. The data analyzing related project is about data preparation. Data scientists spend about 80% of their time on preparing and managing data and from this, data preprocessing shares about 60%.
This section discusses data cleaning, transforming, aggregation techniques as well as feature engineerings such as feature selection and data balancing techniques by using predefined python libraries discussed under tools for machine learning-python programming subsection of chapter three.

4.4.1 *Data Cleaning*

To develop advanced models-intended optimal detection range, it is important to work more on data to make it suitable for machine learning. The two focused important areas in this section are: error detection-identified potential unreliability occurs in a way to machine learning accepted data-sets, and repairing errors-to make data-sets more acceptable through correcting likely errors in model creation.

After loading data into Python with the help of the Panda library, operations and activities discussed below are carried out through data.

Cleaning Duplicates: The zero-variance predictors are input features that have homogeneous data across the entire column of observations. As a result, they contribute no value to the prediction process because the target variable is unaffected by the input value presenting them redundant.
With received roaming CDRs data, columns such as "HMANAGER" roaming home manager-located to the home network for roaming-out, "CDR_TYPE"-types of subscriber usage presented in both data frames (roaming in and out), "SERVICE_TYPE"-types of service used, "DEST_CURRENCY" -a type of currency should be paid for the usage across operators base on the roaming agreement are stored with similar values in the column entirely. Therefore, the columns with such value dropped out from entire data using the Panda drop function- a single line of code and store the modified frames; this process is also working for all data frames referenced in this study.

Handling Missing Values: In python, the data frame with missed value is represented by NaN. Later checking of NaN value exiting in each column and records by using panda's library such isna() function. Mainly such values are treated in two ways: removing columns and rows with NaN values greater than 60%-based on data analyzing basics by using dropna function from panda with setting the threshold at a specified value. Accordingly, the number of columns was reduced to 22 fields. Similarly, the records with less than the defined threshold value are treated by substituting the values with one (1) by using the fillna function from the options presented for NaN value handling and that makes sense for subsequent steps.

Outlier Handling: Outliers are data points that appear differ significantly from other observations in a data-sets. The purpose of outlier detection is to find useful abnormal or irregular patterns masked in large data-sets.
To find outliers in an input feature, we applied Inter-Quartile Range (IQR) technique-each data-set is divided into quartiles(upper quartile and lower quartile). It defines data range as Q1–1.5*IQR for the lower limit and Q3+1.5*IQR for the upper limit. Any data point that falls outside of the range is regarded as an outlier and should need further analysis or treatment. The advantages of IQR are it can easily be visualized on boxplot-python box plotting function and is also not affected by extreme value in data like others such as variance outlier detection technique.
Through plotting outlier values by using boxplot function from the Python Matplotlib library, the columns with an outlier, such as "TOTALBYTE", "TOTALBYTE UP", "TOTALBYTE DOWN", "GGSN", "DURATION" roam-

ing out, as well as others like "FLOWUP1", "FLOWDOWN1" from roaming in the data-set are detected. The cutoff values are computed for outliers in third and first quantile than 1.5 times IQR above and the 75th percentile and less. Then subtract this cut-off from the lower_bound then add it to the upper_bound. To repair data at least near to it's value, values out of ranges are replaced with corresponding limits. This is handled by using the replace_with_thresholds_iqr function.

4.4.2 *Data Transformation*

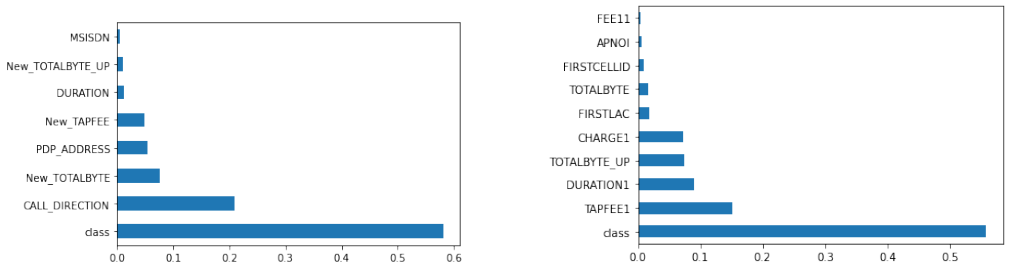
To make up our data fit for machine learning models, further data processing like scaling, integration, aggregations, feature selection, and balancing is required to build optimum model.
Data Scaling: Scaling is the technique of measuring and assigning numbers to independent variables by predetermining standards. The preprocessed data contain attributes with a mixture of character and numbers quantities measures such as kilobyte-subscriber usage volume, seconds-usage duration. To make it understandable for machine learning, these measures and other mixed data types are transformed into some scale one(1) and zero(0) by using label encoding function from sklearn library.

Data Instigation: Data integration refers to transforming features into single data frames. Our data sets were previously segregated into fraud and normal data frames, and we combined them into a single data frame with the peculiarity of maintaining target values by applying the concatenation function from Python's Panda package.

Data Aggregation: Data aggregation is the process of grouping and compiling subscriber's usage behavior. We use the panda's library groupby function to aggregate data based on the two-hour usage as discussed earlier. Accordingly, the records from roaming out and roaming in are shrunk to 170373 and 1584230, respectively.

Feature Selection: The process of finding a subset of significant attributes that most contribute to the prediction model. Irrelevant features in the data can negatively affect the performance of the model for the algorithm trains from less or zero information contributor features.
Additionally, redundant variables lowers the model's decision capabilities and have a similar effect on the model's overall accuracy.
More variables in a model also increases its complexity. The purpose of feature selection in machine learning is to determine the best set of features

for building effective models for stated problem under study.
With the help of the seaborn library, we calculated correlations, which is a measure of the linear relationship between two or more variables, and visualized them using matplotlib in Python. The computed correlation matrix are attached with annex.
For two features are interrelated, the model only requires one of them, as the second one provides no extra information and thus we removed one of them from consideration. As a result, we eliminated features whose information is correlated with 85% or higher-basics of data analysis. Based on this, SelectKBest, a scikit-learn library, provides the most useful features.SelectKBest, a scikit-learn library, provides the optimal learning input features. The higher the score, the more important the feature is in predicting our target feature.
The top features selected from roaming out and merged data-sets are figured by figure 4.1a and figure 4.1b respectively.



(a) Selected Roaming Out Data-set

(b) Selected Feature Merged Data-set

Figure 4.1: Selected Features from roaming out and merged data-set

Data Balancing: An unbalanced data set is one that has skewed-observation in one class exceeds that of in other classes, when classification. Such classification occurs when the class distribution in a data-set is not relatively comparable. There are strong class biases with roaming in data as the detail was stated in data collection section. The ratio of the range is visualized in fig 4.2.

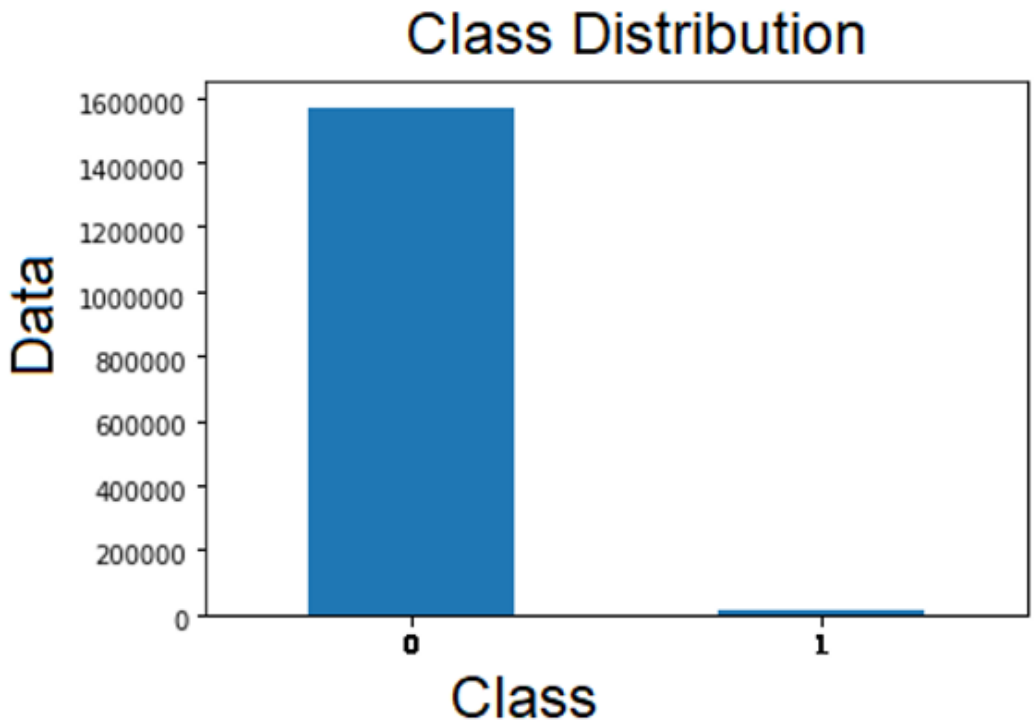


Figure 4.2: Roaming in Class Distribution.

When building the model before data balancing, we achieved a significantly higher accuracy predicting tending to the majority class and we fail to capture the minority compatibly.
To consider out both classes equivalently, Synthetic Minority Oversampling Technique (SMOTE) is applied to the training data-set. The SMOTE creates new training data based on the original training data. It chooses a minority class as the input vector and finds its neighbors (k nearest)-specified as an argument in the function. One of the selected neighbors placed a synthetic point anywhere point under consideration and the process is continued until data is balanced.
Such a method is advanced compared to simple random oversampling because it adds data variety in parallel to increase minority class.
As a result of this transformation, both the fraud and normal classes are predicted in a balanced manner. Model evaluation parameters such as recall, precision, and F1 score are used to clearly show this classification results.

