Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Machine Learning Based Traffic Classification Algorithm for Fixed Network Traffic

A Thesis Submitted to the School of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Engineering

By: Selamawit Bayu

Advisor: Dr. Sosina Mengistu

Declaration

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

Author 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 **Selamawit Bayu**, entitled **Machine Learning Based Traffic Classification Algorithm for Fixed Network Traffic** and submitted in partial fulfillment of the requirements for the degree of Master of Science in 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 _____

Advisor Dr. Sosina Mengistu Signature_____Date _____

Dean, School of Electrical and Computer Engineering _____

Acknowledgment

I would like to express my sincere gratitude to my advisor Dr.Sosina Mengistu for the uninterrupted support of my research, for her patience, motivation, passion, and enormous knowledge. Her guidance helped me in all the time of research and writing of this thesis. Besides my adviser, I would like to thank ethio telecom expertise for their encouragement, insightful comments, and delivering appropriate data. Finally yet importantly, I would like to thank my family: my parents, for giving birth to me in the first place and supporting me spiritually throughout my life.

Abstract

Traffic classification is associating network flows with the applications that generate them. Traffic classification helps ISPs (internet service providers) as the fundamental building block for any traffic management activity, for traffic pricing and treatment (e.g., policing, shaping, etc.) and for security activities. There are various types of methods used for network traffic classification. Port-based and payload-based methods is widely used for application identification in the traffic. In recent years, these methods have not worked well in practice. This is because the number of applications that employ random or non-standard ports have increased dramatically, and payload content encryption is required for security purposes. Therefore, machine-learning techniques have been proposed as solutions in the literature, recently.

In this study, a machine learning method is used for the identification of applications using fixed network traffic data collected from ethio telecom access layer devices. To build the model, two supervised machine-learning algorithms, namely Random Forest and C4.5are selected from the state of the art. The flow level network features extracted from the collected data to train the machine-learning model. This study is unique from existing network traffic classification studies in that it uses two additional new features to train the model. These are the flow index and flow state. The performance of the models analyzed before and after the addition of new features. Finally, application dominance in terms of flow, packet, and byte composition in fixed network traffic is studied.

The experiment results show that Random Forest provided 90.8% and C4.5 provides 88% of the overall accuracy-based on the flow features available in the state of the art. However, after the addition of the flow state and flow index features to build the models, the overall classification accuracy of the Random Forest 95.2% and 94.8 for C4.5. The overall classification accuracy increased by 5.1% for Random Forest and 6% for C4.5.

Finally, this study shows the application's dominance in terms of flow, packet, and byte composition. On the fixed data network, web applications consume approximately 35.5 percent of bytes, 21.5 percent of packets, and 56.6 percent of flow, making them the dominant application.

KEYWORDS: Random Forest, C4.5, Fixed network, Traffic classification, Application identification, Application dominance.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ADSL	Asymmetric Digital Subscriber Line
CDF	Cumulative Distribution
CSV	Common separated values
DB	Database
DNS	Domain Name System
DPI	Deep packet inspection technique
FN	False Negative
FP	False Positive
FTP	File transfer protocol
HTTP	Hypertext Transfer Protocol
IMAP	Internet Message Access Protocol
IP	Internet Protocol
ISO	International Organization for Standardization
ISP	Internet service provider
KNN	k-nearest neighbor's algorithm
LDAP	Lightweight Directory Access Protocol
ML	Machine Learning
NFS	Network File System
NMS	Network monitoring system
P2P	Peer-to-Peer
P3P	Platform for Privacy Preferences
POP3	Post Office Protocol version 3
PRTG	Paessler Router Traffic Grapher
QOS	Quality of service
SNMP	Simple Network Management Protocol
SQL	Structured Query Language
SSH	Secure Shell
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TN	True Negative

