Semi-Supervised Algorithm to Detect Over-The-Top Bypass Fraud: in the case of ethio telecom

By: Wubalem Kinfmichael

Adviser: Dr.-Ing Yihenew Wondie

A Thesis submitted to

School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

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

Addis Ababa University

Addis Ababa, Ethiopia

November 4, 2021

Declaration

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

Wubalem Kinfmichael _____

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Wubalem Kinfmichael, entitled Semi-Supervised Algorithm to Detect Over-The-Top Bypass Fraud: in the case of ethio telecom** and submitted in partial fulfillment of the requirements for the degree of Master of Science Telecommunication Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee

Internal Examiner		Signature		Date	
External Examiner		Signature		Date	
Adviser	Dr.-Ing. Yihenew Wondie	Signature		Date	
Co-Adviser		Signature		Date	

Dean, School of Electrical and Computer Engineering

Abstract

Telecom fraudsters' behavior evolves over time, and their capacity to defraud telecom service providers grows at a rapid pace. Interconnected bypass fraud is one of the most common types of telecom fraud. It is a new sort of telecom fraud that works by intercepting international voice calls and transferring them to VoIP for termination as an OTT call without the knowledge of the telecom provider, caller, or called party.

OTT frauds based on the Mobile Station International Subscriber Directory Number (MSISDN) are becoming more common, posing a threat to telecom firms as the number of smart phones grows and it becomes easier to access OTT services from anywhere. Interconnected bypass fraud is the term for such operations, and OTT bypass is one sort of interconnected bypass fraud. One subtask in the detection of this scam is detecting OTT voice call packets using various network traffic classification techniques.

To categorize network traffic packets, machine learning (ML) techniques are utilized, including the semi-supervised and supervised algorithms Collective Filtered and Decision Tree (DT). For the training and testing of the chosen method, ten cross-fold validation techniques were used. Each ML algorithm's test dataset is correctly prepared. With reasonable model build and evaluation times, both ML algorithms DT and collective filtering achieve better performance of 99 percent and 91 percent accuracy, respectively.

KEYWORDS

OTT bypass fraud, Telecom fraud, Network traffic classification, Machine-learning algorithms

Acknowledgments

First and foremost, I want to express my gratitude to God for providing me with the strength to complete all of the steps. Next, I'd like to show my appreciation to my adviser, Dr.-Ing Yihenew Wondie, for his thoughtful and encouraging comments. I'd also like to express my gratitude to my evaluators, Ephrem Teshale (PhD) and Fitsum Assamnew (PhD), for their constructive criticism throughout the thesis progress presentations. I'd want to thank my employer, Ethiopian Telecom, for providing me with this chance, as well as those who have helped me realize my dream.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
CDR	Call Detail Record
CFCA	Communications Fraud Control Survey
DPI	Deep Packet Inspection
DNS	Domain Name Server
DRS	Domestic Revenue Share
FMS	Fraud Management System
GMM	Gaussian Mixture Model
GSM	Global System for Mobile communications
GUI	Graphical User Interface
ID3	Iterative Dichotomiser 3
IP	Internet Protocol
IQR	Inter Quartile Range
IRSF	International Revenue Share Fraud
K-NN	K Nearest Neighbor
MLP	Multi Layer Perceptron
MSISDN	Mobile Station International Subscriber Directory Number
OTT	Over-the-Top
PBX	Private Branch Exchange
PRS	Premium Rate Service
QoS	Quality of Service
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operating Characteristic
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SPID	Statistical Protocol Identification
SSL	Secure Sockets Layer
SSDP	Simple Service Discovery Protocol
SVM	Support Vector Machine
TCG	Test Call Generation

TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VAS	Value Added Service
VoIP	Voice over IP
WEKA	Waikato Environment for Knowledge Analysis

1

Introduction

On a global basis, telecommunication businesses have lost billions of dollars in income due to scammers around the world [1]. The usage of a telecom operator's infrastructure or services without intending to pay for them is referred to as telecom fraud [2]. As a result of the increased use of telecommunication services, fraudsters have refined their tactics of attack, taking advantage of technological improvements. Interconnect bypass Fraud, International Revenue Share Fraud (IRSF), and Premium Rate Service Fraud (PRSF) are the top three types of telecom fraud currently in use [3]. Over-the-Top (OTT) bypass fraud is one of the most common types of telecom fraud [4].

Instead of being terminated over the traditional voice call telecom infrastructure, OTT bypass redirects conventional voice conversations to Voice over IP over the Internet and converts them to a voice chat application on a smartphone. This type of rerouting (or hijacking) is supported by an international transit operator in collaboration with an OTT service provider, but without specific permission from the caller, called, and their operators [4].

The illegal injection of traffic into a telecom carrier's network, known as telecom interconnect bypass fraud, was the second most common cause of revenue loss, after IRSF, according to the Communications Fraud Control Survey (CFCA) 2017 report [1]. Telecom providers are estimated to have lost $4.27 billion as a result of interconnected bypass schemes. Ethiopia's single telecom service provider, Ethio telecom, is one of the companies most affected by interconnect bypass fraud. The firm now uses a rule-based Fraud Management System to identify and detect interconnect bypass frauds and other telecom scams (FMS). SIMbox fraud is the most well-known linked bypass scheme,

which hijacks an international call transfer to VoIP and terminates it as a local call[5] [6]. OTT bypass fraud is a newer type of telecom interconnect bypass fraud in which a non-OTT call is rerouted through OTT applications and received as an OTT call at the receiver side [5].

OTT bypass fraud [7] has developed as a new paradigm, notwithstanding the well-known Interconnect Bypass fraud. This paradigm shift is being driven by the need for fraudsters to avoid detection and produce more money, the emergence of a large OTT user base, and developments in telecommunication technologies. OTT bypass fraud is a new type of telecom interconnect bypass fraud in which a non-OTT call is rerouted through OTT apps and received as an OTT call at the receiver side [7]. As a result, the call termination cost that is supposed to be paid to the telecom operator at the receiving end will be shared between the OTT service provider and the transit operator engaging in the fraud. This form of scam is frequent in OTT applications like Viber, which require the user's MSISDN during service registration [8].

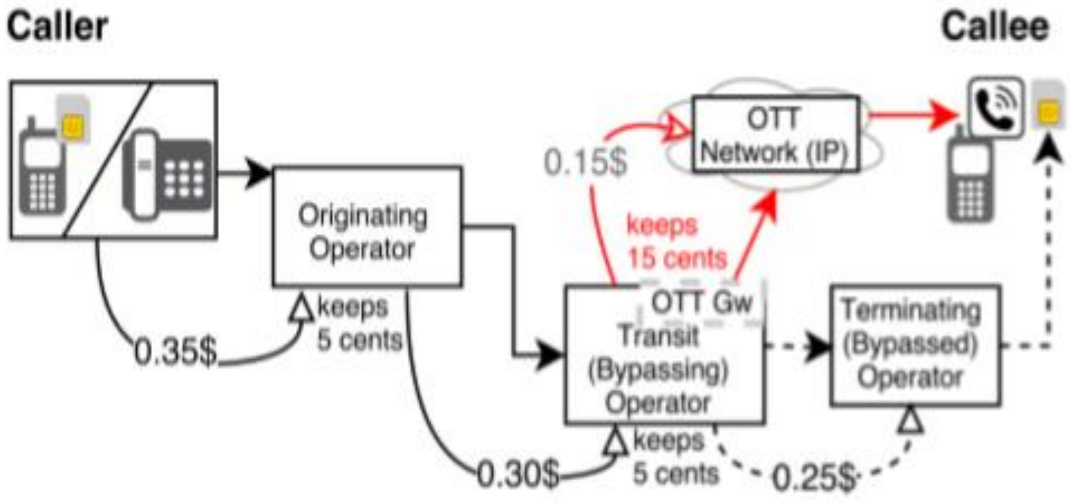

Figure 0-1. A normal call between 2 mobile phones in black and OTT bypass in red [4].

To combat interconnected connection bypass and other sorts of telecom fraud, telecom carriers have developed a number of fraud detection methods. Telecom businesses are losing a substantial amount of income due to the difficulties in recognizing and quantifying the impact of this fraud, even if they are unaware of its existence [8]. Test Call Generation (TCG), traffic data, audio fingerprinting, and Call Detail Record (CDR) are some of the primary approaches for reducing the impact of OTT bypass fraud.

1.1. Statement of the Problem

Telecom service providers invest a significant amount of money to deliver their services. Telecom fraudsters, on the other hand, have a major impact on telecom providers' operations and revenue. Telecom companies have lost a large amount of money as a result of interconnected bypass frauds, as mentioned in the preceding section. OTT bypass fraud is a type of linked bypass fraud that is related to the user's MSISDN number. When registering for OTT apps like WhatsApp, Viber, and Telegram, the MSISDN number was required as a user ID [8].

To my knowledge, Ethiopian telecom's FMS currently has no detection mechanism for OTT bypass fraud, and there is no explicit revenue loss report due to OTT bypass fraud. Since the corporation has lost a significant quantity of foreign currency due to improperly canceled international calls. As a result, such OTT bypass scams must be reduced, and the impact on Ethiopian telecom must be investigated.

Telecom operators gives different types of services by building their infrastructures and make money out of it. As previously explained telecom fraudsters uses telecom operators' infrastructure to make their own money without sharing or paying for it.

Telecom operators, academics, and other interested parties have created many fraud detection processes and approaches, and many comparisons have been made between different detection techniques. Tewodros Hailu recently released a research titled

"Network Traffic Classification Using Machine Learning: A Step Towards Over-the-Top Bypass Fraud Detection" in the context of Ethiopian telecom, which utilised generated data. As a result, ethio telecom will take legitimate network traffic statistics into consideration when detecting Over-the-Top Bypass frauds. There will be a mechanism to calculate the amount of revenue lost by Ethiopian telecom. The results of this study will be compared to those of the previous one.