4.4.3 *Cross Validation Techniques*

Validation Techniques: Cross-validation is an analytical method for determining machine learning model performance. After training the model, it

is important to determine how it really performs by using the testing data-set.
The hold out method used-which entire data-set is divided into: training data and testing data. We split 70 to 30 ratio for training and testing data respectively-based on machine learning for data analyzing basics. From sklearn Python machine learning library, the train_test_split function evaluation is used for the implementation. The weight of train-test split is illustrated in figure 4.3.

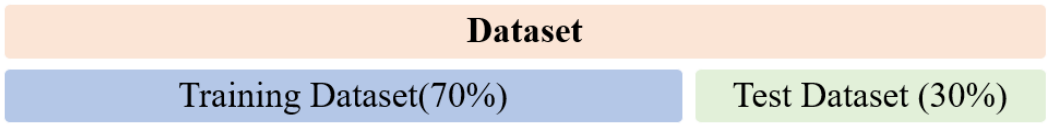


Figure 4.3: Train Test Data Split.

Training Data-set: When the classifier is evaluated, it is based on how well it predicts the class of the cases it is trained on, which is used to fit the model.
Test Data-set: The efficiency of the classifier algorithms is measured by how effectively they predict the type of a set of instances to determine how well a model fits. Figure 4.4 depicts the train test system model that is applied.

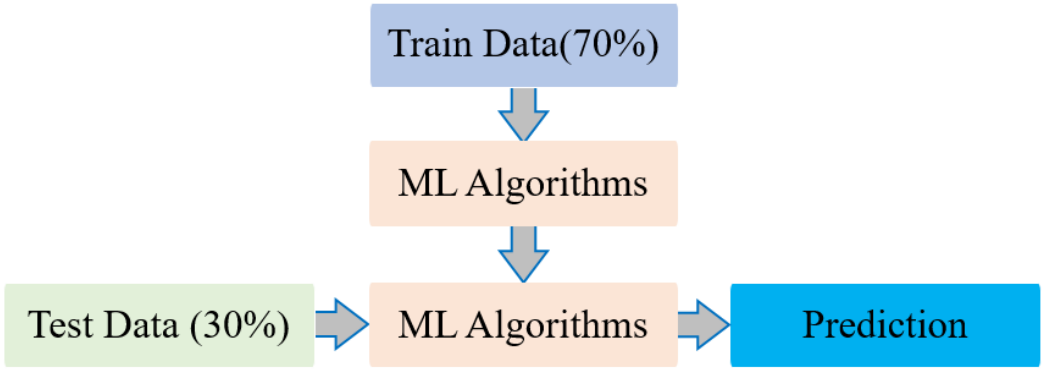


Figure 4.4: Train Test Supply System Model Based on [16].

4.4.4 *Algorithm Training*

A total of nine experiments performed with training and test data set options. For model_1, model_2, and model_3 mentioned earlier, 119261,

1108959, and 898633 instances are used to train the model using the fit function to fit the predictor and target, correspondingly. The function used differs ways depending on the classifiers' library built-in python. Based on the data clarifying techniques and methodology applied, the DecisionTreeClassifier from sklearn.model is the quickest and most well-trained classifier in this experiment.
To evaluate how well fit is predicted, 51112 instances for model_1, 475268 cases for model_2, and 385129 data for model_3 are utilized to forecast test data set.
The evaluation is based on various measurement metrics. The details of model evaluation metrics used for this study are discussed in section 4.5.

4.5 PERFORMANCE EVALUATION METRICS

It is important to measure how well our classification models predicts the desired outcome when building and optimizing it. The inclusive model evaluation parameters for classification are employed to get a complete picture of our models.

4.5.1 *Confusion Matrix*

Confusion matrix is a 2x2 matrix that describes the model's performance. It provides a more detailed view that includes not only the performance of a predictive model, but also which classes are predicted correctly and wrongly, as well as the types of errors produced. In our demonstration, we deals with the following four classification metrics.

- **True Positive(TP):**The cases in which we predicted positive and the actual output was also positive.
- **True Nagative(TN):** The cases in which we predicted Negative and the actual output was Negative.
- **False Positives(FP):** The cases in which we predicted Positive and the actual output was Negative.
- **False Negatives (FN):** The cases in which we predicted Negative and the actual output was Positive.
  The confusion matrix working principle is is clearly given in the con-

fusion matrix table 4.1. However this constraint can be computed in ROC evaluation parameter.

	Actual Positive	Actual Negative
Predicted Positive	True Positive(TP)	False Positive(FP)
Predicted Negative	False Positive(FP)	True Negative (TN)

Table 4.1: Confusion Matrix [16]

4.5.2 *Accuracy*

Accuracy is one of the metrics for evaluating classification models that depict single class accuracy measurement. It is a common performance measurement parameter in many machine learning comparisons and it is computed as the ratio of correctly predicted observation to the total observations (rate of total correct classification). Mathematically it is computed as equation 4.1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

However, if the data-sets are not balanced—both negative and positive classifications have varying numbers of data instances, accuracy may not be sufficient.

4.5.3 *Precision*

Precision is the positive predictive value or ability to only predict positive samples as positive. Precision considered as a measure that indicates us, how well our model performs in terms of false positives..The estimated precision value will be computed with equation 4.2.

$$\text{Presicion} = \frac{TP}{TP + FP} \tag{4.2}$$

4.5.4 *Recall*

The recall is known as the actual positive rate or the number of positive test samples that will be classified as positive. Recall, gives information about the performance with regards the false negatives. The recall value for the single class will be computed as the equation in 4.3.

$$Recall = \frac{TP}{TP + TN} \tag{4.3}$$

In our case scenario, it would show us the proportion of mobile roaming data-sets that report to the normal and that were predicted correctly by the model.

4.5.5 *F1_Score*

The weighted harmonic mean of the precision and recall of the test is the F1_Score (F-measure), which is a measurement of a test's accuracy. The F1_score is computed as equation 4.4.

$$F1\_Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4.4}$$

4.5.6 *Receiver Operator Characteristic (ROC)*

AUC - ROC is one of the critical evaluation parameter for evaluate the effectiveness of any classification model at various threshold settings. It indicates how well the model can distinguish between classes. The greater the AUC, the more accurate the model is at predicting negative classes as negative and positive classes as positive. An exceptional model has an AUC close to one, implying that it has a high level of separability. A poor model has an AUC close to zero, showing that it has the weakest measure of separability.

5

RESULT AND DISCUSSION

This section presents the details of the experiments carried out for each model in the building as well as comparative analysis among classifiers and models.
The experiments are carried out on selected classifier algorithms (J48, ANN, and SVM)-discussed under chapter_3-ML algorithms subsection. As a consequence of experimentation with every possible combination of settings, a total of nine (9) models were created.
The implementation is conducted on the anaconda3 jupyter notebook python environment which provides several packages for experimentation. The details are discussed in the chapter_3 machine learning tools section.
The model evaluation parameters discussed before are applied in the evaluation of each classifier's performance as well as models.

5.1 RESULTS COMPARISON OF ALGORITHMS AND MODELS

It is advisable to measure performance evaluation for the specified classifiers to determine the more suitable algorithm for the detection of the stated problem.
Furthermore, performance comparison for proposed approaches should be carried out in order to suggest one with a good outcome. For each modeling technique, the analysis is conducted on particular algorithms.

5.1.1 *Result Comparison of Algorithms*

The performance metrics that were used for comparisons are accuracy, precision, recall, F-Measure, Receiver Operator Characteristic (ROC). For this binary classification, a value of zero(0) represents non-fraud-normal usage of data, whereas a value of one (1) indicates fraud usage-irregular usage of data. However, for such an unbalanced data-set, it is challenging to classify both classes equivalently or nearly at the same level. Table 5.1 shows the

classifications performance results of the algorithms for Mode_1.

Classifiers	Accuracy	Target Class	Precision	Recall	F1_score	ROC
J48 DT	98.35%	Normal(0)	98.00%	99.00%	97.00%	98.70%
		Fraud(1)	100.00%	98.00%	99.00%	
		Weighted Average	97.00%	99.00%	98.00%	
ANN	98.29%	Normal(0)	95.00%	99.00%	97.00%	98.65%
		Fraud(1)	100.00%	98.00%	99.00%	
		Weighted Average	97.00%	99.00%	98.00%	
SVM	94.27%	Normal(0)	83.00%	99.00%	93.00%	95.70%
		Fraud(1)	100.00%	92.00%	96.00%	
		Weighted Average	92.00%	96.00%	93.00%	

Table 5.1: Model_1 with the Three Classifiers: Summarized Performance Results

According to the considered data-set in the building of model_1 as well as research methodology followed with selected algorithms and concerning evaluation metrics, a J48 DT classifier scores accuracy of 98.35% and ROC of 98.70%, which is better performance compared to the described classifiers. SVM, on the other hand, ranked lower on the criteria mentioned for accuracy and ROC, with scores of 94.27% and 95.70%, respectively. With an accuracy of 98.29% and ROC of 98.65%, ANN performs in the mid-range when comparing to the decision tree and SVM classifiers.
In respect of F1_score, which evaluates insight precision and recall, ANN was found similar to J48 DT, which achieved 98.0%. However, using such metrics from SVM, the classification result is 93.0%, which is significantly small compared to others. The variations in classification performance are due to the better classifier functionality composed from and data approach fitting with the technique relatively.

To illustrate our comparison criteria more comprehensively and concisely in that one can be described by another, we imply the F1 score as well as precision and recall- concerning chapter 4. Similarly, when we evaluate

ROC, on other hand, it does mean to evaluate the true positive rate and true negative rate, which optimizing is affect the model's accuracy positively. In addition, accuracy is taken into account to evaluate the model's capacity to classify the classes. Figure 5.1 provides a comprehensive summary of performance results relatively.