1. Introduction

Modern network environments are becoming more and more complex and diverse due to the emergence of a large number of new applications [1]. Because of the dynamic and diverse introduction of a huge number of new applications, determining which applications cause traffic congestion is the most difficult task. Complexity of network traffic composition and dynamics, which are difficult to manage and supervise, especially for ISP's (Internet Service Provider) network. In general, the characterization of network traffic provides insights for various network management activities, such as capacity planning and provisioning, traffic engineering, fault diagnosis, application performance, anomaly detection, and pricing [2]. In order to acquire a better knowledge of how people utilize their networks, it is necessary to study and identify many types of network applications that generate network traffic. There are different methods of identifying network applications. The traditional traffic classification methods include port-based and payload-based methods [3] [4]. Port-based identification of network traffic is quite simple and relies on mapping applications to well-known port number[5]. These methods are effective for application identification with The Internet Assigned Numbers Authority (IANA)registered port numbers. However, these approaches fail with the increase of applications that use dynamic port assignment. Payload-based methods, also known as deep packet inspection techniques (DPI), analyze the payload of the IP packets and look for each application or protocol characteristic signature[6]. This method is accurate and reliable in most classification cases [7]. However, the method fails to identify and classify applications in the network traffic due to payload content encryption and user privacy information protection.

A number of researchers are looking particularly closely at the application of machine learning (ML) techniques (a subset of the wider Artificial Intelligence discipline) to IP traffic classification in recent days [8]. ML methods have been using prioritized sets of features, which lead to different dynamical behaviors during training and classification. Since these features are both port-independent and payload-independent, the ML methods provide much more flexibility. ML techniques may be supervised or unsupervised. The supervised techniques need a complete labeled data set to classify unknown classes. The supervised learning techniques trains the model with some labeled data sets and then it will produce prediction output on new data samples [3]. While

in unsupervised, there is no need for a complete labeled dataset. The result of this technique does not identify instances in predefined classes

This thesis work shows fixed network traffic classification by capturing raw data from ethio telecom access layer devices in order to determine the dominating applications in the network and to select the best algorithm for classification.

1.1 Statement of Problem

The rapid growth of the internet in size, complexity, and traffic type makes it a challenging task for ISP to manage their network, to provide better quality of service, and to have good knowledge of applications. There is a huge amount of network traffic flow caused by different kinds of applications. However, there is no better understanding of the current and emerging applications that generate network traffic. Due to this, ISP face different kinds of problems, such as security problems, lack of knowledge about who is generating traffic, and a failure to know the nature and type of network traffic. This case is also the same for ethio telecom as an ISP company. In ethio telecom there is a growing network traffic flow, but different kinds of applications that generate the network traffic flow are not exactly known. However, knowledge of network traffic patterns is critical for gaining a better understanding of the applications that drive traffic flow, which assists ISP as input for quality of service, security, and market forecasting. To get this benefit, accurate classification and identification of applications is important.

There are different methods used to identify and classify applications running in network traffic, such as port-based or deep packet inspection techniques. However, this method fails to use because of the growing number of applications that use dynamic port assignment and encryption packet payload for security purposes, respectively. Because there is a large amount of data traffic and various types of applications running on the ISP network, and the drawbacks of both the port-based and DPI methods mentioned above, which make them difficult to use in the classification of applications running on the ISP network.

In this research, machine-learning techniques used to classify network traffic by recognizing statistical patterns in externally observable attributes of data packets. This method is both port-independent and payload-independent and efficient for dealing with ISPs network since it have larger amounts of raw traffic data.

1.2 Research Question

This thesis work used to answer the following main research questions:

- Which applications are dominant in ethio telecom network?
- Which are the best features for the classification?
- Which machine-learning algorithm provides a better classification accuracy?

1.3 Objective

1.3.1 General Objective

The main objective of this study is to identify applications running in ethio telecom fixed network traffic using a machine-learning algorithm.

1.3.2 Specific Objectives

- To select relevant packet data from fixed network traffic that used for this study.
- To select the appropriate machine learning algorithm for classification.
- To identify the best features that would be suitable to train the machine-learning model.
- To develop network classification models based on machine learning techniques.
- To analyze the performance of the classification models.
- Finally, the fixed network application trend analysis and dominance analysis carried out.