1.2. Objective

1.2.1. General Objective

The main objectives of the research is to detect MSISDN based OTT bypass fraud using Semi-supervised machine algorithm.

1.2.2. Specific Objectives

The specific objectives of the research are:

- Identify the characteristics and features of OTT traffic.
- For network data traffic classification, choose specialized machine learning methods.
- Using the chosen machine learning algorithm, classify the gathered traffic.
- Evaluate the model's performance of the chosen machine learning method and make an appropriate recommendation.
- Based on the findings, make a conclusion and a recommendation.

1.3. Contributions of the Research

As far as we know, no work has been done on detecting MSISDN-based OTT applications using semi-supervised machine learning techniques on network traffic data obtained in a controlled laboratory environment. The following points will be discussed as a result of the research:

- To increase QoS and manage network administration by giving priority to applications that run on the network.
- Will provide support when new security policies for OTT services or fraud avoidance are developed or strengthened.
- Policymakers should be aware of the bandwidth utilization ratio of OTT applications and establish policies accordingly.
- For future researchers interested in OTT traffic categorization tasks, the labeled OTT network traffic data collected and used in this study can be used.

1.4. Literature Review

To detect MSISDN-based OTT, the researchers employed three machine learning algorithms: Adaptive Booster (AdaBoost) + J48, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and Support Vector Machine (SVM). To generate 1.7 million labeled packets for examination, the Author [8] used 10 cross-fold and distinct test data validation procedures. When comparing the outcomes of each machine learning method, the AdaBoost + J48 strategy outperformed the others for MSISDN-based OTT fraud categorization in both assessment methodologies (ten cross-fold validation and independent test data validation). According to Jayeeta Datta et al. [9], bleaves traffic classification is a critical challenge for offering differentiated service quality to applications.

Three algorithms for machine learning MSISDN-based OTT is detected using Adaptive Booster (AdaBoost) + J48, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and Support Vector Machine (SVM), which takes a sample from Viber, Tango, and Telegram. To generate 1.7 million labeled packets for examination, the Author [7] used ten cross-fold and distinct test data validation procedures. When evaluating the outcomes of each machine learning method for MSISDN-based OTT fraud categorization, the AdaBoost + J48 strategy outperformed the others in both assessment methodologies (ten cross-fold validation and independent test data validation). According to Jayeeta Datta et al. [8] bleaves traffic categorization is a

critical challenge for offering differentiated service quality to applications. Several approaches that use machine learning techniques to identify application traffic have been developed (with varying degrees of success). The experiment uses Google Hangout as a case study to explore the detection outcome. Google Hangout is a peer-to-peer video chat program that connects two people via the internet. The experiment used 2.5 million packets of network traffic, three distinct classification algorithms, and a 10-fold cross validation technique to quantify performance.

Over-the-Top (OTT) bypass fraud is an example of interconnect telecom fraud. Merve Sahin et al. [4] take a large chunk of the call charge, resulting in a huge loss of revenue for the skipped operators. Because any networked bypass fraud degrades telecom service quality while delivering no value to subscribers. According to the authors [4] OTT bypass affects up to 83 percent of calls. They also show that OTT bypass degrades service quality and can collide with other fraud strategies, making quality more challenging.

Internet service providers (ISPs) and telecommunications regulators are interested in detecting VoIP conversations in order to either ban unlawful commercial VoIP or favor paid consumers' VoIP calls. Due to the security complexity and tunneling method, signature-based, port-based, and pattern-based VoIP detection techniques are not more accurate or efficient in identifying MSISDN-based OTT fraud. As a result, the authors [10] provide a generic, robust, and efficient statistical analysis-based approach for recognizing encrypted, non-encrypted, or tunneled VoIP using a threshold-based rule-based approach. High-speed real-time network traffic may also be handled by the solution.

Four packet features are utilized for classification: packet length, delta time (packet inter-arrival time), cumulative byte, and relative time. These features are claimed to be employed for the first time in internet traffic categorization. According to the researchers, AdaBoost has the highest overall accuracy (98.3%), whereas MLP has the lowest accuracy (84.2%) when it comes to identifying the five applications. According

to the test findings, AdaBoost is the best classifier for spotting VoIP applications in particular when compared to the others [8].

Datta et al. [9] employed Naive Bayes, J48 decision tree, and AdaBoost machine learning approaches to detect Google Hangout traffic in other Google Hangout studies. They employed a dataset of 2.5 million packets generated as separate traffic for training purposes: 1, 984, 954 and 689,025 packets for non-Google Hangout and Google Hangout traffic, respectively. The dataset is collected and analyzed using the Wireshark and WEKA tools, respectively. The classification uses packet length, protocol, source and destination port numbers, packet type, and a DNS response from Google as packet features.

The author of this paper [11] considers OTT pass fraud detection using Skype network traffic. This aims to categorize Skype service flows such phone calls, skypeOut, video conferencing, chat, file upload, and download. For the statistical analysis values, a classification approach based on the Statistical Protocol IDentification was used to classify Skype encrypted traffic (SPID). They tested their model on a representative dataset and found that it performed exceptionally well in terms of Precision and Recall.

1.5. Methodology

The methodology that included in this research:

- Conducting extensive literature reviews which is mainly focuses on telecom frauds and especially related to OTT bypass telecom frauds.
- Continuous discussions held with Domain experts and select suitable classification technique for OTT traffic.
- Collecting VoIP traffic and capture the traffic using network traffic capturing tools such as Wireshark.
- With the selected machine learning tool to classify the traffic and analyze the performance of the selected classification algorithm.

1.6. Thesis Organization

This thesis research organized as follows; Chapter 2 discusses what OTT bypass fraud is and depict different OTT traffic classifier and detections techniques. Additionally, machine learning algorithms' concepts and categories explained in detail. Chapter 3 shows the experimental analysis of this research. Data preprocessing, feature selection, algorithm training and evaluation of building system model. Chapter 4 discusses the results obtained by the performance evaluation of the algorithms' model. At the final chapter, Chapter 5, conclusion and recommendations are stated.

2

BACKGROUND

2.1. OTT Bypass Fraud

Interconnected byPass fraud has been a big difficulty for telecommunication companies since their inception, and it is one that they are continuously battling to overcome. The fraudsters illegally utilize the telecom operator's infrastructure without paying the required fees, causing the telecom operator to lose money. This is commonly referred to as a "free rider" problem in economics [[8] [12] The hefty interconnection cost for ending an international call is the major enabler for this form of scam. In a contemporary sort of telecommunications connection bypass fraud, OTT bypass fraud is also known as OTT hijacking [7].

As described in the preceding section, such call rerouting is carried out by telephone transit operators in collaboration with the OTT service provider, without the caller's or telephone service provider's authorization on the receiver side [8]. OTT services use packet switching to transmit audio, text, and video content across data networks that are not under the authority of internet service providers [13] [14].

For service registration, several OTT applications require the user's MSISDN [9]. Skype and Yahoo Messenger are two popular OTT services that require a user account, whereas Viber and Tango require the user's MSISDN to register. OTT bypass fraud is most commonly carried out with OTT applications that require the user's MSISDN during service registration. Telecom fraud has an impact on telecom companies. In addition to income loss, service quality has deteriorated. There is also a cost associated with declining to provide value added services (VAS) [7].

To combat fraud, appropriate mitigation strategies have been devised. To detect fraudulent activity, Test calls, Network Traffic Analysis, CDR analysis, and audio fingerprinting analysis are used. There is a list of mitigation approaches for network traffic classification, some of which are as follows: To detect OTT by pass fraudulent activities, port-based techniques, payload-based techniques, and statistical-based techniques are used.

2.2. Machine Learning Algorithms

Machine learning (ML) is a computational method that employs particular instructions and rules to comprehend significant ideas of information and services from large amounts of data. Those principles, however, are not devised by computer programmers [6] [15]. Machine learning algorithms extract knowledge or information without sequential steps of instruction by learning from experience sample data. Machine learning was created to perform jobs that are difficult and beyond the capability of humans [16]. Machine learning algorithms used to address problems, in addition to these difficult challenges, necessitate adaptive nature solutions. Fraud detection from any transaction and forecasts are two further reasons why machine learnings are preferred over other programming languages. Machine learning, which is a subset of Artificial Intelligence (AI), assists in the extraction of knowledge patterns from large amounts of data. Machine learning is a technique for learning and improving an issue based on previous experiences [17] [18]. The core idea behind machine learning is to train and evaluate a model in order to build a new set of rules based on inference from source data [6].

Data mining, pattern recognition, and knowledge discovery from data (KDD) are more like ML. ML uses many mathematical formulas to extract information from earlier or historical data. The most common classifications for MLs are Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning [18] [19]. In the last year, machine learning has gained traction in a variety of industries, particularly those with large amounts of data. Telecommunications corporations are

inundated with many forms of massive data. As a result, such businesses are employing machine learning to forecast business losses or profits, detect fraud, anticipate customer churn, and assess customer happiness [8] [20] [21].

Machine learning is applied in different domains. The common application areas of machine learning algorithms include [8]:

- **Web Page Ranking :**

  It is a scenario where search engines return web pages in precedence of relevance based on user's query.

- **Collaborative Filtering:**

  It is a situation where systems predict user's need without the need for users to have explicit query. This query prediction is done using users previous experience. Good examples for collaborative filtering can be on-line book stores recommending users to buy additional books, and query prediction in Google search.

- **Automatic Document Translation:**

  Manual translation of documents from one language to other is tedious and error prone task. However, machine learning can be used for this task just by training the algorithms providing language translation examples.

- **Named Entity Recognition:**

  Is a process of identifying entities, such as places, titles, names, and actions, from a given document. Good example for this task is Apple's mail application where addresses are extracted from email and filled to the address book.