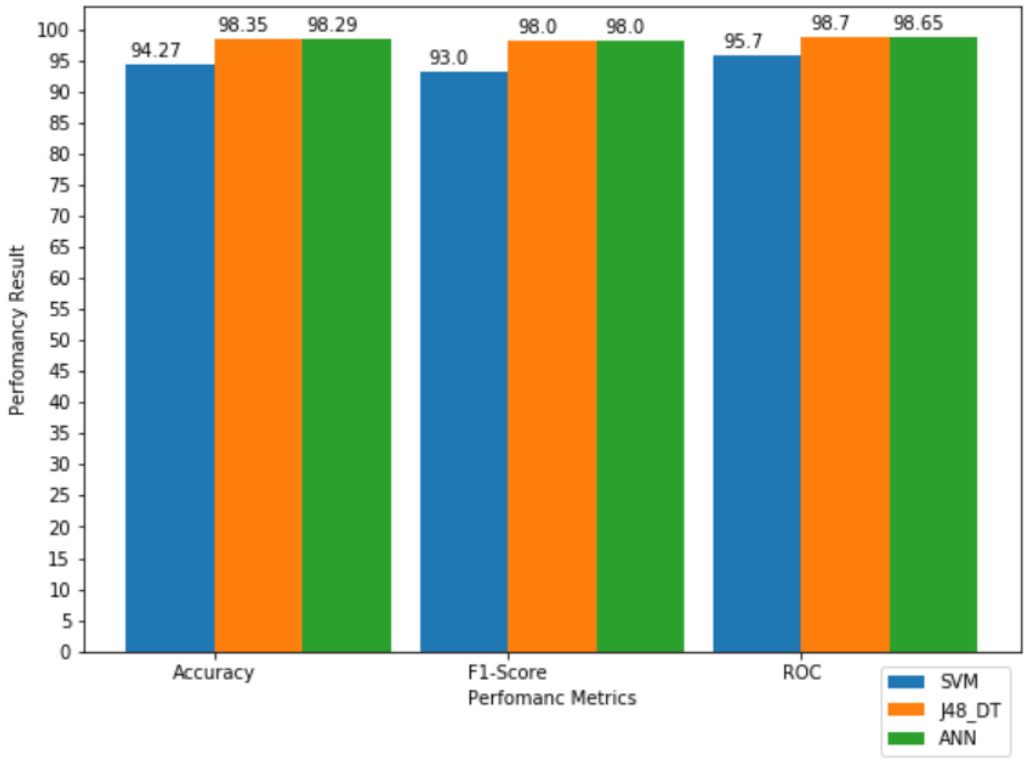


Figure 5.1: Model_1-Summary Comparison of Performance Results

To develop and evaluate Model_2, we supply data-set from roaming in with mounts of outlined and train-set split approach discussed under algorithm training subsection of chapter 4.
The same classifiers, approach and evaluation parameters from model_1 is used while building this model. Accordingly, the achieved results regarding to each algorithms and metrics are discussed in table 5.2.

Classifiers	Accuracy	Target	Precision	Recall	F1_score	ROC
J48 DT	99.24%	Norma(1)	98.0%	100.0%	97.0%	98.24%
		Fraud(0)	100.0%	98.0%	99.0%	
		Weighted Average	100.0%	100.0%	98.0%	
ANN	98.98%	Norma(1)	99.0%	100.0%	99.0%	97.79%
		Fraud(0)	100.0%	96.0%	98.0%	
		Weighted Average	99.0%	98.0%	99.0%	
SVM	98.29%	Norma(1)	97.0%	99.0%	98.0%	98.50%
		Fraud(0)	99.0%	98.0%	98.0%	
		Weighted Average	98.0%	99.0%	98.0%	

Table 5.2: Model_2 with the Three Classifiers: Summarized Performance Result

In general, all algorithms within consideration performed better than the performance given while analyzing model_1 with nearly all parameters in the perspective. When classifiers are analyzed separately, there are some difference in actual performance.
When compared to ANN, which has an accuracy of 98.98% and a ROC average weighted true positive and true negative of 97.7%, the J48 DT classifier performs well in this scenario, with an accuracy of 99.24%, and a ROC of 98.24%. When applying the accuracy measures, SVM results were nearly ranked lower than the other classifier algorithms presented, with a difference of 0.95 from J48 DT and 0.69 from ANN.

The SVM algorithm, on the other hand, rated 98.50% in ROC, which is higher than results from equivalent measures of J48 DT by 0.26 and ANN by 0.71. We could conclude from this that, although data training takes a while, the algorithm has a greater capacity to optimize true positive and true negative rate than other algorithm in comparison accordingly to this parameter.

In comparison, ANN has achieved an optimum result of 99% in the average of F1-score-consisting of precision and recall, whereas both J48 DT and ANN perform equal 98% in this case. This clearly shows that using the hidden layer function of ANN delivers nearly all predicted actual positive rates

that tend the model well predictor. Figure 5.2 illustrates a comprehensive and brief comparison of Model_2's classifier performance.

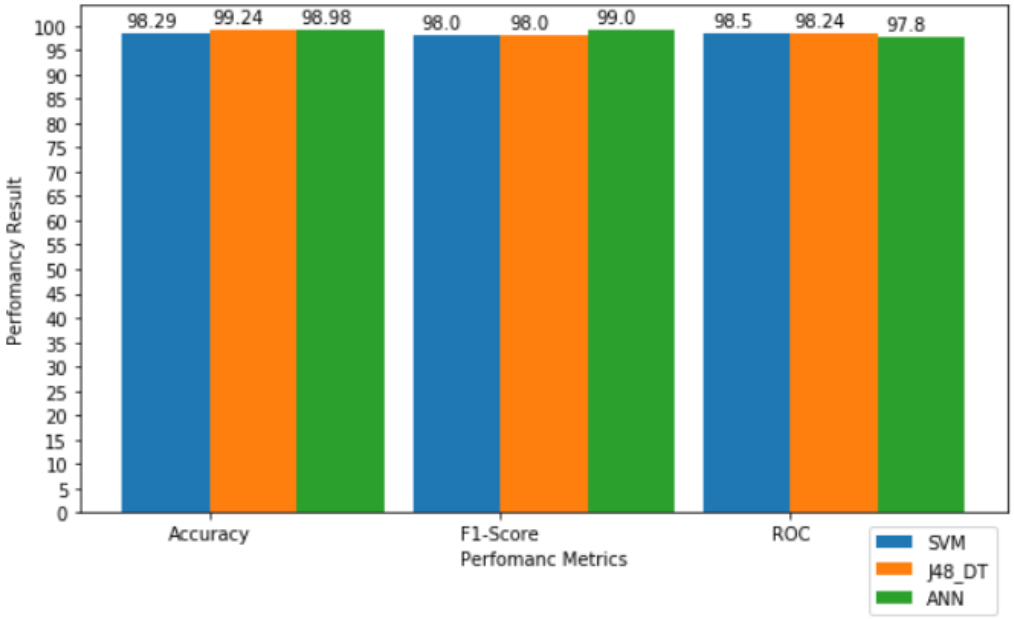


Figure 5.2: Model_2-Summary Comparison of Performance Result

Model_3 is built from shared features from roaming out and roaming in. The classifiers are fed based on the train-test split rate provided in the algorithm training section of chapter_4.
The evaluation techniques and the parameters used to measure the performance of the model follows the same steps in prior models evaluation principles. Accordingly, the detail achieved results are discussed in the table 5.3.

Classifiers	Accuracy	Target	Precision	Recall	F1_score	ROC
J48 DT	99.50%	Norma(1)	99.0%	100.0%	100.0%	99.30%
		Fraud(0)	100.0%	99.0%	99.0%	
		Weighted Average	100.0%	99.0%	99.0%	
ANN	99.41%	Norma(1)	98.0%	100.0%	99.0%	98.00%
		Fraud(0)	100.0%	96.0%	98.0%	
		Weighted Average	99.0%	98.0%	98.0%	
SVM	98.49%	Norma(1)	96.0%	100.0%	98.0%	98.70%
		Fraud(0)	100.0%	98.0%	98.0%	
		Weighted Average	98.0%	99.0%	98.0%	

Table 5.3: Model_3 with the Three Classifiers: Summarized Performance Result

Comparatively, the overall performance result of all classifiers are resulted better with near all metrics stated. When limited to a model, J48 DT performs better with the accuracy of 99.50%, average F1_Score 99% and ROC 99.30%. Although the result from ANN and SVM regarding these metrics are not neglected. ANN has performed next to J48 DT algorithm with accuracy resulted 99.41%. It is computed equivalent result in the average F1_score with SVM which is 98%. The SVM scored accuracy result 98.49% this is relatively ranked less compared to other classifiers.
On other hand, SVM acting middle in ROC with result of 98.70%, which is less ranked compared to J48 DT with resulted 99.30% and more than ANN resulted 98%. The summarized and inclusive comparison metrics are illustrated in figure 5.3.