1.4 Literature Review

The author [2] uses two machine-learning algorithms for internet traffic classification. Support Vector Machine (SVM) is a supervised algorithm, whereas K-means clustering is an unsupervised algorithm [2]. The aim of the classification of network traffic is to map an unknown traffic sample to given categories. The author uses SVM to construct a hyperplane and group of hyperplanes for traffic classification in a high dimensional or infinite dimensional space. The kernel function applied in this algorithm to avoid high-dimensional operations. On the other hand, unsupervised, K-means classification allocates the samples into a fixed number of partitions based on a similarity measurement. An overall accuracy of both algorithms of over 95% achieved. Meanwhile, the system performance further improved with model tuning and feature selection. The author [6]

worked on the clustering flow label propagation technique, synthetic flow feature generation algorithm, and feature selection technique to classify the network traffic with a real dataset from a large-scale supercomputer. The clustering flow label propagation technique can used to extract correlation application flows, increasing the number of labeled flows. The author also uses a synthetic-flow feature generation algorithm to extract and select the most effective features in raw flows. The author also used supervised learning algorithms such as Random Forest, C4.5, and KNN to test the performance of traffic classification and application identification. The results show that the overall classification accuracy of this model is about 99%. The author [[9]uses a system model that shows step by step techniques for identifying unknown network traffic classes using machine learning techniques. These steps are network traffic capture, feature14 extraction selection, training process sampling, implementation of machine learning algorithm and result. Then WWW, DNS, FTP, P3P, and TELNET application traffic duration of 1 minute using the Wire Shark tool and extracting 23 features using the Netmate tool. After that, traffic classified using four machine-learning algorithms. Experimental results show that the C4.5 decision algorithm gives high accuracy results as compared to other Support Vector Machine, bayes Net and Naive Baes machine learning classifiers.

The author [5] proposes a method of classifying traffic flowing on the network according to the application generated by the end host. To perform these tasks, the raw data in this study consists of two packet level traces collected on an ADSL platform by a major ISP in France [5]. The author uses this data to perform different options of ADSL user profiling. Finally, based on network traffic classification using a machine-learning algorithm, the author did label a client with their dominating application [5]. Indeed, the dominating application in terms of bytes usually generates the vast majority of users' total volume [5]. Customers with the same dominating applications are clustered together [5].

The author [8]considered popular end-user application classification by using machine-learning algorithms and compared four kinds of machine learning algorithms such as j48, random forest, k-NN, and baye net. To increase the accuracy of the results for both datasets, and to reduce the computational complexity, the authors apply the Chi Squared AttributeEval feature selection method. Chi Squared Attribute Eva evaluator gave the most satisfying result with a 90% reduction of 111 features into 12 selected features for both datasets [6]. Applying the feature selection method increases the algorithm accuracy by 2%.

1.5 Methodology

- Review related literature to identify the appropriate network traffic classification technique and algorithm.
- Select the appropriate and well-known classification algorithms and data mining tool. Network traffic capture by open-source tools from ethio telecom access layer devices. Then, in order to improve the performance and minimize the time taken to build the models, the data-preprocessing task has been done.
- After the data preprocessing performed, the next major role was the selection of a relevant feature subset by removing redundant and irrelevant features. Which helps to improve the accuracy of a classification algorithm and reduce the time taken to build a model. Finally, make it ready for training and testing the models.
- When the data processing and feature selection are finished, train the selected classification ML algorithms by using two different kinds of network flow feature. The first one is the network flow feature that exists in the state of art. The second one is the network flow feature that exists in the state of art with two additional new features that are flow state and flow index. Then, the performance of these algorithms evaluated before and after the addition of the new feature set.
- Finally, in the fixed network, application trend and dominance analysis performed.

In this research work, Wireshark, tranalyzer2, MS-excel, and weka 3.9.5 tools used for data capture, extraction, visualization, and machine learning processes.

1.6 Scope and Limitation

1.6.1 Scope of the thesis

This thesis focuses on fixed network traffic classification based on data captured from ethio telecom access layer devices and classifies them into eight user applications. Then, in the fixed network, an analysis of the dominant applications performed.