- **Telecommunication:**

Machine learning algorithms are applied in different tasks such as failure prediction of telecom Business Support System (BSS), detecting telecom frauds [8], designing customer churn prediction model where satisfaction level of a customer is analyzed and potential customers who may leave for another telecom service providers are identified , and intrusion detection systems where detection of anomalous traffics is done [8].

2.2.1. Supervised Learning

A labeled training dataset is used by supervised machine learning algorithms, and once the training is complete, the created model may predict the provided class labels for fresh datasets that were not used during the training phase. The main group of supervised learning algorithms is classification and regression. Some examples of supervised learning algorithms are; Decision Tree, Naive Bayesian Classifier, Rule-Based Classifier, the K-Nearest Neighbors Classifier, Neural Network, Linear Discriminant Analysis and Support Vector Machine [18]. Some of the algorithms discussed in detail in the coming sections.

2.2.1.1. Support Vector Machine (SVM)

SVM is a type of supervised machine learning approach that can be used to classify and predict data. The ability to evaluate and recognize patterns in a given input dataset is provided by SVM. The working principle of the algorithm (SVM) is to maximize the space margin between two classes to provide the maximum possible space between them. The vectors of the supplied instances are divided into two classes by a hyperplane [22]. A hyperplane is employed for two-dimensional qualities, while the kernel technique is used for high-dimensional features [18] [23]. Eq. 2.2.1.1-1 shows how to obtain the maximum distance between two vectors using a linearly discriminant function [24].

$$S(x) = WTX + b \quad (2.2.1.1\text{-}1)$$

Where,	S(x)	Linearly discriminant function.
	x	Feature vector (input vector).
	w	Adjustable weight vector to control direction of the hyperplane.
	b	Bias which control the hyperplane position.

On the training stage, SVM changes the weight of the input and combines the bias values to split each class, placing examples of class one (C1) on one side of the hyperplane and instances of class two (C2) on the other. Set the class of the new dataset instances using Eq. 2.2.1.1-2 and Eq. 2.2.1.1-3; according to the outcome of Y(x), the class is classified as C1 if Y(x) > 0 and C2 if Y(x) 0 [25]

$$Y(x) = w^{T}x + b > 0 \quad (2.2.1.1\text{-}2)$$

$$Y(x) = w^{T}x + b < 0 \quad (2.2.1.1\text{-}3)$$

The SVM kernel trick technique is utilized to fix the problem because the prior situation does not work for nonlinear separable datasets. Eq. 2.2.1.1-4 generates a smooth separating nonlinear decision boundary. SVM constructed a model that categorizes fresh examples into one of two categories using training datasets [18].

$$K(x_i, x_j) = e^{-(||x_i - x_j||^2/\delta^2)} \quad (2.2.1.1\text{-}4)$$

2.2.1.2. Decision Tree

Another popular supervised machine learning technique is Decision Tree. The flowchart is made with DT, which is a hierarchical and statistical process for categorizing datasets into the given categories. The statistical measure was used to select a root node characteristic from the experimental dataset, and the process was repeated until the end nodes, also known as leaf-nodes, were reached. Entropy and information gain are statistical metrics that DT employs in the development of the algorithm's tree structure. ID3, C4.5, and CART are the most common decision tree algorithms. DT creates a model that predicts the values of the target variable by learning simple decision rules learned from data properties. Eq. 2.2.1.2-1 is used to calculate information homogeneity using entropy.

$$H(x) = -\sum_i P(x_i) \log_2^{P(x_i)} \quad (2.2.1.2\text{-}1)$$

Where, H(x) Entropy of the dataset, and

$P(x_i)$ Probability of an instance being labeled with certain class given a specific feature.

As demonstrated in Eq. 2.2.1.2-2, an attribute's information gain is determined by subtracting the dataset's entropy from the target attribute/feature.

$$\mathrm{IG}(\mathrm{A},\mathrm{X}) = \mathrm{H}(\mathrm{x}) - \sum_{\mathrm{t}} -\mathrm{P}(\mathrm{t}) * \mathrm{P}(\mathrm{t}) \qquad (2.2.1.2\text{-}2)$$

Where,	H(x)	Entropy of the dataset, and
	IG(A,X)	Information gain of a specific attribute 'A'
	T	Subset created from splitting set S by attribute A

2.2.1.3. Artificial Neural Network (ANN)

A system based on a biological neural network, such as the brain, is known as an artificial neural network (ANN). According to [5] [26], neural networks are a brain symbol for information processing. ANNs can learn from their surroundings by iteratively altering their synaptic weight and bias level. The three types of neural networks include feed forward neural networks, recurrent neural networks, and self-organizing maps. There are three levels in each NN's network: input, hidden, and output [6].

2.2.2. Unsupervised Learning

Model evaluation is problematic with unsupervised machine learning methods because they do not employ labeled datasets during model training [5]. Clustering is a popular density estimation method that uses an unlabeled input dataset to try to figure out how many clusters or groups there are [25]. K-Means and Gaussian Mixture Model (GMM), two common unsupervised machine learning techniques, are examples [27] [28] [18].

2.2.2.1. K-Means

K-Means is a popular clustering technique that divides 'n' observations or instances into 'k' groups over a number of repetitions. The number of clusters is determined by end users, and each instance in a dataset is assigned to the cluster with the closest mean. Mean values are used as a cluster prototype in the K-Means algorithm. K-NN, a supervised machine learning algorithm, has a tenuous relationship with K-Means. By adjusting the value of k to 1 in k-NN, a 1-NN classifier can be used to clusters formed by k- Means to categorize new data into existing clusters (Nearest centroid classifier or Rochhio algorithm) [8].

The K-Means algorithm clusters data points or instances by selecting random 'k' means and then organizing them into 'k' clusters based on their distance from the means. Each iteration updates the mean values for each cluster depending on the data points within the clusters, and data points are grouped into the cluster with the closest mean. This procedure is done until the mean values and cluster data points do not change. Different clusters and means can be constructed each time the k-Means approach is used due to the random selection of initial k means, even though each data point belongs to exactly one cluster [8].

2.2.2.2. GMM

GMM is a clustering model for a mixture of 'M' Gaussian distributions with the goal of determining the best three model parameters for a particular dataset. The Gaussian representations' Mean and Covariance, as well as the weight of each Gaussian, are the model's parameters. The model that best matches the data is then found via GMM. Calculate the posterior probability of data instances using each component, and then assign each instance to a cluster based on the determined likelihood [8]. The definition of GMM with 'M' components is given in Eq. 2.2.2.2-1.

$$p(x|\theta_k) = \sum_{k=1}^{m} w_k \, p(x|\theta_k) \qquad (2.2.2.2\text{-}1)$$

Where, w_k Weight of the k^{th} component

$p(x|\theta_k)$ Covariance of the k^{th} component

θ_k Mean of the k^{th} component

The Expectation Maximization (EM) algorithm is used by GMM, and it necessitates the computation of initial mean and covariance estimations. The problem with EM is that if the initial estimates are wrong, it can get stuck in local optima. To avoid this problem, the K-Means algorithm can be used to calculate mean and covariance, which can then be supplied as input to EM. Two of the most common GMM application fields are speaker identification and biometric verification [8].

A range of machine learning technologies and frameworks have been used to create data mining jobs. WEKA, Matlab, Encog, IBM SPSS modeler, KNIME, LIONsover, Mlpy, SAS business miner, and Oracle data miner are some of the most often used tools and libraries [39]. WEKA, an open source tool with both a GUI and a command line option, was used in this investigation. [39].

For standard data mining tasks, WEKA offers a number of cutting-edge machine learning algorithms and data pretreatment tools built in. These operations include classification, clustering, association rule mining, regression, and attribute selection. The program assists with data collection, statistical evaluation of learning schemes, and display of the input and output of the learning process [8].

2.2.3. Semi-Supervised Learning

Labeled and unlabeled datasets are mixed together in this dataset. The majority of cases in semi-supervised machine learning labeled data are rare; yet, the model built by the semi-supervised ML algorithm predicts dataset classes better than the model generated by the supervised ML algorithm [18]. The given data in this sort of learning are a combination of classified and unclassified data. This mixture of labeled and unlabeled data is utilized to create a data categorization model that is acceptable. Labeled data is in short supply in most cases, while unlabeled data is plentiful. The goal of semi-supervised classification is to create a model that can better predict future test data classes than a model created just from labeled data [18].

Semi-supervised learning algorithm assumes the following concepts

- **Continuity Assumption**: This assumes that the points that are closest to each other will have the same output label.

- **Assumption of Clustering**: The data can be separated into discrete clusters, and points in the same cluster are more likely to have the same output label.

- **Manifold Assumption**: the data are roughly distributed over a manifold with a lower dimension than the input space. This assumption permits distances and densities defined on a manifold to be used.

There are two ways of learning process

- **Transductive Learning**

Does not generalize to unknown data during training and does not provide label data during training.

  - Assume labels
  - Train classifier on assumed labels

- **Inductive Learning**

This time, the system does generalization to previously unseen data and generates labels data during training, as well as applying manifold assumptions to the final classifier.

The following are some practical uses of semi-supervised machine learning methods.

- Speech Analysis
- Internet Content Classification
- Protein Sequence Classification

2.2.4. Reinforcement Learning

One technique for a machine learning system to collect data from environment interactions is to use it to take action in order to maximize rewards while minimizing risk [18].

The steps that reinforcement learning takes are as follows:

- The agent observe the input state
- The decision-making function is employed to motivate the agent to act.
- After the action is completed, the agent receives a reward or reinforcement from the environment.
- Information about the reward is recorded in the state-action pair.