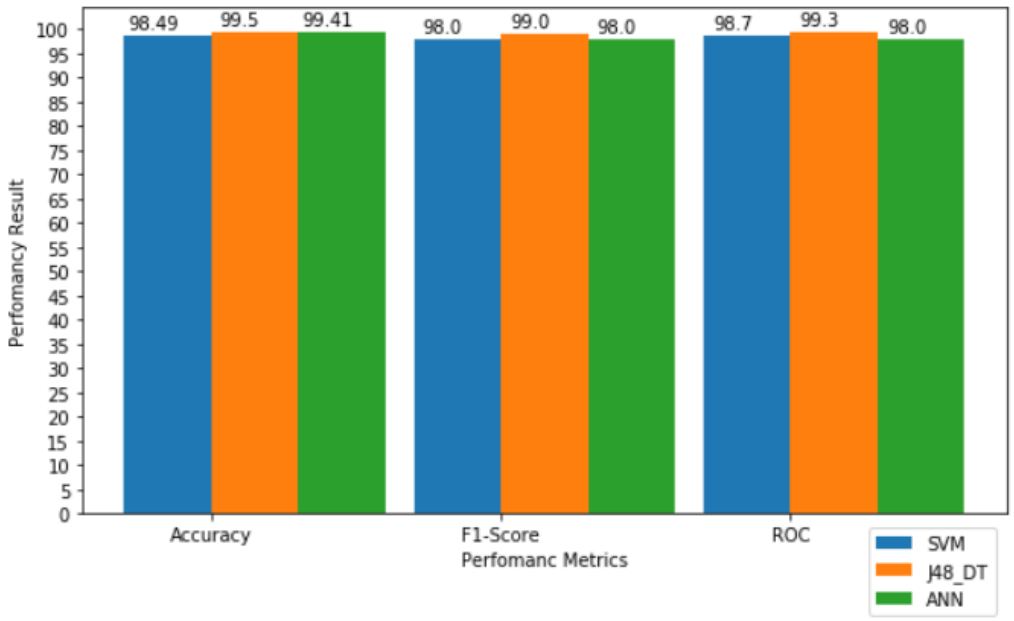


Figure 5.3: Model_3-Summary Comparison of Performance Result

5.1.2 *Result Comparison of Models*

Since the goal of goal this study extends beyond comparative of classifier algorithms, it is interesting to conduct a comparison modeling approaches based on the stated constraint in parallel with the comparison of the algorithm's performance .
As discussed early in each model comparison with different constraints or metrics, the J48 DT classifier has resulted in better performance compared to the others mentioned. Although the performance result shown in this algorithm is even different in almost all metrics under consideration. The detailed results are discussed in table 5.4 comparatively.

Classifier	Models	Accuracy	Precision	Recall	F1_score	ROC
J48 DT	Model_1	98.35%	97.00%	99.00%	98.00%	98.70%
	Model_2	99.24%	98.00%	100.00%	98.00%	98.24%
	Model_3	99.50%	100.00%	99.00%	99.00%	99.30%

Table 5.4: Result Comparison of Model Based on Data-sets.

This variety of performance of the algorithm with the same parameters is due to the flexibility of models in the detection of mobile data roaming.

Here also the viewpoints are narrowed down into accuracy, F1_score, and ROC which can inclusively and concisely analyze the performance of models. As results expressed in figure 5.4 show, the performance of Model_1 delivered less rank in almost all metrics mentioned compared to Model_3. However, it is observed largely in only ROC when compared to Model_2 whereas they achieved the same result 98.00% in average F1_score. Similarly, Model_2 has computed high with an accuracy of 99.24% compared to Model_1 which is accurate to 98.35%. In summing-up, Model_3 has computed better in accuracy, average F1_scores, and ROC with performance results of 99.50%, 99.00%, and 99.30% respectively. The respected detail is discussed in figure 5.4.

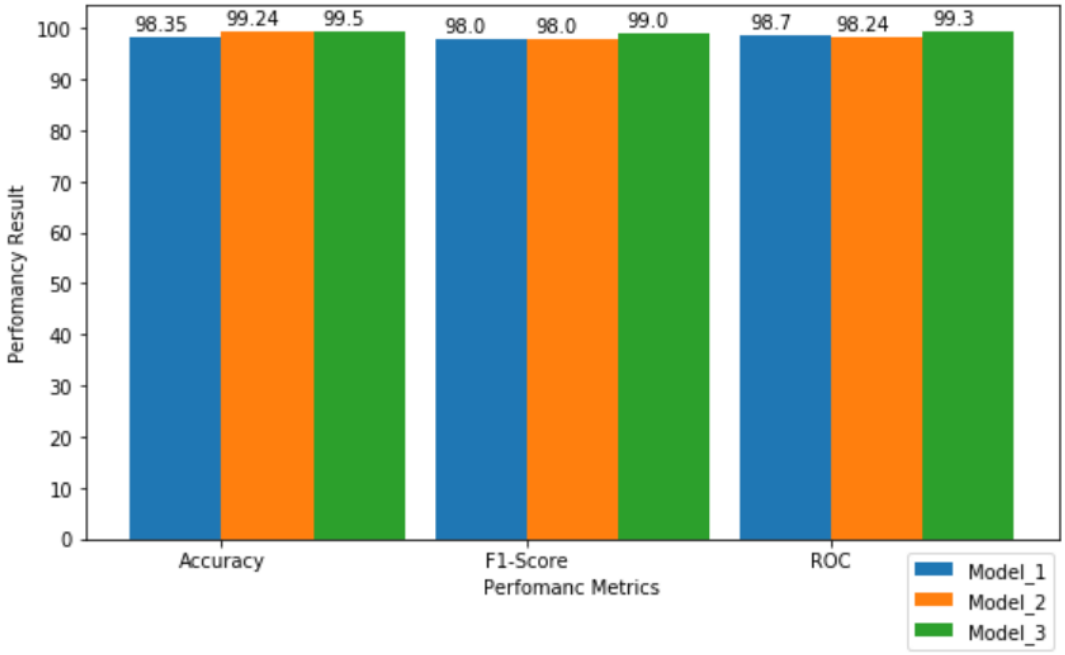


Figure 5.4: Model_4 The Summary Comparison of Models

This variety can be reflected in two ways in addition data preparation methods followed under chaper_4. Firstly, rather than threatening roaming data independently, algorithms fit to learn more from mixed data-set in addition to saving time and storage. Secondly, the inbuilt classification functionally of tree-based classification such as J48 DT is better in training and classifying data from supervised machine learning considered.

6

CONCLUSION AND FUTURE WORKS

Regarding this thesis results and techniques followed, some concerns need to be addressed. There are areas and suggestions for further studies for more improving the prevention and detection techniques of roaming fraud. This section perceived with a general conclusion and suggests future study areas regarding the problem.

6.1 CONCLUSION

For the flexibility of fraudsters and intelligence in adapting to new and complex technologies, it is a continued challenge to secure telecommunications services. Beyond various security checks in place, mobile data roaming fraud is still a significant source of revenue loss for telecom operators and service providers such as ethio telecom.

To close out this security gap, this study is concerned with analyzing roaming mobile data usage to distinguish between legitimate and fraudulent usage. To solve the problem, roamer CDR data from ethio telecom network elements are collected and the detection models are built after applied steps of data preparation techniques. J48 DT-tree-based classifiers, ANN-neural function centered classifier, and SVM-marginal oriented classifier are the three supervised machine learning classifier algorithms debated.

To make the detection mechanism more comprehensive, three alternative models were built and evaluated independently. The evaluation is based on the performance of the classifiers regarding each modeling approach. The evaluation of models was carried out using key model performance measurement metrics such as accuracy, precision, recall average F1 score, and ROC. The true positive and true negative rates are measured in ROC and precision and recall are measured in F1_Score, which are in tension with each other. In comparison to other classifiers, J48 DT performed better overall.
Throughout the process of evaluating the model, the success model from each data-set is not neglected. However, the model built from a mixed data-

set is a good fit in detecting roaming mobile data fraud. This implies that organizations better periodic analysis of data rather than waiting for TAP file-user usage from the visited network. Additionally, it is also important to enhance the roaming business process such as maximizing TAP file transfer time and network security.

For compiled usage behavior exceeds the detection of such fraud, organization better to periodically analysis of data rather than waiting for TAP file-user usage from visited network.

In this way, it is important to detect and prevent roaming fraud related to mobile data and possible to minimize losses, maximizing profits and customer trust and the company's reputation can be improved.
Improving roaming agreements among operators which shall includes the tendency to preventing roaming frauds is important. It is also necessary to improve the security holes on signaling system (SS7 or S9) for 3G networks and 4G or long term evolution (LTE) signaling system.

6.2 FUTURE WORKS

This study used mobile roaming CDR data from network elements with a fixed month dataset, and increasing the data size can improve the accuracy of the approaches. However, researchers should worry about the selection of classifiers and the machine performance that they can avail for some algorithms such as SVM, which falls or takes a long training time with a large data-set and ordinary computer performance.

For the study, only CDR data from various systems were used, but another researcher can conduct the same study using live signaling data. However, before dealing with signaling data, researchers needed special intermediate tools to label or structure data.

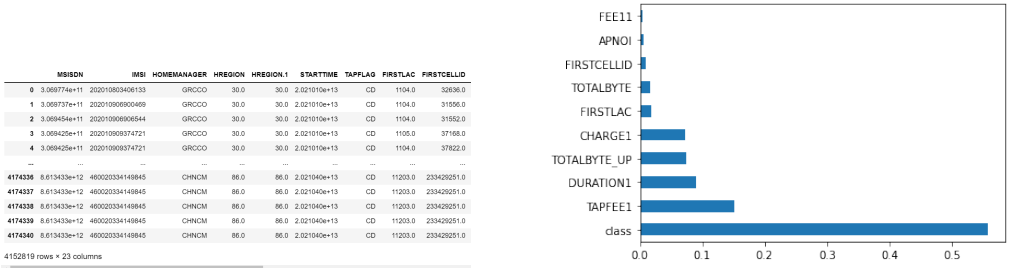
Assessments on improving roaming agreements among operators, which include the tendency to preserve roaming frauds and improving the security of hole signaling systems are also another suggested area.