1.6.2 Limitation of Thesis

This thesis limited itself to capturing the network traffic from three sites only and the data traffic of fixed network analysis and classified into end-user applications, not including the voice traffic.

1.7 Contribution

This research has contribution to ethio telecom mainly to fixed network internet service:

- Allow ethio telecom to perform optimization of network and improve network Performance.
- Study of application running in the network traffic helps ethio telecom in better Understanding of end user of the service.
- Network traffic classification also used as input for varies study such as security analysis, network management and market prediction.
- Knowledge of application helps ethio telecom as input for quality of experience.

1.8 Thesis Organization

This thesis organized into five chapters. Chapter one deals with the introduction, statement of the problem, objectives of study, and the methodology to achieve the objective of this study, scope and limitations of the thesis, literature review and contribution of the thesis.

The second chapter covers the fundamentals of network traffic analysis, including why it's necessary and how it's currently handled.

The third chapter introduces the traffic classification model and its processes, such as data collecting, data preparation, feature selection, algorithm selection and model training.

The results of the machine-learning algorithm described and reported in chapter four.

Finally, in Chapter five, there is a conclusion and suggestions for future research.

2. Network Traffic Analysis

Network traffic analysis is the process of capturing the network traffic and inspecting it closely to determine what is happening on the network [10].For ISPs, network analysis plays a critical role in responding to support of their diverse business objectives. This may help them to have

immediate access to data flowing through their networks and have detailed knowledge of the composition of traffic as well as the identification of application usage, which required by operators for better network design, provisioning, and quality of service (QOS) solutions [11]. First, in the network analysis, split traffic into distinct classes and then prioritize and treat them differently based on different criteria. Second, recognize the application to which classes belong. This traffic classification output helps to differentiate between the class of billing and the verification of service level agreements (SLA's). Moreover, network traffic classification used as a significant stage for developing successful congestion control schemes, and to differentiate out normal and malicious packets [12]. Network analyzed at different levels [13]. These levels are packet level, flow level, and network level for security management.

2.1 Flow Level Analysis

A network flow defined as the set of IP packets passing through an observation point in the network during a certain time interval, such that all packets belonging to a particular flow have a set of common properties [14]. These common properties called "standard five tuples" and are typically contained in the packet header [15]. These are the source IP, source port, destination IP, destination port, and port id. The flow definition does not place restrictions on the five standard tuples. Beyond these properties, they include the duration of the flow, volume of data, number of packets per flow, average packet size per flow, and sometimes mac address and vlan tags combined with the Ip address taken as flow[16]. In any case, flow has a network packet with the same properties.

Network analysis based on flow level aggregates information from different packets into a flow. The purpose of aggregating packets into flows is to convert raw packets communicating over the network into meaningful information elements about interaction. This interaction can provide a high-level understanding of the network behavior based on its statistical information without seeing the payload of the packet.

Different authors suggested that the performance of the traffic analysis depends on lots of factors, such as link utilization, pattern of packet arrival, number of flows, etc[11][17]. Among them, the number of flow counts influences the whole phases of the traffic analysis system. It is critical to have a thorough understanding of flow-based IP traffic characteristics in order to comprehend network traffic behavior and improve traffic performance.

2.2 Packet Level Analysis

Packet-level traffic analysis expresses traffic flows in terms of inter-packet time, length of the packet, packet size, mean and variance of the packet length, square of the root mean, etc. Packet level analysis is simpler and more general than apart from allowing us to analyze network traffic from a different perspective, they have the important advantage that their application in traffic simulation and generation lets us measure and study network parameters like delay, jitter, packet loss, packet corruption[18]

Level	Feature Set
Flow	➢ Packet length statistics within the flow: Mean and variance in forward and backward direction ➢ Time statistics within the flow: duration of flow and interarrival time the packet within the flow. ➢ Data volume statistics within the flow: Volume of data, number of packets per flow, bytes count ➢ Protocol based statistics within the flow: Initial Advertised Window bytes, advertised window byte, number of packets with PUSH bit set, Number of out of order packets, Time to live, Response flags, Number of canonical names, Query name rank
Packet level	➢ Length of the packet ➢ Mean of the packet length ➢ variance of the packet length, ➢ Square of root mean of packet length ➢ packet size ➢ Inter arrive time of the packet ➢ payload size