3

System Design

The general experimental process employed in this investigation is examined in this chapter. Figure 0.1 shows the model's detailed experimental approach, which includes data collection, data pre-processing, and classification. The next section goes over each activity completed on this module in detail.

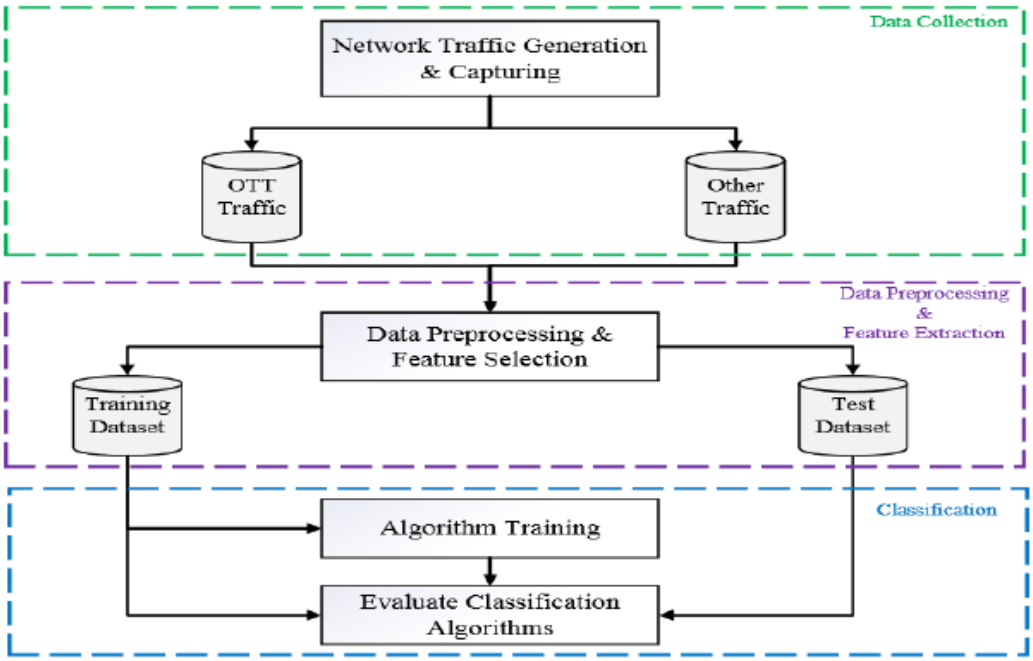

Figure 0-1. Overall experimental process [8]

3.1. Data Collection

Classification algorithms require labeled or unlabeled data, or a combination of the two. As a result, in order to perform this and related study, MISISDN-based OTT network traffic must be produced and captured [29] [30] [8]. As a result, network traffic data packets are collected and categorized [9] [31]. The Author [8] uses MSISDN-based OTT traffics and the architecture shown in Figure 3.1-1 to generate and capture network packets.

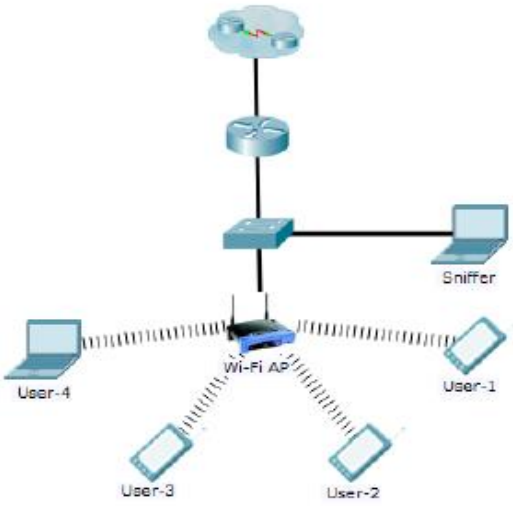

Figure 3.1-1. Network capturing architecture [8]

The two types of created network traffics are MSISDN-based OTT and non-MSISDN-based OTT. Tango, Viber, and Telegram, three of Ethiopia's most popular apps, were employed to collect MSISDN-based OTT traffics for this study. Other non-MSISDN-based OTT services used to generate network traffic packets include Skype, Yahoo Messenger, and other web-based packets like YouTube, Gmail, and Facebook.

The following site was used to generate and capture network traffic. The list of devices and software used while the infrastructure is being created is shown in Table 3.1-1.

Table 3.1-1. Infrastructure used for network traffic generation & capture

Type/Device	Quanti	Purpose
Dell Laptop (4 GB RAM and CPUs with 2.9 GHz clock rate	2	Traffic Generatin & Capturing
Samsung Smart-phone	2	Traffic Generatio
Huawei Smart-phone	1	"
Huawei Router	1	"
Cisco Switch	1	Traffic Generatin & Capturing
Wi-Fi Access Point	1	"
EPON Internet connection	100 Mb	
Wireshark 2.4.3	-	Traffic Capture

3.2. Data Preprocessing and Feature Selection

Network traffic datasets are manually generated for this type of study. To maintain data quality and reduce the amount of noisy data, preprocessing is required. Users' Internet Protocol (IP) based filtering, manual attribute selection, and outlier removal were all conducted as part of the data preprocessing methods.

3.3. IP based filtering

When using the Wireshark application to capture network traffic, all network traffic packets are gathered, and all transactions are captured, as previously indicated. As a result, not all packets reported in this study are meaningful. As a result, we only keep packets that originate from the user's IP address and reject any other network traffic packets that are unnecessary to the classification process.

3.4. Manual attribute selection

Wireshark is one of the 50 or more packet attributes that can be gathered with the tools listed above.

The previously described software, Wireshark, collects the properties of at least 50 packets. Only five of these characteristics were chosen, and they were based on a recent study [8] and previous relevant research assessments [31]. A list of qualities is provided below, along with a brief description of each [8].

- **Packet length**: The length of each packet crossing the real network, which is determined by the network's Hardware and Software. It has a numeric data type.

- **Delta time** is the difference in arrival times between two consecutive packets. In other words, delta time is the amount of time that has passed since the last packet arrived, and it can be used to calculate network roundtrip and server response times. It is a numeric data type and is also known as packet inter arrival time.

- **Relative time**: This is the amount of time that has passed since the first packet and the current packet. To put it another way, it's the total time it takes to capture the last packet after capturing the first. It has a numeric data type and is also known as cumulative time.

- **Cumulative byte**: When a huge block of data crosses the network, the quantity of data that can be communicated between the sender and recipient is referred to as the cumulative byte. It is the scale that is used to calculate the total number of bytes transmitted in a certain time interval from the collected traffic. It also has a numeric data type.

- **Protocol**: It specifies the protocol that is utilized in each packet. TCP, UDP, DNS, SMTP, Simple Network Management Protocol (SNMP), Simple Service Discovery

Protocol (SSDP), and Secure Sockets Layer (SSL) are some of the protocols that Wireshark captures.

3.5. Outliers detection and removal

A dataset with outliers, often known as noisy data, comprises numbers that do not fit the rest of the dataset. In order to improve the classification performance of algorithms, certain variables in a dataset must be identified and deleted. A few instances are [6] [8] and [32]. In this study, outliers are identified using the Inter Quartile Range (IQR) approach. Individual outcomes that deviate from the main trend of the datasets are known as outliers.

The IQR begins by sorting the entire dataset and separating it into four equal pieces. After that, determine the three quartile values (Q1, Q2, and Q3). The median and second-quarter readings are nearly identical. IQR seeks to find the upper and lower limits, often known as the fence, because outliers are detected outside of a defined border. Eq. 1 is used to compute the IQR value. 3.5-1.

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1 \tag{3.5-1}$$

The boundaries are calculated using an outlier's factors which has a basic value of '1.5', Eq. 3.5-2 and Eq. 3.5-3 shows the upper and lower boundary respectively.

$$\mathrm{Upper\_Limit} = \mathrm{Q}_1 + (1.5 * \mathrm{IQR}) \tag{3.5-2}$$

$$\mathrm{Upper\_Limit} = \mathrm{Q}_1 - (1.5 * \mathrm{IQR}) \tag{3.5-3}$$

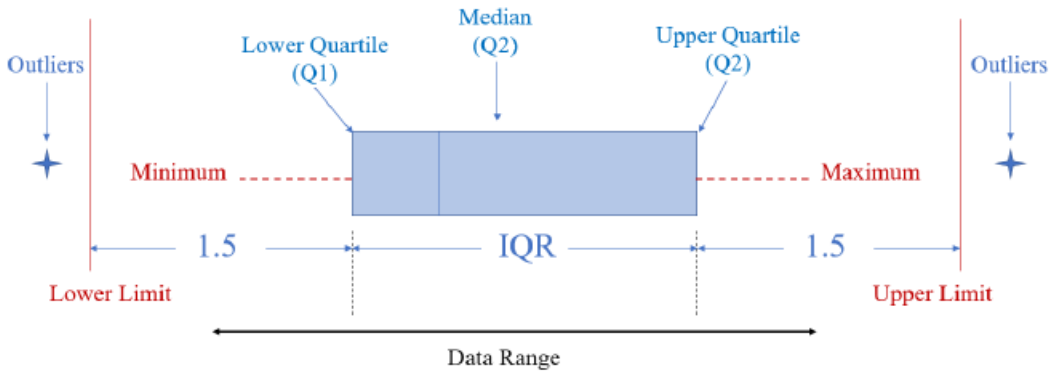

Figure 3.5-1 Pictorial Presentation of IQR [6] [8].

Following the above-mentioned data cleaning or preprocessing operations on the initial data, the datasets listed below are ready for the research's next model building process.