Annex: 1

Attributes	Descriptions
FIRSTLAC	VN location area code
APNOI	APN operator identifier
CALL_DIRECTION	Address to which used
CHARGE11	Charge unit/cycle
DURATION(2)	Total measured time
FEE11	Total usage in byte
FIRSTCELLID	1st position connected
FLOWUP1	Active byte for roam in
GGSN	Gateway address
MSISDN	User service number
PDP_ADDRESS	Context request adress
TAPFEE(2)	Cost for TAP
TOTALBYTE	Total usage in byte
TOTALBYTE_UP	Total active in byte

Table A.1: Features Description

Annex: 2



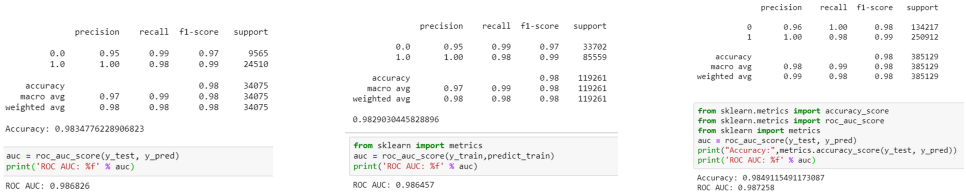
(a) Sample Data Visualazation

(b) Sample Feature Selection

Figure A.1: Sample data visualization and feature selection

Annex: 3

Here blows are sample results achieved with selected classifiers and different data-sets.



(a) Result of J48 DT with 1st model

(b) Result of ANN with 1st model

(c) Result of SVM result with 3rd model

Figure A.2: Results from roaming out and merged data-set

Annex: 4

Here Under Annex: 4 are the Publishable Manuscript Ready Along with Main Document.

Mobile Roaming Fraud Detection Based on User Behaviour: In case of ethio telecom*

1st Samuel Mekasa Michael
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
sm1481@gmail.com

2nd Ephrem Teshale Bekele (PhD): Assistant Professor
School of Electrical and Computer Engineering
Addis Ababa University
Addis Ababa, Ethiopia
ephremteshale@gmail.com

January 24, 2022

***Abstract*—The convergence of heterogeneous technology and flexibility of fraudster behaviors introduces security challenges like a fraud in the telecommunication industry. The mobile roaming data-internet fraud committed on visitor networks is a continued challenge and significant source of revenue losses for telecommunications societies. The introduced prevention and detection mechanisms such as minimizing usage detail transfer time and building the model using one-way data have limitations. In this study, we used three data sets to build the model for detecting mobile roaming data fraud by using three different supervised machine learning algorithms. The model is evaluated by different metrics and the model with merged data-set (roaming in and roaming out) achieved better performance and similarly, J48 DT has resulted greater with an accuracy of 99.50, average F1_Score 99.00, and ROC 99.30.**

Organizations are better at the periodical analyzing of data rather than waiting for TAP file-user usage from visited networks in addition to revising roaming agreement.

***Index Terms*—User behavior, Mobile data roaming, Fraud detection, Machine learning algorithms, Machine learning tools, Home network, Visited network**

I. Introduction

The increased communication service demand pushes global telecommunication networks to encompass heterogeneous communication technologies including the internet. These complexity and flexibility of fraudster behaviors, introduced security challenges such as fraud in telecommunication society [4] [9] [6].

Telecommunication fraud is the use of telecommunication networks to avoid payment- with incorrect payment, no payment or someone else pays according to International Telecommunication Union (ITU) [5]. Others such as [8], articulate it as, the stealing or use of telecommunication services to commit other form of frauds.

International mobile data roaming is one of the highly affected areas by such frauds. The ongoing challenge of its fraud decreases user satisfaction and experience, as well as an organization's reputation. ethio telecom who is currently the sole telecommunication operator and service provider in Ethiopia is in part of this challenge. For the difficult understanding of the complex ecosystem and rapid flexibility of fraudsters, a comprehensive understanding and avoiding fraud is a challenging task in the sector. Nevertheless, industry experts have a partial view because they usually specialize in the fraud type detected the businesses [4]. Telecommunication fraud is a major source of revenue loss for the service provider and their customers [2][1][8]. As the expert's feedback, telecommunication providers are losing 3% to 10% of their income due to frauds challenges in the sector [3].

According to Communications Fraud Control Association (CFCA) report 2019 shows, global fraud loss estimated $28.3 billion (USD) and billions of dollar losses associated with roaming fraud. [8][1][12]. The recent three years (2015, 20117, and 2019) of global telecommunication operators and losses from ethio telecom annual report due to frauds are discussed in table 1.

	Revenue Losses in case of Fraud		
	[Global Telecom]		[Ethio Telecom]
Years	$ Billion (USD)	% of Loss	$ Million (USD)
2015	38.1	1.69	33
2017	29.2	1.27	89
2019	28.3	1.74	∼48

TABLE I
TELECOM OPERATORS REVENUE LOSS IN CASE OF FRAUD [1][8].

Penetrative technology, employee disappointment, organizational inefficiencies, weakness of business models, financial crimes (money laundering), geopolitical and socioeconomic influences are some indications for the motivation of fraudsters [4][9].

To fight such fraud, telecommunications organizations like ethio telecom rely on using antiquated system-rule-based systems like the Fraud Management System (FMS). However, it is not advanced because it generates a high number of false positives. Studies were conducted to detect and prevent for minimizing fraudster's impact on the sector.

In [4], the taxonomy that distinguishes frauds into the origin of fraud, the exposures, the fraud types the exploitation techniques, and the way they benefit was introduced through study root causes of frauds, the vulnerabilities of industries. The proposed implementation of Near Real-Time Roaming Data (CDR) Exchange (NRTRDE) back to the home network for the prevention of roaming data still takes about four hours which is a long enough time gap for the fraudsters to make a profit.

Studies in Maciá-fernández, G.[10] and [9], the roaming fraud attack and defense strategies focused on the telecom service and their network security threats were discussed. It explores the weakness of the business process in securing the

service and the vulnerability network in telecommunication. This proposed approach ranged from a statistical model to more complex methods such as data mining and machine learning such as neural networks. However, it needs further data analysis not implemented.

Others such as [13],[11] and [19] were built an algorithm that determines suspected fraud numbers from normal to increase the income of the telecommunication service provider through the detection of international roaming fraud. However, it is a rule-based fraud management system in which is threshold has been set by the operator and usages compared to it.

Recently, in [18] and [16] build a predictive model used to detect international mobile roaming fraud and subscription fraud respectively. In the process of developing the roaming model, roaming traffics was used and local usage is used for subscription fraud and it states that all frauds are started from subscription fraud. However, roaming fraud is not limited to subscription fraud characteristics as committed both on visited and home networks.

The roaming fraud detection build has a substantial solution for the problem. However, it is confined to roaming out subscribers with a togethering of all services currently provided. But, there are notable frauds with roaming in on the visited network and considering many service types, may limit from deep sight for a specific service type's fraud as usage behaviors are contributed from many services as features. The study also utilized TAP file-out roamer's usage from the visited network forwarded through third parties, which takes a long time and maybe enough time for a fraudster to commit fraud.

Additionally, other roaming fraud prevention and detection mechanism are proposed like in [14][17][7] aimed to show possible security vulnerabilities of signaling elements or SS7 and cyber attacks for roaming networks and implement machine learning against rule-based filtering for detection of SS7 attacks. Specifically, the article in [17] relies on an inclusive review of the SS7 expected attacks and provides mitigation techniques such as a machine learning based framework to detect anomalies in the SS7 network with comparative to rule based filtering. However, the real user data was not considered in the scenario which may make fail to show the real behavior of roaming fraudsters.

In parallel, in the study of [14], a fake base station is installed to establish a connection to a subscriber through the air interface. The IMSI (International Mobile Subscription Identity) is captured using this fake station. To explore the network-network communication an emulator-based (jSS7 simulator) LTE testbed is used. The author has investigated how Diameter messages can be manipulated over the S9 interface to perform a fraud or DoS attack using the IMSI number and in its result shows the difference between abnormal and legal usage. On the other hand, analyzing using labeled data from real network traffic is better realistic than using such a fake base station.

The main goal of this research is to detect mobile data roaming fraud based on the user behaviors using machine learning algorithms through:

- Insight the international roaming fraud nature specifically mobile data fraud with the tendency to ethio telecom.
- Selecting relevant usage fields or feature used in building the fraud detection model.
- Comparing the ML algorithms based on their performance results.
- Build model from alternative data-sets to select the optimal approach for better detection of roaming data fraud.

The compressed system model of the study is like in figure 1

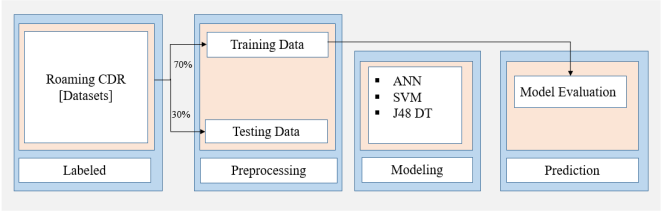


Fig. 1. System Model of the Study

The rest of the paper is organized as follows. Section II explains the data-set and approaches followed in developing the model. Section III discuss detail about the result achieved with each data-set. Finally, the conclusion and future works points are discussed in section paper IV.

II. Methodology

The mobile roaming data-internet subscriber's usages and fraud data are collected from ethio telecom systems in every two hours separately. All the collected data are Comma-Separated Values (CSV) format- a file format that allows us to save the tabular data and much useful for large data-sets. First, we examine the roaming out data set to build model used to detect roaming out fraud. In this case, we consider 12 months of consumption from June first, 2020 to May end 2021 with a total of 295,446 records. Considering much month's usage in past may handle a variety of user behaviors increase the model's quality.