Table 2. 1 feature set for network analysis level

2.3 Network Traffic Classification Techniques

Network Traffic Classification is the process to identity the network applications or protocol that exists in a network. In order to classify unknown classes of applications in traffic, a number of network traffic classification techniques have been developed. The first technique called Port based techniques. This technique includes a port, which is firstly registered in Internet Assign Number Authority (1ANA)[19]. However, this technique failed due to an increase in applications, which use dynamic port number assignment. The other technique is the payload-based methods, which also called deep packet inspection techniques (DPI). It aims to analyze the payload of the IP packets and look for each application or protocol's characteristic signatures[20]. Due to payload content encryption and user privacy information protection, DPI techniques also failed to identify real-time network applications. Machine learning techniques overcome the drawbacks of port based and DPI by recognizing statistical patterns in externally observable attributes of the traffic. Their ultimate goal is either clustering IP traffic flows into groups that have similar traffic patterns, or classifying one or more applications of interest[21].

2.4 Machine Learning Algorithm

Nowadays, Machine Learning (ML) techniques are a very popular approach to identify and classify patterns in different fields of science [22] . The main objective of ML approach is to give the computer automatic learning capabilities, where the machines are able to extract knowledge from a process under certain conditions [22] . Generally speaking, the ML

Model should give the current state of the process given a number of incoming inputs. Machine learning algorithm mainly have four categories: supervised, unsupervised, semi supervised and reinforcement

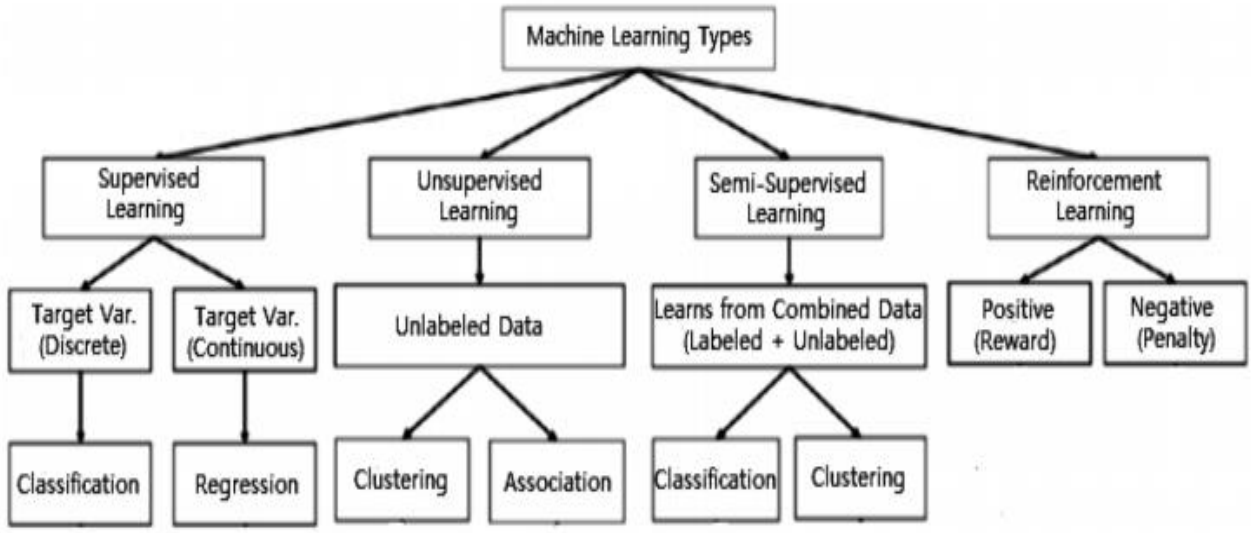

Figure 2. 1 Type of machine learning techniques[23]

2.4.1 Supervised Learning

Under supervised learning, a set of examples or training modules are provided with the correct outputs and on the basis of these training sets, the algorithm learns to respond more accurately by comparing its output with those that are given as input[24]. It means that the supervised learning technique trains the model with some labeled data set and then it will produce prediction output in new data sample. Example of supervised learning Algorithms are: Random Forest, C4.5 also known as decision tree, regression Analysis and support vector machine (SVM).