Table 3.5-1. Size of dataset after data preprocessing task

Packet Type	**Initial Data**	**Outlier Data**	**Traffic Size**
MSISDN-based OTT	444,877	12,957	431,920
Other (Non-MSISDN-based OTT	445,725	17,143	428,582

3.6. Algorithm Training

The next step is to train the given algorithm on the preprocessed data and generate a classification model after the preprocessing work is completed. For this investigation, two ML algorithms from two different ML algorithm categories were used (Supervised and Semi-supervised). Two supervised ML categories are DT Decision Tree and Collective. The semi-supervised machine learning algorithm category was narrowed down.

Table 3.6-1. Total dataset size after data preprocessing

Packet Type	Traffic Size
MSISDN-based OTT [**Y**es]	431,920
Other (Non-MSISDN-based OTT) [**N**o]	428,582

Semi-supervised machine learning techniques are employed with a few labeled datasets and a large number of unlabeled or unknown class datasets, as detailed in the previous sections. The whole dataset is shown in Table 3.6-1, from which we randomly chose 10% of datasets from each class MSISDN-based OTT and other type, which are labeled or known class types. Semi-supervised ML methods is a labeled class that covers both MSISDN-based OTT and other semi-supervised ML techniques (non MSISDN-base OTT). After the random selection, the remaining datasets are classified as unknown datasets, and they contain both dataset classes. The detailed percentage values for each dataset class are shown in Table 3.6-2.

Table 3.6-2. Semi-supervised ML Algorithm's dataset

Traffic Type	Data Size
10% of **Y**es	43,192
10% of **N**o	42,858
Unlabeled (**Y**es and **N**o)	774,452
Total Traffic	860,502

If there are enough labeled datasets for the experiment up to 40% of the dataset to be taken as OTT traffic class labeled, and the remaining 60% belongs to the other class label, None OTT traffic, for supervised ML algorithm labeled data ratio relies on the number of total data size. However, in the case of the Semi-supervised ML technique,

the ratio of labeled datasets is determined by the dataset's availability. When labeled datasets become limited, it is nearly impossible to determine the ratio for unlabeled datasets as well.

Table 3.6-2 shows that 20% of the dataset is labeled (including both OTT and non-OTT traffic), while the remaining datasets are not class labeled, making them suitable for Semi-supervised ML algorithms.

Table 3.6-3. Semi-supervised ML Algorithm's dataset

Packets Type	**Test dataset (60%)**	**Training dataset (40%)**
MSISDN-based OTT (Yes)	259,152	172,768
Other (Non-MSISDN-based OTT) (No)	257,149	171,433
Total Packets	516,301	344,201

3.7. Algorithm Evaluation

The primary goal of this study is to assess and compare the performance of categorization algorithms. The algorithms are evaluated using a ten-fold cross-validation technique. The ten cross-fold validation procedure divides the dataset into ten equal halves, with one serving as the training dataset and the remaining nine as the testing dataset. Each partition serves as a test and training dataset, and the process is repeated ten times. Many assessment measures are used to evaluate the performance of all machine learning models. The Confusion Matrix, Classification Accuracy, F-measure, Recall, Root Mean Squared (RMS), and ROC curve are some of the most frequent assessment matrices. The next sections go over the evaluation metrics indicated earlier.

3.7.1. Confusion Matrix

As the term says, when the algorithm becomes confused between the detected classes. A 2X2 matrix stores the True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values of categorization class labels.

Figure 3.7.1-1. Confusion Matrix

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

Where TP Number of instances correctly classified as class ‘Y’.

FP Number of instances that belongs to class ‘Y’ but classified as class ‘N’).

FN Number of instances that belongs to class ‘N’ but classified as class ‘Y’.

TN Number of instances correctly classified class ’N’.

3.7.2. Classification Accuracy

The classification accuracy is calculated using the ratio of successfully categorized (class 'Y' and class 'N') items to the total test dataset. Eq. 3.7.2.1 represents the fraction of properly classified events, demonstrating classification accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3.7.2\text{-}1)$$

3.7.3. F-Measure

The F-measure is a harmonic mean of precision and recall that is calculated as shown in Eq. 3.7.3.1.

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.7.3\text{-}1)$$

Recall The ratio of correctly classified cases as Normal (class 'Y') divided by the sum of correctly and wrongly classified instances of normal (class 'Y') as (class 'Y') and (class 'N') is calculated. Fraudulent (class 'N') recall was calculated in the same way using Eq. 3.7.3-2.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.7.3\text{-}2)$$

Precision for class ’A’ denoted by Eq. (3.7.3-3) measures the ratio of instances correctly classified as class ’A’ with respect to total number of instances classified as class ’A’. Precision for class ’B’ is calculated in the same way.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.7.3\text{-}3)$$

3.7.4. ROC curve

The ROC curve graphically depicts the True Positive Rate (TPR) and False Positive Rate (FPR) (FPR). The FPR and TPR are represented on the X and Y axes, respectively. When the ROC curves are too close to the top-left corner of the area, the technique is utilized as a perfect classifier. The algorithm is categorized as a low-level classifier when the ROC curves are below the graph's linear line (X=Y).

In addition to the assessment metrics described above, the time spent by the algorithms to classify the provided dataset is taken into account when comparing algorithm performance.

4

Results and Discussion

This Chapter describes the experiment results of the research with ten cross-fold validation technique for both selected supervised and Semi-supervised machine learning algorithm.

4.1. Model Evaluation

Due to the lack of OTT traffic dataset, the main purpose of this study is to classify MSISDN-bass OTT traffic packets using a semi-supervised machine learning technique. Compare the findings to those of previously chosen supervised machine learning methods. Because we are using the WEKA tool, which is a GUI software, we can easily change the dataset. Despite the fact that WEKA is well-known for supervised and unsupervised machine learning algorithm dataset types, working with the Semi-supervised mix of the two ML dataset kinds caused some challenges. As a result, the University of WEKA has created a package that can deal with datasets produced by semi-supervised machine learning algorithms. For this research project, the packages must be imported into the WEKA tool.

We used a semi-supervised algorithm dataset to evaluate the model's overall accuracy in this study. Both semi-supervised and supervised machine learning methods were used in the experiment. Both algorithms are evaluated using the ten cross-fold assessment technique. The Collective filter semi-supervised machine learning algorithm is outperformed by the Decision tree of a supervised machine algorithm. The decision tree had a 99 percent accuracy rate, but the Collective filter classifier only had a 91 percent accuracy rate. The major goal of this study is to come up with a novel

method for categorizing and detecting MSISDN-based OTT packets, as well as to look at how they are classified and discovered. Another technique for handling these challenges is to use semi-supervised algorithms, which can help you get started with semi-supervised datasets. Due to the difficulty in collecting OTT traffic packets, this work can be regarded a step forward in detecting MSISDN-based OTT frauds. The overall accuracy of the chosen machine learning algorithms is depicted in Figures 4.1-1 and 4.1-2, respectively (DT and Collective Filtered). The numerical performance values of the selected approaches are shown in Table 4.1-1.

Table 4.1-1 Confusion Matrix of the two algorithm

DT

Confusion Matrix	Class 'Y'	Class 'N'
Class 'Y'	431,736	184
Class 'N'	288	428,294

Collective Filtered

Confusion Matrix	Class 'Y'	Class 'N'
Class 'Y'	391,429	40,491
Class 'N'	35,524	393,058

In addition to the numerical evaluation data the confusion matrix, overall accuracy can be represented in graph and it shows on the bellow two graphs.

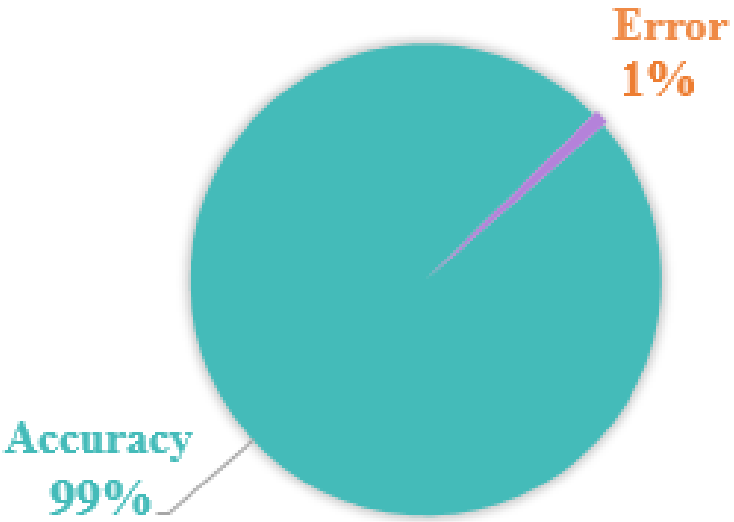

Figure 4.1-1. Overall accuracy Performance accuracy of DT

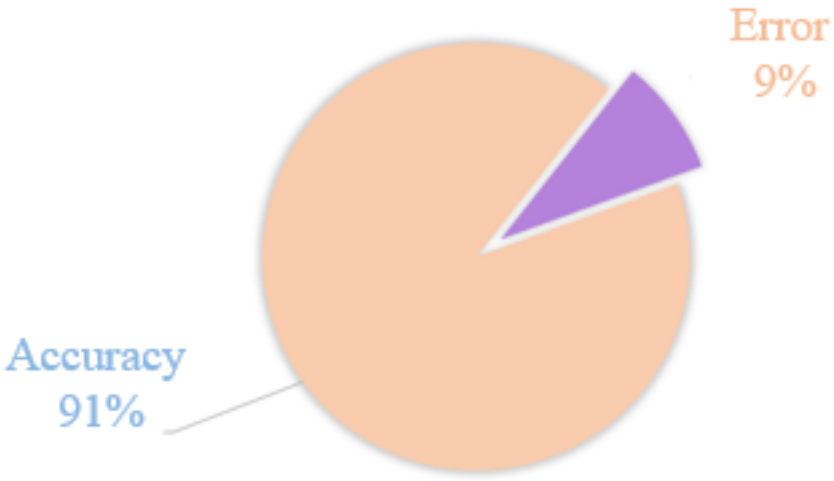

Figure 4.1-2. Overall accuracy Performance accuracy of Collective Filtered

F-masseur is one way of evaluating metrics of classification algorithm; the bellow Table 4.1-2 shows the performance of the algorithms with the relationship of Recall and Precision of the algorithm.