In addition to the CDR data, fraud service numbers for the corresponding months are collected and which is numbered to 1536 unique data. The pandas (python libraries)-filter data frame method on anaconda3 navigator version 2.0.4 of jupyter notebook 6.0.1 version with base root environment is used and find 26,701 fraud-related CDR data for discussed months and know we have labeled data-set.

Secondly, we consider roaming in usage or data to build a model used to detect roaming in fraud. It is also collected from the organization and the three months CDR data (January, February, and April) of 2021 traffic usage are considered to develop the model. Accordingly, 3,138,055 data usage or CDR are found which is very high traffic compared to roaming in our data. A similar approach from the same system in roaming out fraud number collection is used to acquire related fraud numbers and can find 32,068 fraud-related CDR are extracted

and labeled accordingly.

Thirdly, two data-sets from roaming in and roaming out used. The shared fields appear common for both data-sets are discovered.

To make data suitable for machine learning, data cleaning activities are followed with help of panda libraries.

Cleaning Duplicates: The zero-variance predictors are input features that have homogeneous data across the entire column of observations and they contribute no value to the prediction process because the target variable is unaffected by the input value presenting them redundant. Fields stores similar values in the entire column is dropped out from data with the help of the Panda drop function. The process is also carried out in all data frames referenced in this study.

Handling Missing Values: Later checking of NaN value exiting in each column and records by using panda's library such isna() function. These values are treated in two ways: removing columns and rows with NaN values greater than 60%-based on data analyzing basics. Accordingly, the number of columns was reduced to 22 fields. Then, the records with less than the defined threshold value are treated by substituting the values with one (1) by using the fillna().

Outlier Handling: There are data points that appear to differ significantly or irregular patterns from other observations in our data-sets. To detect it, the Inter-Quartile Range (QR) technique-each data-set is divided into quartiles(upper quartile and lower quartile). For features that appear different, we detect outliers by using boxplot function on python Matlib libraries. To repair data at least near to its value, values out of ranges are replaced with corresponding limits.

Feature Selection: Irrelevant features in the data can badly affect the performance of the model. With the help of the seaborn library, we computed correlations of variables and visualized them using the matplotlib function in Python. The correlation matrix of the merged data-set is visualized in figure 2.

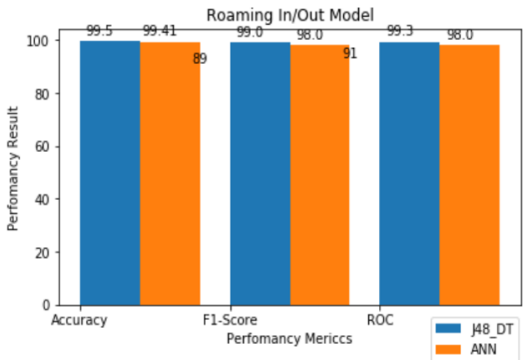


Fig. 2. Correlation of model one in Class Distribution.

For two features are interrelated, the model only requires one of them, as the second one provides no extra information and thus we removed one of them from consideration. As a result, we eliminated features whose information is correlated with 85% or higher-basics of data analysis. Based on this, SelectKBest, a scikit-learn library, provides the most useful features.SelectKBest, a scikit-learn library, provides the optimal learning input features. The higher the score, the more important the feature is in predicting our target feature. Figure 3 shows the top features selected from merged data-set.

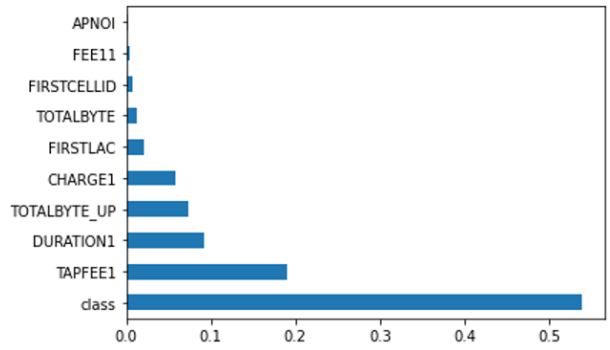


Fig. 3. Selected Features in Building the Model.

Data Balancing: The data are skewed-observation in one class exceeds that of in other classes and this leads strong class biases. Class Distribution of Roaming in data is visualized in figure 4. Building the model with such data, the accuracy predicting tending to the majority class and we fail to capture the minority compatibly. To threat equivalently, Synthetic Minority Oversampling Technique (SMOTE) is applied to the training data-set. The SMOTE creates a new training data-set based on the original training data. It chooses a minority class as the input vector and finds its neighbors (k nearest)-specified as an argument in the function. Such a method is advanced compared to simple random oversampling as using data variety in parallel to increase minority class.

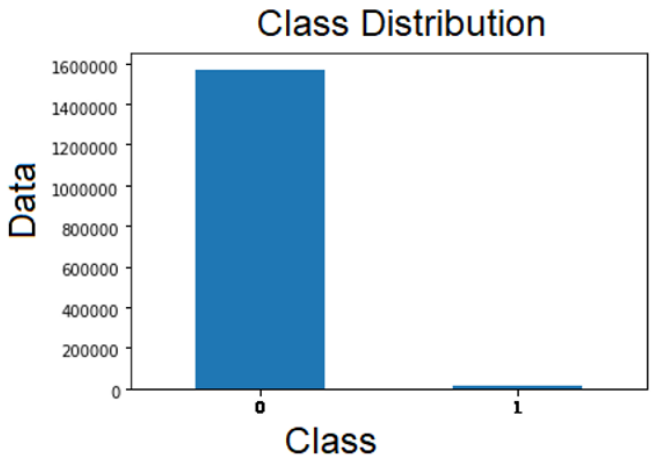


Fig. 4. Roaming in Class Distribution.

Validation Techniques: To determine how the model is really performing, the hold out cross-validation techniques used-which entire data-set is divided into training data and testing data. We split 70% to 30% ratio for training and testing data respectively-ML for data analyzing basics. From the

sklearn Python machine learning library, the train_test_split function evaluation is used for the implementation.

Training Data-set: When the classifier is evaluated, it is based on how well it predicts the class of the cases it is trained on, which is used to fit the model.

Test Data-set: The efficiency of the classifier algorithms is measured by how effectively they predict the type of a set of instances to determine how well a model fits. Figure 5 depicts the train test system model that is applied.

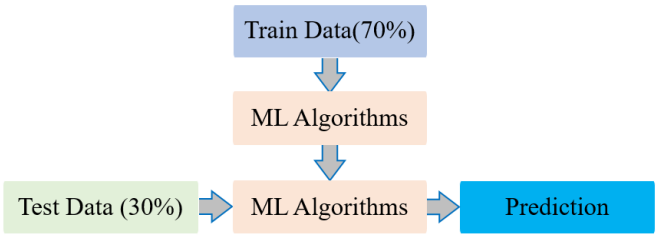


Fig. 5. Train Test Supply System Model Based on [16].

Training the Model: A total of nine experiments are conducted with three data-set options. Each data-set is categorized into the number of training and test data-sets. From the first, second, and third data-sets, 119261, 1108959, and 898633 instances are respectively used for the train using the fit function-based on the classifiers' library built-in python and in parallel, we noticed that the DecisionTreeClassifier from sklearn.model is the quickest and most well-trained classifier in this experiment. To evaluate(test) how well fit is predicted, 51112, 475268, 385129 instances from the first, second and third data-sets are used respectively.

J48 Decision Tree: It considers the best attribute as the root from the entire data-set and then divides it into subsections to build a tree. It produces unambiguous findings when the data is largely categorical and conditional. However, affected by noise and unsuitable for very large data-sets.

Support Vector Machine: Creates decision border-hyperplane at the extreme point and separates n-dimensional space into classes. The classier uses only a small subset of the total training set for classification, thus reducing the computational complexities through the use of kernel trick and over-fitting of data is avoided by classifying with a maximum margin [15]. On the other hand, training time headache with large data sets and not in effect with noisy data.

Artificial Neural Network: Data forwarded from input to output and hidden layers are the computational engine of the Multi-layer Perceptron (MLP)-ANN contains more than 1 hidden layer. It can solve non-linearly separated problems and tolerate noise data. However, the data hungriness, dependency on data quality, and sensitivity to overfitting are some of its drawbacks.

Performance Evaluation Metrics: The inclusive model evaluation parameters for classification such as: Confusion matrix-provide the detailed view of model performance including predicting degree. Accuracy- computes the rate of total correct classification. Precision-measure that indicates how well our model performs in terms of false positives. Recall- gives information about the performance with regards to the false negatives. For both recall and precision are tension for each F1-score- weighted harmonic mean of the precision and recall is used to measure both under one and Receiver Operator Characteristic (ROC)- indicates how well the model can distinguish between classes are applied.

III. Result and Discussion

The experiments are carried out with selected algorithms on an anaconda3 jupyter notebook python environment which provides several packages for experimentation. For this binary classification, a value of zero(0) represents non-fraud or normal usage which is the 1^{st} row of trebles whereas a value of one (1) indicates fraud usage-irregular usage of data denoted in 2^{nd} row. However, for such an unbalanced data-set, needs special threat to classify both classes nearly at the same level.

Result Comparison of Algorithms: To show the detailed performance of the model, metrics that were used are, accuracy, precision, recall, F-Measure, Receiver Operator Characteristic (ROC). However, to make it more comprehensive one can be described by another, we need deep through F1_score (precision and recall) and for ROC(true positive rate and true negative rate) and accuracy.

According to the methodology followed and applied metrics in the building of the model from the first data-sets, the detail is summarized in table 2.