Random Forest

Random Forest is an ensemble classifier, which constructs a group of independent and non-identical decision trees based on the idea of randomization[25][26]. In this algorithm the Bagging principle is also used with another randomization technique called Random feature Selection [27]. In random forest, every decision tree made from randomly selected train dataset. Since the feature

selected randomly split in each decision tree, the correction between each decision tree reduced, which improves the classification accuracy. Ensemble and bagging techniques of random forest algorithm also helps to overcome the problem of over fitting. The Random Forest is appropriate for high dimensional data modeling because it can handle missing values and can handle continuous, categorical and binary data[28].

C4.5 Decision Tree

The C4.5 algorithm constructs a model based on a tree structure, in which each internal node represents a test on features, each branch representing an outcome of the test, and each leaf node representing a class label.

C4.5 algorithm consists of two process, preparation of decision tree and make the rules. Then, calculate the entropy and information gain with the highest attribute is selected [29] . Information gain ratio measures the correlation between two random variables[30]. In case of this work, it measures the correction between a feature and class label. The random variables X and Y, the gain ratio defined [29] .

$$Gain\ Ratio(X/Y) = \frac{H(X) - H(Y)}{H(X)}$$

Where: $H(X) = -\sum_{xi} p(xi) \log p(xi)$

$$H(X) = -\sum_{j} p(yi) \sum_{i} p\left(\frac{xi}{yi}\right) \log\left(\frac{xi}{yi}\right)$$

C4.5 has been widely used in traffic classification[5].

2.4.2 Unsupervised Learning

Unsupervised learning refers to the process of grouping data into clusters using automated methods or algorithms on data that has not been classified or categorized [31] . In this situation, algorithms must “learn” the underlying relationships or features from the available data and group cases with similar features or characteristics[32].The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc[23].

K-Means Clustering

The unsupervised learning method K-means cluster categorizes a dataset into a predefined number of clusters (assuming clusters), for example, to classify network flow. The basic concept is to pick centroids at random for each cluster. Each input represented as a coordinator by examining the features values, which made up of a group of points. Each point assigned to the closest centroid, and each cluster of points assigned to a centroid is measured. The centroid of each cluster later updated depending on the points assigned to the cluster. The procedure repeated by updating the stages until no changes made to the clusters, or until the centroids remain the same. Network flows are represented by points in a P-dimensional space (dimension refers to features such as packet size), with each packet having its own dimension; the size of packet p in the flow is represented by the coordinate on dimension p. The technique repeated, with the stages updated, until no clusters change, or the centroids remain the same.

2.4.3 Semi-Supervised Learning

Semi-supervised learning is a machine learning (ML) technique that combines supervised and unsupervised learning schemes. The main objective of Semi-supervised learning is to overcome the drawbacks of both supervised and unsupervised learning [33]. However, supervised learning requires a huge amount of training data to classify the test data, which is a cost-effective and time-consuming process [34].

2.4.4 Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency[35]. The RL approach based on interacting with the environment, as opposed to supervised learning, which is based on given sample data or instances. The problem to be solved in reinforcement learning (RL) is defined as a Markov Decision Process (MDP)[35].Which is all about making judgments consecutively. A typical RL problem has four components: Agent, Environment, Rewards, and Policy.

It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing and supply chain logistics, however, not preferable to use it for solving the basic or straightforward problems

3. Traffic Classification Model

This chapter focuses on the model development procedure that carried out in this study. There are several steps in the procedure. The first section of the discussion focused on capturing fixed network traffic from ethio telecom access layer devices. Then the traffic flow feature generated

and labeled with the end-user application type. After that, the data preprocessing process, which includes methods such as data cleaning and feature selection, takes place. The selection of a classification algorithm and model validation approaches discussed in depth. Figure 3 depicts a summary of the experimental workflow.