Table 4.1-2 F-measure of the algorithm

Evaluation Metrics	Algorithms Result	
	DT	Collective Filtered
Recall $\frac{TP}{TP + FN}$	0.99	0.91
Precision $\frac{TP}{TP + FP}$	1	0.92
F-measure $\frac{2 * Precision * Recall}{Precision + Recall}$	0.99	0.91

Another way to measure algorithm performance is to see how long it takes to create and analyze a model. Figure 4.1-3 depicts the entire model creation and evaluation process. When compared to Collective filtering, developing and evaluating a model with the Decision tree algorithm takes less time. Despite having a lower outcome than the DT algorithm, the collective filtering algorithm's overall performance and model construction time are appropriate. The performance of each method is influenced by the fact that each has a different training and testing dataset.

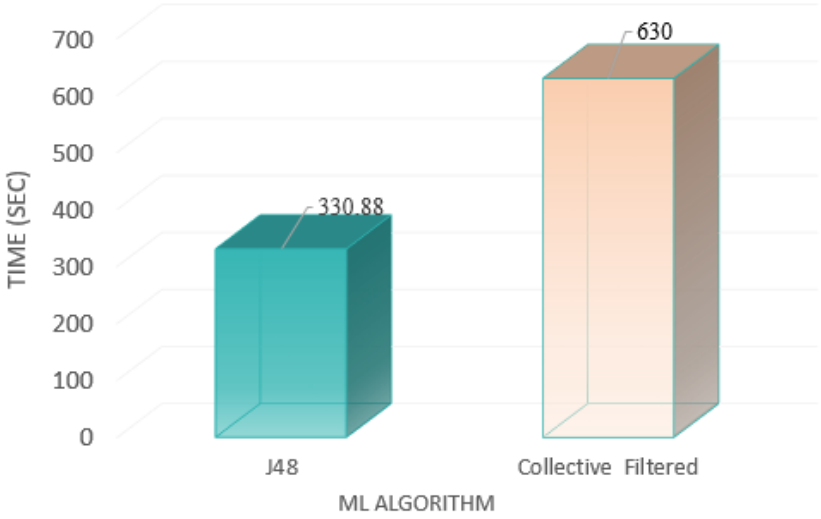

Figure 4.1-3 Model built time of the algorithms

5

Conclusion and Recommendation

5.1. Conclusion

Tariffs from international calls help telecommunications providers in underdeveloped countries cover the costs of expansion. Interconnected by pass fraud takes advantage of this arrangement by charging lower rates to callers and taking revenue away from operators. Their actions have resulted in customer dissatisfaction, income loss, and security concerns. Interconnected bypass fraudsters are targeting telecom service providers as call interconnection prices rise. To detect OTT network traffic packets, machine learning algorithms are applied in this study. Detecting and terminating fake OTT calls is one of the 21/59 scam remedies offered. This research looked into the applicability of machine learning approaches for detecting OTT voice call packets, as well as the detection mechanism.

The supervised algorithm, DT., uses labeled datasets, while the semi-supervised technique uses collective filtering. Both algorithms were evaluated using the ten cross-validation process. Packets are identified with the appropriate class during the traffic production phase in the controlled laboratory setting. On the other hand, each training and testing dataset is built independently.

In the performance assessment, both the DT collective algorithm and the DT algorithm have a high performance value; nevertheless, when the two algorithms are compared, the DT algorithm outperforms the collective algorithm. Similarly, it takes less time to build and evaluate the DT algorithm than it does to develop and analyze the collective filtered algorithm. The Collective Filtered ML algorithm achieves a high level of accuracy, with 91 percent accuracy. Getting such a useful result with a small amount of

labeled data and a vast amount of unlabeled data is an excellent example of how semi-supervised may be used to solve a major problem. As a consequence, the Collective filtered ML approach may be rated the top performer in detecting voice call packets from OTT data while using ten cross-fold validation.

5.2. Recommendations for Future Work

To improve the technique's performance and accuracy, we recommend that data aspects of network packets be refined further in future study. While the focus of this thesis is on OTT by pass fraud network traffic packet classification and detection, considering the large number of frauds in the market, similar study on other fraud kinds and approaches is encouraged.

Because this study uses the same attribute types as the previous one, additional and different attributes could be added to the attribute list. Different assessment procedures can also be employed to assess the outcome.

References

[1] Lin and C. Z. a. Z, "Study on Fraud Detection of Telecom Industry Based Rough Set," p. pp. 15–19, 2018.

[2] C. Xing and D. M. Isaacowitz, "FRAUD DETECTION IN MOBILE COMMUNICATION NETWORKS USING DATA MINING," Motiv. Emot, vol. 30, p. 243–250, 2006.

[3] I. Ighneiwa and H. Mohamed, "Bypass fraud detection: Artificial intelliger approach," no. arXiv preprint arXiv:1711.04627, 2017.

[4] M. Sahin and A. Francillon, "Over-the-Top bypass: Study of a recent telephony fraud," M. Sahin and A. Francillon, “Over-the-Top bypass: Stud of a recent telephony fraud,” in Proc. SIGSAC Conf. on Comp. and Comm Security, ACM, 2016, pp. 1106–1117, no. in Proc. SIGSAC Conf. on Com and Comm. Security, p. 1106–1117, 2016.

[5] K. Hagos and A. Ababa, "SIM-Box Fraud Detection Using Data Mining Techniques : The Case of ethio telecom," 2018.

[6] F. Tesfaye and A. Ababa, "Near-Real Time SIM-box Fraud Detection Usi Machine Learning in the case of ethio telecom," 2020.

[7] M. Sahin and A. Francillon, "Over-the-Top bypass: Study of a recent telephony fraud," in Proc. SIGSAC Conf. on Comp. and Comm. Security, ACM, p. 1106–1117, 2016.

[8] T. Hailu and A. Ababa, "Network Traffic Classification Using Machine Learning: A Step Towards Over-the-Top Bypass Fraud Detectio," 2018.

[9] J. Datta, N. Kataria and N. Hubballi, "Network traffic classification in encrypted environment: A case study of Google Hangout," in Nat. Conf. o Commun. , no. Mumbai, India, pp. 1-6, 2015.

[10] M. Rathore, A. Paul, A. Ahmad, M. Imran and M. Guizani, "High-speed network traffic analysis: Detecting VoIP calls in secure big data streaming in IEEE Conf. on Local Comput. Netw., Dubai, UAE, pp. 595-598, 2016.

[11] M. Korczy´ nski and A. Duda, "Classifying service flows in the encrypted skype traffic," in Int. Conf. on Commun. (ICC), IEEE, p. 1064–1068, 201

[12] B. Reaves, E. Shernan, A. Bates, H. Carter, and P. Traynor, "Boxed out: Blocking cellular interconnect bypass fraud at the network edge," in in USENIX Symp. on Security, p. 833–848, 2015.

[13] T. K. Sawe, "Emergence of OTT communication services and sustenance revenue among kenya telcos," Int. J. of Innov. Sci., Eng. and Technol, vol. no. No. 8, p. 377–381, 2016.

[14] J. Sujata, S. Sohag, D. Tanu, D. Chintan, P. Shubham, and G. Sumit, "Imp of Over-the-Top (OTT) services on telecom service providers," Indian J. o Sci. and Technol, vol. 8, no. No. S4, p. 145–160, 2015.

[15] I. Society, "Artificial intelligence and machine learning : Policy paper," 2017.

[16] S. Shalev-Shwartz and S. Ben-David, "Understanding Machine Learning: From Theory to Algorithms," Cambridge: Cambridge University Press, 2014.

[17] S. Marsland, "Machine Learning An Algorithmic Perspective," 10.1017 / CBO9781107298019, 2014.

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

[19] T. Oladipupo, "Types of machine learning algorithms," New Advances in Machine Learning, no. doi: 10.5772/9385, 2010.

[20] J. Y. Lee, J. H. Lee, J. S. Yeo, and J. J. Kim, "A snp harvester analysis tobetter detect snps of ccdc158 gene that are associated with carcass qualitytraits in hanwoo," Asian-Australasian Journal of Animal Sciences, vol. 26, no. no. 6, p. 766–771, 2013.

[21] N. T. A. S. a. A. Sharma, "A review of supervised machinelearning algorithms," 2016.

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

[23] N. T. A. S. a. A. Sharma, "A review of supervised machine learning algorithms," 2016.

[24] "Neural Networks and Learning Machines," McMaster University Hamilto Ontario Canada, vol. 3, no. Pearson Education, 2009.

[25] Y. B. a. M. Ozuysal, "Introduction to Machine Learning Second Edition," vol. 1107, no. Second, p. 105–28, 2014.

[26] P. Gaur, "Neural networks in data mining," International Journal of Electronics and Computer Science Engineering, vol. 1, no. 3, 2013.

[27] E. F. M. H. a. C. P. I. Witten, Practical Machine Learning Tools and Techniques, San Francisco: Morgan Kaufmann, 2016.

[28] A. S. a. S. Vishwanatha, Introduction to Machine Learning, Cambridge : Cambridge University Press, 2008.

[29] R. A. a. N. Zincir-Heywood, "Machine learning based encrypted traffic classification: Identifying ssh and skype," for Security and Defense Applic no. (CISDA),, pp. 1-8, 2009.

[30] N. Z.-H. a. A. A. F. R. Alshammari, "Performance comparison of four rule sets: An example for encrypted traffic classification," in World Congr. on Privacy, Security, Trust and the Manage. of e-Business, no. IEEE, pp. 21-2 2009.