Classifiers	Accuracy	Precision	Recall	F1_score	ROC
J48 DT	98.35%	98.0%	99.0%	97.0%	98.70%
		100%	98.0%	99.0%	
		97.0%	99.0%	98.0%	
ANN	98.29%	95.0%	99.0%	97.0%	98.65%
		100%	98.0%	99.0%	
		97.0%	99.0%	98.0%	
SVM	94.27%	83.0%	99.0%	93.0%	95.70%
		100%	92.0%	96.0%	
		92.0%	96.0%	93.0%	

TABLE II
Model from 1^{st} Data-set

According to the considered data-set in the building of model as well as research methodology followed with selected algorithms and concerning evaluation metrics, a J48 DT classifier scores accuracy of 98.35% and ROC of 98.70%, which is better performance compared to the described classifiers. SVM, on the other hand, ranked lower on the criteria mentioned for accuracy and ROC, with scores of 94.27% and 95.70%, respectively. With an accuracy of 98.29% and ROC of 98.65%, ANN performs in the mid-range when comparing to the decision tree and SVM classifiers.In respect of F1_score, which evaluates insight precision and recall, ANN was found similar to J48 DT, which achieved 98.0%. However, using such metrics from SVM, the classification result is 93.0%, which is significantly small compared to others. The variations in classification performance are due

to the better classifier functionality composed from and data approach fitting with the technique relatively.

To develop a model from second data-set, we supply data from roaming in with mounts of outlined and train-set split approach discussed. The achieved results regarding to each algorithms and metrics are discussed in table 3.

Classifiers	Accuracy	Precision	Recall	F1_score	ROC
J48 DT	99.24%	98.00%	100%	97.00%	98.24%
		100%	98.00%	99.00%	
		100%	100%	98.00%	
ANN	98.98%	99.00%	100%	99%	97.79%
		100%	96.00%	98%	
		99.00%	98.00%	99%	
SVM	98.49%	97.00%	99.00%	98.00%	98.50%
		99.00%	98.00%	98.00%	
		98.00%	99.00%	98.00%	

TABLE III
MODEL FROM 2^{nd} DATA-SET

Almost all algorithms performed better than the performance in the previous data-set with nearly all parameters. However, there is some difference when treating in actual performance figures.
When compared with accuracy, J48 DT has resulted better with 99.24%. On another way, ANN with an F1_score result of 99.0% performed great. Similarly, J48 DT again achieves a higher result of 98.24% in ROC. On the other hand, SVM rated 98.50% in ROC, which is higher than results from ANN and J48 DT and this clearly shows that, although data training takes a while with SVM, the algorithm has a greater capacity to optimize true positive and true negative rate and similarly, using the hidden layer function of ANN delivers nearly all predicted actual positive rates.

Another model is built from shared features from roaming out and roaming in. The evaluation techniques and the parameters used to measure the performance of the model follow the same steps in prior data-sets mode. The detail achieved results are discussed in table 4.

Classifiers	Accuracy	Precision	Recall	F1_score	ROC
J48 DT	99.50%	99.0%	100%	100.0%	99.30%
		100%	99.0%	99.0%	
		100.0%	99.0%	99.0%	
ANN	99.41%	98.0%	100.0%	99.0%	98.0%
		100.0%	96.0%	98.0%	
		99.0%	98.0%	98.0%	
SVM	98.49%	96.0%	100.0%	98.0%	98.70%
		100.0%	98.0%	98.0%	
		98.0%	99.0%	98.0%	

TABLE IV
MODEL FROM 3^{rd} DATA-SET

Comparatively, the overall performance result of all classifiers has resulted better. When limited to a data-set model, here also J48 DT performs better with the accuracy of 99.50%, average F1_Score 99%, and ROC 99.30%. Although the result from ANN and SVM regarding these metrics are not neglected. ANN has resulted equal in the average F1_score with SVM which is 98%. The SVM scored accuracy result 98.49% this is relatively ranked less compared to other classifiers.
Result Comparison of Model Based on Data-sets: Since the goal of goal this study extends beyond comparative classifier algorithms, it is interesting to conduct a comparison of the model from different data-sets. As discussed early in each model comparison with different constraints or metrics, the J48 DT classifier has resulted in better performance compared to the others mentioned. Although the performance result shown in this algorithm is even different in almost all metrics under consideration. The detailed results are discussed in table 5.

Classifier	Model from:	Accuracy	F1_score	ROC
J48 Dta	1^{st} Data-set	98.35%	98.0%	98.70%
	2^{nd} Data-set	99.24%	98.0%	98.24%
	3^{rd} Data-set	99.50%	99.0%	99.30%

TABLE V
RESULT COMPARISON OF MODEL BASED ON DATA-SETS.

The viewpoints are narrowed down into accuracy, F1_score, and ROC which can inclusively and concisely analyze the performance of the model. The performance of the model for the first data-set ranked less in almost all metrics compared to that of the third data-set. However, it is observed larger in only ROC when compared to model performance from the second data-set and they achieved the same result 98.00% in average F1_score. Similarly, model performance from the second data-set computed high with an accuracy of 99.24% compared to first which is accurate to 98.35%. In summing-up, the model from the third data-set has computed better in accuracy, average F1_scores, and ROC with performance results of 99.50%, 99.00%, and 99.30% respectively.
These varieties can be reflected in two ways in addition data preparation methods followed. Firstly, rather than threatening roaming data independently, algorithms fit to learn more from mixed data-set or variety of feature contributes more information about user behavior in addition to saving time and storage. Secondly, the inbuilt classification functionally of tree-based classification such as information gain or entropy in J48 DT is better in training and classifying data from supervised machine learning considered.

IV. CONCLUSION AND FEATURE WORKS

For the flexibility of fraudsters and intelligence in adapting to new and complex technologies, it is a continued challenge to secure telecommunications services. Beyond various security checks in place, mobile data roaming fraud is still a significant source of revenue loss for telecom operators. To close out this security gap, this study is concerned with building a model for the detection of roaming mobile data fraud using three alternative data-sets. Model from compiled usage behavior exceeds in the result and therefore organization better to periodically analyzing usage data rather than waiting for TAP file to detect and prevent such fraud and then possible to minimize losses, maximizing profits and customer trust, as well as the company's reputation, can be improved.

In the future, the researcher can conduct the same study using live signaling data. However, before dealing with signaling data, it needed special intermediate tools to label or structure data and worry about selecting classifiers for some algorithms such as SVM, which falls or takes a long training time with a large data-set. Assessments on improving roaming agreements among operators, which include the tendency to preserve roaming frauds and improving the security of hole signaling systems are also another suggested area.

BIBLIOGRAPHY

[1] Olivier FESTOR. "Understanding Telephony Fraud as an Essential Step to Better Fight It." PhD thesis. TELECOM ParisTech, 2017.

[2] Alae Chouiekh and EL Hassane Ibn EL Haj. "Convnets for fraud detection analysis." In: *Procedia Computer Science* 127 (2018), pp. 133–138.

[3] Gabriel Macia-Fernandez, Pedro Garcia-Teodoro, and Jesus Diaz-Verdejo. "Fraud in roaming scenarios: an overview." In: *IEEE Wireless Communications* 16.6 (2009), pp. 88–94.

[4] Vanita Jain. "Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining." In: *International Journal of Information Technology* 9.3 (2017), pp. 303–310.

[5] Peter Hoath. "The cost of fraud to the industry." In: *TAF Regional Seminar on Costs andTariffs*. 2008.

[6] Godfred Yaw Koi-Akrofi et al. "Global telecommunications fraud trend analysis." In: *International Journal of Innovation and Applied Studies* 25.3 (2019), pp. 940–947.

[7] NJ Bedminster. *COMMUNICATIONS FRAUD CONTROL ASSOCIATION ANNOUNCES RESULTS OF 2019 GLOBAL TELECOM FRAUD SURVEY (CFCA)*. Tech. rep. CFCA, 2019.

[8] Pablo A Estévez, Claudio M Held, and Claudio A Perez. "Subscription fraud prevention in telecommunications using fuzzy rules and neural networks." In: *Expert Systems with Applications* 31.2 (2006), pp. 337–344.

[9] mobileum. *Roaming Fraud: Fraud Attacks in Roaming Environment is Still High*. Ed. by mobileum. 2020. URL: https://rb.gy/kwpqyh.

[10] Gabriel Maciá-Fernández. "Roaming fraud: assault and defense strategies." In: *Roaming fraud: assault and defense strategies* (2009).

[11] GSMA. "International Roaming Explained." In: *GSMA-Mobile-roaming-web-English*.

[12] Yousef Alraouji and Arif Bramantoro. "International call fraud detection systems and techniques." In: *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*. 2014, pp. 159–166.

[13] Hussein M Marah, Osama Mohamed Elrajubi, and Abdulla A Abouda. "Fraud detection in international calls using fuzzy logic." In: *International Conference on Computer Vision and Image Analysis Applications*. IEEE. 2015, pp. 1–6.

[14] Qianqian Zhao et al. "Detecting telecommunication fraud by understanding the contents of a call." In: *Cybersecurity* 1.1 (2018), p. 8.

[15] Constantinos S Hilas and Paris As Mastorocostas. "An application of supervised and unsupervised learning approaches to telecommunications fraud detection." In: *Knowledge-Based Systems* 21.7 (2008), pp. 721–726.

[16] D. Tekeste. "A Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection : The case of ethio telecom." MA thesis. AAU, 2020.

[17] J. Bawa. *Leverage machine learning for telecom fraud detection*. Ed. by J. Bawa. 2019. URL: https://rb.gy/dbdbla..

[18] Tarikua Worku. "Predictive Modeling for InternationalRoaming Fraud Detection." MA thesis. AAU, 2018.

[19] Janvier Omar Sinayobye, Fred Kiwanuka, and Swaib Kaawaase Kyanda. "A state-of-the-art review of machine learning techniques for fraud detection research." In: *2018 IEEE/ACM Symposium on Software Engineering in Africa (SEiA)*. IEEE. 2018, pp. 11–19.

[20] Walid Moudani and Fadi Chakik. "Fraud detection in mobile telecommunication." In: *Lecture Notes on SOftware Engineering* 1.1 (2013), p. 75.

[21] Roman Chuprina. *How to Use AI and Machine Learning in Fraud Detection*. Ed. by Roman Chuprina. 2020. URL: https://rb.gy/kkj8uk.

[22] Isha Singh et al. "Signaling security in LTE roaming." In: *Signaling Security in LTE Roaming* (2019).

[23] Kaleem Ullah et al. "SS7 Vulnerabilities—A Survey and Implementation of Machine Learning vs Rule Based Filtering for Detection of SS7 Network Attacks." In: *IEEE Communications Surveys & Tutorials* 22.2 (2020), pp. 1337–1371.

[24] Bob Kamwendo. "Vulnerabilities of signaling system number 7 (SS7) to cyber attacks and how to mitigate against these vulnerabilities." PhD thesis. 2015.

[25] Data-Fair. *What is Data Science? A Complete Data Science Tutorial for Beginners*. 2018. URL: https://bit.ly/2KcX4AJ.

[26] Domaka N. Nanwin Ledisi G. Kabari and Edikan Uduak Nquoh. "Telecommunications Subscription Fraud Detection usingArtificial Neural Networks." In: *Machine Learning Arteficial Intelegence* (2015).

[27] GHAYAS. *WHAT IS THE DIFFERENCE BETWEEN PREPAID AND POSTPAID? Accessed on: 08, 02, 2021*. Ed. by A GHAYAS. URL: https://rb.gy/mkjscb.

[28] ethio telecom. *Customer Statistics*. Ed. by ethiotecom. 2020. URL: https://rb.gy/6tiryl.

[29] Bindu Rao. *Prepaid simcard for automatically enabling services*. US Patent App. 11/388,865. 2007.

[30] allaboutETHIO. *Mobile Postpaid: Accessed on 2021, 02, 08*. Ed. by allaboutETHIO. URL: https://rb.gy/dvpcz8.

[31] ethio telecom. *Roaming Service*. Ed. by ethio telecom. 2020. URL: https://rb.gy/fluhvc.

[32] James H. Lawrence Thomas R. Suozzi. "Telecommunications Fraud DEFINITION." In: *Nassau County Police Department: Telecommunications Fraud* ().

[33] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." In: *Journal of Network and Computer Applications* 68 (2016), pp. 90–113.

[34] K. Hagos. "SIM-Box Fraud Detection Using Data Mining Techniques: The Case of ethio telecom." MA thesis. AAU, 2018.

[35] Mais Arafat, Abdallah Qusef, and George Sammour. "Detection of wangiri telecommunication fraud using ensemble learning." In: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE. 2019, pp. 330–335.

[36] Vodafone. *Wangiri Japanese term telecom fraud also referred to as one ring cut*. Ed. by Vodafone. 2020. URL: https://rb.gy/atar3p.

[37] Jonathan Spruytte et al. "International roaming in the EU: Current overview, challenges, opportunities and solutions." In: *Telecommunications Policy* 41.9 (2017), pp. 717–730.

[38] Vaishali Advani. *What is Machine Learning? How Machine Learning Works and future of it?* Ed. by Vaishali Advani. 2020. URL: https://rb.gy/qdxfic.

[39] Avantika Monnappa. *Data Science vs. Big Data vs. Data Analytics*. Ed. by Avantika Monnappa. URL: https://rb.gy/bs6iho.

[40] Dr. Michael Tamir. *Machine Learning: How machine learning works*. Ed. by IBM Cloud Education. URL: http://bitly.ws/g6L7.

[41] Stephen J Mooney and Vikas Pejaver. "Big data in public health: terminology, machine learning, and privacy." In: *Annual review of public health* 39 (2018), pp. 95–112.

[42] IBM Cloud Education. *Supervised Learning*. Ed. by IBM Cloud Education. URL: https://rb.gy/bgbnyj.

[43] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.

[44] Bo-Hao Chen, Jia-Li Yin, and Ying Li. "Image noise removing using semi-supervised learning on big image data." In: *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*. IEEE. 2017, pp. 338–345.

[45] YCAP Reddy, P Viswanath, and B Eswara Reddy. "Semi-supervised learning: A brief review." In: *Int J Eng Technol* 7.1.8 (2018), p. 81.

[46] IBM Cloud Education. *Unsupervised Learning*. Ed. by IBM Cloud Education. URL: https://rb.gy/r0zuta.

[47] Hardeep Singh. "Performance analysis of unsupervised machine learning techniques for network traffic classification." In: *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. IEEE. 2015, pp. 401–404.

[48] Francesco Musumeci et al. "An overview on application of machine learning techniques in optical networks." In: *IEEE Communications Surveys & Tutorials* 21.2 (2018), pp. 1383–1408.

[49] Błażej Osiński and Konrad Budek. *What is reinforcement learning? The complete guide Accessed on: 17, 02, 2021*. URL: https://rb.gy/84a1qx.

[50] Javapoint. *Reinforcement Learning TutorialAccessed on: 17, 02, 2021*. Ed. by Javapoint. URL: https://rb.gy/y2qo2v.

[51] Anirudh Janagam and Saddam Hossen. *Analysis of network intrusion detection system with machine learning algorithms (deep reinforcement learning algorithm)*. 2018.

[52] James Cannady. "Next generation intrusion detection: Autonomous reinforcement learning of network attacks." In: *Proceedings of the 23rd national information systems security conference*. 2000, pp. 1–12.

[53] Rajesh Gupta et al. "Machine learning models for secure data analytics: A taxonomy and threat model." In: *Computer Communications* 153 (2020), pp. 406–440.

[54] Larhmam. "Support-vector machine." In: *Support-vector machine* (2018).

[55] Sharmila Subudhi and Suvasini Panigrahi. "Use of fuzzy clustering and support vector machine for detecting fraud in mobile telecommunication networks." In: *International Journal of Security and Networks* 11.1-2 (2016), pp. 3–11.

[56] Oliver Knocklein. *Classification Using Neural Networks*. Ed. by Picks(Towards Data Scinece). URL: https://bit.ly/2WgF670.

[57] Oludare Isaac Abiodun et al. "State-of-the-art in artificial neural network applications: A survey." In: *Heliyon* 4.11 (2018), e00938.

[58] By Great Learning Team. *Supervised Machine Learning with Three Models: Processes involved in Decision Making*. Ed. by Greate Learning. URL: shorturl.at/kqEL0.

[59] Jahnavi Mahanta. *Introduction to Neural Networks, Advantages and Applications*. Ed. by Jahnavi Mahanta. 2017. URL: https://rb.gy/rptm6j.

[60] Akash Patel. *Terminologies in Decision Tree :Attribute Selection Measures (ASM)*. Ed. by medium. URL: shorturl.at/vxKN8.

[61] RAM DEWANI. *5 Popular Data Science Languages – Which One Should you Choose for your Career?* Ed. by RAM DEWANI. 2020. URL: https://bit.ly/342GxXj.

[62] Nitin Garg. *10 Powerful Programming Languages For Doing Machine Learning*. Ed. by Nitin Garg. 2020. URL: https://bit.ly/3oDnM4D.

[63] Vikash Kumar. *Python Vs R: What's Best for Machine Learning*. Ed. by Vikash Kumar. 2019. URL: https://bit.ly/340Aena.

[64] Anmol Bansal and Satyajee Srivastava. "Tools used in data analysis: A comparative study." In: *International Journal of Recent Research Aspects* 5.1 (2018), pp. 15–18.

[65] Garrett Grolemund. *Quick list of useful R packages: Recommended Packages*. Ed. by Rstudio. URL: shorturl.at/kovK6.

[66] Amelec Viloria et al. "Comparative Analysis Between Different Automatic Learning Environments for Sentiment Analysis." In: *International Symposium on Distributed Computing and Artificial Intelligence*. Springer. 2020, pp. 134–141.

[67] Jaswitha Abbineni and Ooha Thalluri. "Software Defect Detection Using Machine Learning Techniques." In: *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE. 2018, pp. 471–475.

[68] Ceyhun Ozgur et al. "MatLab vs. Python vs. R." In: *Journal of Data Science* 15.3 (2017), pp. 355–371.

[69] Sydney Mambwe Kasongo and Yanxia Sun. "A deep long short-term memory based classifier for wireless intrusion detection system." In: *ICT Express* 6.2 (2020), pp. 98–103.

[70] Omar Almomani. "A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms." In: *Symmetry* 12.6 (2020), p. 1046.

[71] LAMIDO YAHAYA, IBRAHIM HASSAN, and ABBAS MUHAMMAD RABIU. "A SURVEY OF PERFORMANCES OF SOME SELECTED MACHINE LEARNING ALGORITHMS FOR CARDIOVASCULAR DISEASE PREDICTION." In: ().

[72] Nikita Duggal. *Top 10 Python Libraries For Data Science for 2021 Accessed on: 18, 02, 2021*. Ed. by Nikita Duggal simplilearn. URL: https://rb.gy/kvgjex.

[73] John Terra. *Why is Python Essential for Data Analysis? Accessed on 2021, 02, 18*. Ed. by John Terra. URL: https://rb.gy/wtostw.

[74] Rising Odegua and Festus Ikpotokin. "DataSist: A Python-based library for easy data analysis, visualization and modeling." In: *arXiv preprint arXiv:1911.03655* (2019).

[75] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[76] Sebastian Raschka. *Python machine learning*. Packt publishing ltd, 2015.

[77] Samadrita Ghosh. *A Comprehensive Guide to Data Preprocessing*. Ed. by Neptune. URL: https://shorturl.gg/mNu.

[78] Gil Press. *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Ed. by Gil Press. URL: https://shorturl.gg/AUz.