[31] M. A.-K. N. A.-S. a. S. S. G. Al-Naymat, "Classification of VoIP and non-VoIP traffic using machine learning approaches," J.of Theoretical and App Inform. Technol, vol. 92, no. no. 2, p. 403, 2016.

[32] S. M. A. K. Z. H. A. R. B. S. Qayyum, "Fraudulent call detection for mobi networks," in Int. Conf. on Inform. and EmergingTechnologies (ICIET), n IEEE, p. 1–5, 2010.

Appendix

A. IEEE version of this research

Semi-Supervised Algorithm to Detect Over-The-Top Bypass Fraud: in the case of ethio telecom

Wubalem K/Michael
Addis Ababa Institute of Technolog
Addis Ababa University, Addis Abab
Ethiopia
wubalem8281@gmail.com

Yihenew Wondie
Addis Ababa Institute of Technology
Addis Ababa University, Addis Ababa,
Ethiopia
yihenew.wondie@aait.edu.et

Abstract

Telecom frauds' behavior is changing through time and their capability as well growing rapidly to cop-up with telecom service providers. Inter connected bypass fraud is one of the top telecom fraud category in the world, Over-The-Top (OTT) bypass fraud is one family of interconnected bypass telecom frauds. It is a recent type of telecom fraud type, which works by hijacking international voice calls and transfer to VoIP to terminate as an OTT call type without the knowledge of telecom operator, caller and called party.

Mobile Station International Subscriber Directory Number (MSISDN) based OTT frauds increase being threats for telecom companies, due to the increasing of smart phones and being easy to get access OTT service anywhere. Such activities are called interconnected bypass fraud, OTT bypass also one of the interconnected bypass fraud type. Detecting OTT voice call packets through different network traffic classification techniques is one subtask in the detection of this fraud.

Machine learning (ML) algorithms are used to classify the network traffic packets, Semi-supervised and supervised algorithm Collective Filtered and Decision Tree (DT). Ten cross-fold validation technique for the training and testing the selected algorithm. Test dataset is properly preparing for each ML algorithm. Both ML algorithms DT and collective filtered achieves better performance of 99% and 91% accuracy respectively with acceptable model build and evaluation time.

KEYWORDS: OTT bypass fraud, Telecom fraud, Network traffic classification, Machine-learning algorithms

I. Introduction

Telecommunication companies had suffered with revenue losses above billions of dollars on a world scale because of fraudsters [1]. In general, telecom fraud defined as using operator's telecom infrastructures or services with no intention paying for it [2]. Due to the rapid growth of telecommunication services fraudsters as well increased their ways of attack using the advantages of technology advancement. Currently telecom frauds are a lot in number, the top three telecom fraud types are International Revenue Share Fraud (IRSF), Interconnect bypass Fraud and Premium Rate Service Fraud (PRSF) [3]. Over-The-Top (OTT) bypass fraud is one of interconnect telecom fraud family [4].

In OTT bypass, a normal voice calls are directed to Voice over IP through Internet and converted to a voice chat application on a smartphone, instead of being terminated over the normal voice call telecom infrastructure. This kind of rerouting (or hijacking) supported by an international transit operator in coordination with the OTT service provider, but without explicit authorization from the caller, called and their operators [4].

Ethio telecom, which is the sole telecom service provider in Ethiopia, is among the telecom companies highly affected due to interconnect bypass fraud. Currently, the company is using a rule based Fraud Management System (FMS) to detect and prevent interconnect bypass frauds and other telecom frauds. SIMbox fraud is the commonly known interconnected bypass fraud type, which also

highjacks an international call transfer to VoIP and terminate it as local call [5] [6]. Over-The-Top (OTT) bypass fraud is also among the recent type of telecom interconnect bypass fraud where a call initiated as non-OTT call is rerouted through OTT applications and received as an OTT call at the receiver side [5].

II. Literature Review

Three machine learning algorithms Adaptive Booster (AdaBoost) + J48, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and Support Vector Machine (SVM) is used to detect MSISDN-based OTT which takes a sample from Viber, Tango and Telegram. The Author [7] generate 1.7 million labeled packets for evaluation using Ten cross-fold and separate test data validation techniques. The result of each machine learning algorithms' compared with each other, AdaBoost + J48 algorithm achieved better perform on both evaluation techniques (ten cross-fold validation and separate test data validation) MSISDN-based OTT fraud classification.

Over-The-Top (OTT) bypass fraud is one of interconnect telecom fraud family. Merve Sahin et al. [4] collect a large share of the call charge and induce a significant loss of revenue to the bypassed operators. Since any interconnected bypass frauds degrades the quality of the telecom service without any benefits the users. The authors [4] find out up to 83% of calls being subject to OTT bypass. Additionally, they show that OTT bypass degrades the quality of service, and sometimes collide with other fraud schemes, exacerbating the quality issues.

Four packet features; namely, packet length, delta time (packet inter-arrival time), cumulative byte, and relative time; are used for classification claiming that these features have been used for the first time in internet traffic classification. The scholars have stated that AdaBoost achieves the best overall accuracy (98.3 %) while MLP is the least performer with an accuracy of 84.2 % in classifying the five applications. From the test results reported, AdaBoost is also the best classifier in detecting VoIP applications specifically compared to the others [8].

Telecom operators have used different fraud detection approach to overcome interconnected bypass and other telecom fraud types, due to the difficulty in detecting and quantifying the impact of this fraud, telecom companies are losing a huge amount of revenue even without knowing the existence of this fraud [7]. The major approaches are Test Call Generation (TCG), Call Detail Record (CDR), data traffic, and audio fingerprinting is required to minimize the impact of OTT bypass fraud

III. Statement of the problem

Telecom service providers investing a huge amount of money to provide their services. However, telecom operators' operation and revenues highly affected by telecom fraudsters. Telecom providers lost a large amount of revenues due to interconnected bypass frauds as mentioned on the previous section. OTT bypass fraud is one type of interconnected bypass fraud that is associated with users' MSISDN number. OTT services app like WhatsApp, Viber, and Telegram required MSISDN number at the time of registration as a user ID [7].

With the best of my knowledge, currently ethio telecom's FMS have no detection mechanism for OTT bypass fraud and there is no explicit revenue lose report caused by OTT bypass fraud. Since the company, losing a huge amount of foreign currency because of international calls terminated inappropriately. So, such OTT bypass frauds needs to be mitigated and study how ethio telecom is impacted with such fraud types.

Telecom operators gives different types of services by building their infrastructures and make money out of it. As previously explained telecom fraudsters uses telecom operators' infrastructure to make their own money without sharing or paying for it.

Telecom operators, literatures and other concerned bodies have made different fraud detection mechanisms and techniques, also there are many comparisons have made among with different detection technique. Recently there is a research done on OTT bypass fraud in the case of ethio telecom, by Tewodros Hailu using generated data

with the title “Network Traffic Classification Using Machine Learning does a recent research: A Step Towards Over-the-Top Bypass Fraud Detection”. Therefore, ethio telecom shall take the real network traffic data as consideration in the process of Over-the-Top Bypass frauds detection. There will be a way to compute how revenue loss that ethio telecom faced. This research result will compare with the previous research result.

IV. Methodology

This chapter discusses overall experimental process conducted through this research. Figure 1 shows the experimental process of the model;

Data collection, Data Pre-processing, and Classification. Each detail tasks performed under these modules are describe in the next section.

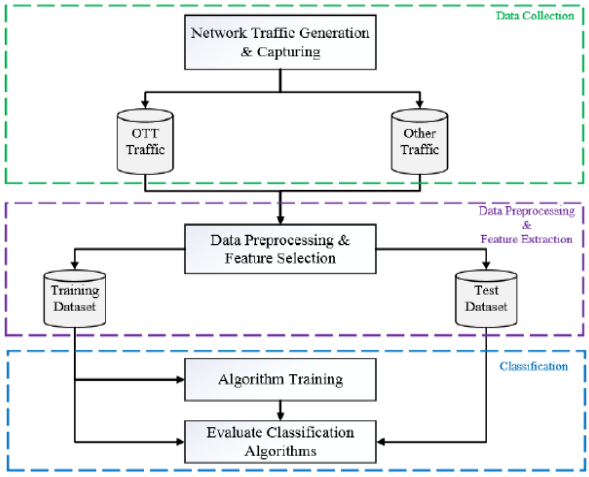

Figure 1. Overall experimental process [7]

1. Data Collection

Classification algorithms requires either labeled or unlabeled data, or the combination of the two data types.

To generate and capturing the network traffic the following resource have been used. Table 1 shows the list of device and software used while the infrastructure is built.

Table 1. Infrastructure used for network traffic generation & capture.

Type/Device	Quantity	Purpose
Dell Laptop (4 GB RAM and CPUs with 2.9 GHz clock rate	2	Traffic Generating & Capturing
Samsung Smart-phone	2	Traffic Generation
Huawei Smart-phone	1	"
Huawei Router	1	"
Cisco Switch	1	Traffic Generating & Capturing
Wi-Fi Access Point	1	"
EPON Internet connection	100 Mb/s	
Wireshark 2.4.3	-	Traffic Capture

2. Data Preprocessing

Network traffic datasets are generated manually for the need of such research type. To maintain the data quality and reduce the level of noisy data, preprocessing task required.

3. Feature Selection

User’s Internet Protocol (IP) based filtering, manual attribute selection and Outlier removals have done under the data preprocessing tasks.

- **IP based filtering**

All network traffic packets are captured as mentioned above, while capturing network traffic using Wireshark tool, all transactions are captured. Therefore, not all captured packets are useful for this research. Due to that, we keep only packets coming from user’s IP address and remove the rest of the network traffic packets because they are not relevant for the classification process.

- **Manual attribute selection**

Among of 50 and more packets attributes that are captured using previously stated tools, which is called Wireshark. Previously stated tool that is call Wireshark capture 50 and more packets' attributes.

Table 2. Size of dataset after data preprocessing

Packet Type	Initial Data	Outlier Data	Traffic Size
MSISDN-based OTT	444,877	12,957	431,920
Other (Non-MSISDN-based OTT)	445,725	17,143	428,582

4. Algorithm Training

Once the preprocessing task is completed, the next step is training the selected algorithm using the preprocessed data and build classification model. For this research, two ML algorithms are selected from two different ML algorithm categories (Supervised and Semi-supervised). From supervised ML category DT

Decision Tree and Collective Filtered from Semi-supervised ML algorithm category.

5. Algorithm Evaluation

The main target of this research is to evaluating and comparing performance of classification algorithm. Ten cross-fold validation technique is used to evaluate the algorithms. Ten cross-fold validation technique is working as follows, the dataset is divided in to 10 equal parts, then only one of the part is used as training dataset and the remaining nine part of the section is used for testing. This process is repeated ten times until each partitions used as both test and training dataset. All ML models evaluate their performance using different evaluation metrics. The common evaluation matrices are Confusion Matrix, Classification accuracy, F-measure, Recall, Root Mean Squared (RMS) and ROC curve. The listed evaluation metrics are discussed in the coming pages.

Confusion Matrix

As the name indicates when the algorithm gets confused between the labeled classes. A 2X2 matrix contains True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values of classification class labels.

Table 3. Confusion Matrix

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

Where,

TP	Number of instances correctly classified as class 'Y'.
FP	Number of instances that belongs to class 'Y' but classified as class 'N').
FN	Number of instances that belongs to class 'N' but classified as class 'Y'.
TN	Number of instances correctly classified class 'N'.

Classification Accuracy

Classification accuracy is used to measure the ratio of correctly classified (class 'Y' and class 'N') with respect to the whole test dataset. Classification accuracy denoted by Eq.1 shows the percentage of correctly classified instances

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

F-Measure

A harmonic mean of precision and recall is the F-measure and the measurement is calculated as shown below Eq. 2.

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Recall Measures the ratio of correctly classified instances as Normal (class 'Y') divide by the sum of correctly and incorrectly classified instances of normal (class 'Y') as (class 'Y') and (class 'N'). Recall for Fraudulent (class 'N') computed the same way using Eq**.** (3).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Precision for class 'A' denoted by Eq. 4 measures the ratio of instances correctly classified as class 'A' with respect to total number of instances classified as class 'A'. Precision for class 'B' is calculated in the same way.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

V. Results and Discussion

This Chapter describes the experiment results of the research with ten cross-fold validation technique for both selected supervised and Semi-supervised machine learning algorithms.

Model Evaluation

The main target of this research is to classify MSISDN-bass OTT traffics packets using Semi-supervised ML algorithm, which is the fact that OTT traffic dataset are scarce. In addition, compare the result with previously selected supervised ML algorithm results. Since we are using WEKA tool, which is a GUI software, enable us easily to manipulate the dataset. Even though WEKA is popular for supervised and unsupervised machine learning algorithm dataset types, but there were some difficulties to handle the combination of the two ML dataset types which is called Semi-supervised. Due to that, University of WEKA has developed a package, which can support semi-supervised ML algorithm datasets. So, the packages need to be imported in to the WEKA tool for this research experiment.

In this research, we have used semi-supervised algorithm dataset and measure the overall accuracy of the model. Not only semi-supervised, supervised machine learning algorithm as well used for the experiment. Both algorithms are evaluated using 10 cross-fold evaluation technique. A supervised machine algorithm's Decision tree achieves better performance than Collective filter semi-supervised machine learning algorithm. Decision tree performance achieved 99% of accuracy and Collective filter classifier achieve 91% of performance. The main target of this research to show the alternative way to classify and detect MSISDN-based OTT packets and show how being Semi-supervised algorithms are another way of obtaining solution for such types of research solution and helps to start detection using semi-supervised datasets. Since getting OTT traffic packets is difficult, this research can be considered as another step to detect MSISDN-based OTT frauds.

Table 4. Confusion Matrix of the two algorithm

Confusion Matrix	Class 'Y'	Class 'N'
Class 'Y'	431,736	184
Class 'N'	288	428,294

Confusion Matrix	Class 'Y'	Class 'N'
Class 'Y'	391,429	40,491
Class 'N'	35,524	393,058

In addition to the numerical evaluation data the confusion matrix, overall accuracy can be represented in graph and it shows on the bellow two graphs.

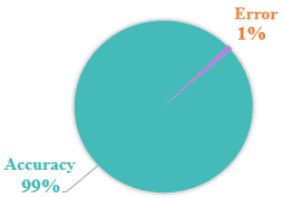

Figure 2. Overall accuracy Performance accuracy of

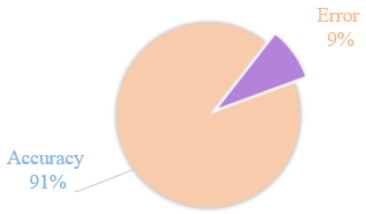

Figure 3. Overall accuracy Performance accuracy of Collective Filtered

F-measurement

F-masseur is one way of evaluating metrics of classification algorithm; the bellow Table 0-2 shows the performance of the algorithms with the relationship of Recall and Precision of the algorithm.

Table 5. F-measure of the algorithm

Evaluation Metrics	Algorithms Result	
	DT	Collective Filtered
Recall = $\frac{TP}{TP + FN}$	0.99	0.91
Precision = $\frac{TP}{TP + FP}$	1	0.92
F-measure = $\frac{2 * Precision * Recall}{Precision + Recall}$	0.99	0.91

Model build and evaluation time is another way to measure the performance of algorithm and compare with each other. Decision tree algorithm takes less amount of time to build and evaluate the model as compared with Collective filtered. Even if the figures show less result for collective filtered algorithm compared with DT algorithm, its overall performance and model build time is acceptable result. Since we are using different training and testing dataset for each algorithm, their performance as well influenced.

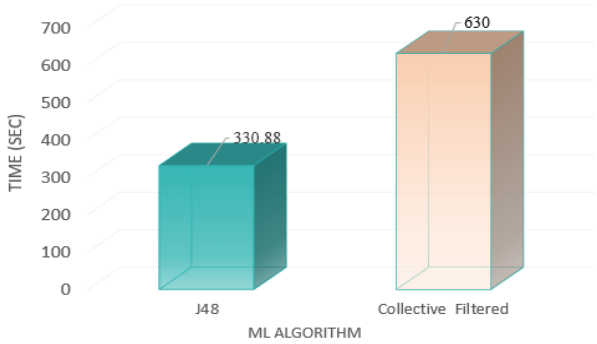

Figure 4. Model built time of the algorithms

Conclusion

Telecommunication operators in developing countries subsidize their cost of expansion by tariffs collected from international calls. Inter connected by pass fraud abuse this scenario by delivering less expensive price to callers and divert the revenue from operators. Their impact ranges from customer dissatisfaction to revenue loss and security issues. Interconnected bypass fraudster targets telecom service providers with higher call interconnection fee. In this research machine learning algorithms in the detection of OTT network traffic packet. Detecting and blocking fraudulent OTT calls is one of the recommended solution of this fraud. In this research, the applicability of machine learning

In the performance evaluation, Both DT collective algorithm perform high performance value, when we compare each algorithm, DT has a higher value than collective algorithm. Similarly, the time that takes to build and evaluate is less value is recorded for DT than collective filtered algorithm. The recorded 91% accuracy is a good performance

achievement of the Collective Filtered ML algorithm. Achieving such valuable result with less amount of labeled dataset and

Technique. Labeling of packets with the corresponding class is done during the traffic generation phase while using the controlled laboratory environment. However, each training and testing datasets are prepared independently. huge amount of unlabeled dataset, a good way of showing Semi-supervised is one way of a huge problem solving mechanism. So, Collective filtered ML algorithm can be considered as a best performer in detecting voice call packets through OTT traffic using ten cross-fold validation.

References

[1] Lin and C. Z. a. Z, "Study on Fraud Detecti of Telecom Industry Based on Rough Set," p. p 15–19, 2018.

[2] C. Xing and D. M. Isaacowitz, "FRAUD DETECTION IN MOBILE COMMUNICATIC NETWORKS USING DATA MINING," Motiv Emot, vol. 30, p. 243–250, 2006.

[4] I. Ighneiwa and H. Mohamed, "Bypass frau detection: Artificial intelligence approach," no. arXiv preprint arXiv:1711.04627, 2017.

[5] M. Sahin and A. Francillon, "Over-the-Top bypass: Study of a recent telephony fraud," M. Sahin and A. Francillon, “Over-the-Top bypass Study of a recent telephony fraud,” in Proc. SIGSAC Conf. on Comp. and Comm. Security, ACM, 2016, pp. 1106–1117, no. in Proc. SIGS Conf. on Comp. and Comm. Security, p. 1106–1117, 2016.

[6] K. Hagos and A. Ababa, "SIM-Box Fraud Detection Using Data Mining Techniques : The Case of ethio telecom," 2018.

[7] F. Tesfaye and A. Ababa, "Near-Real Time SIM-box Fraud Detection Using Machine Learning in the case of ethio telecom," 2020.

[8] T. Hailu and A. Ababa, "Network Traffic Classification Using Machine Learning: A Step Towards Over-the-Top Bypass Fraud Detectio, 2018.

[9] J. Datta, N. Kataria and N. Hubballi, "Netw traffic classification in encrypted environment: case study of Google Hangout," in Nat. Conf. o Commun. , no. Mumbai, India, pp. 1-6, 2015.