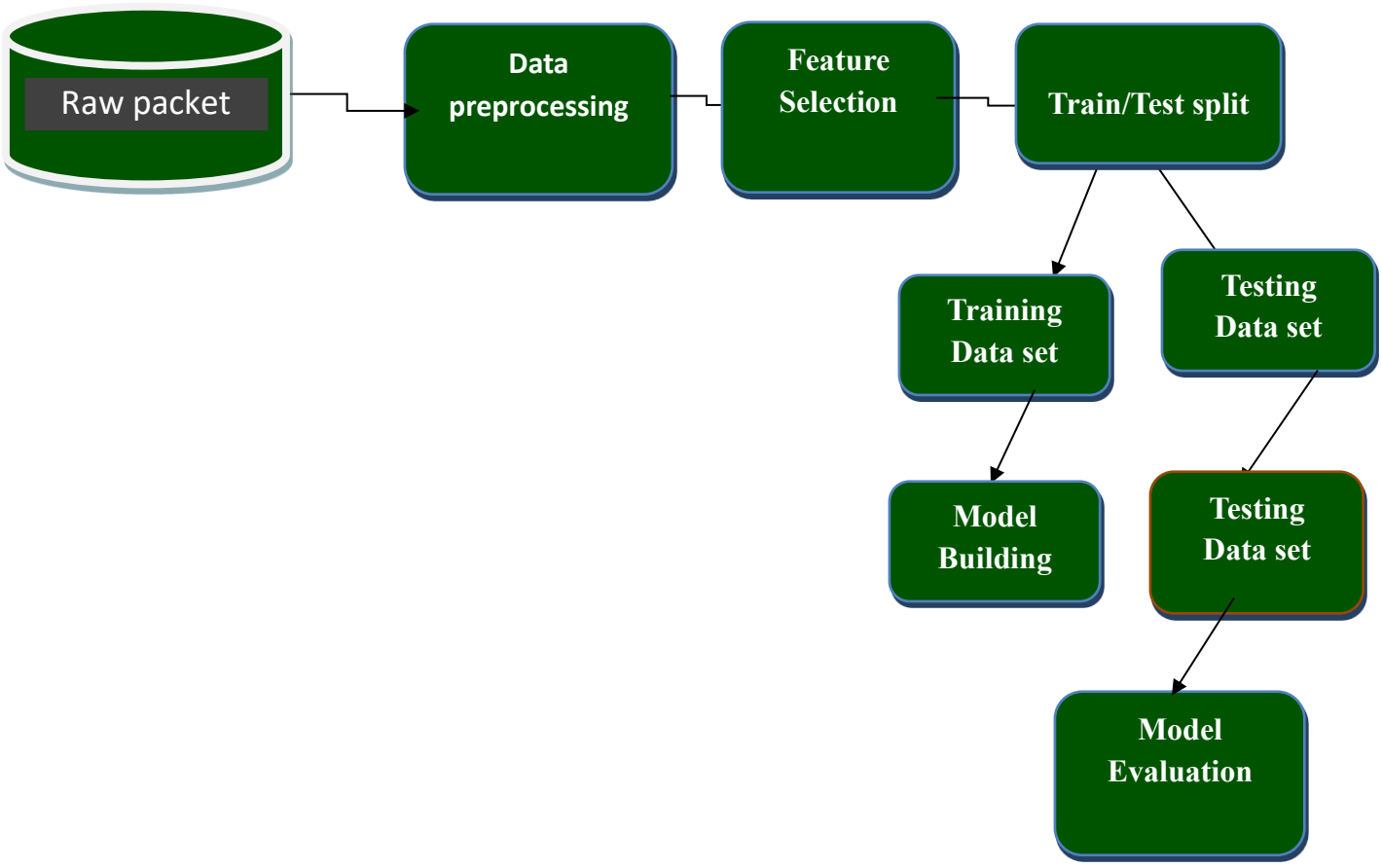

Figure 3. 1 Summary of experimental analysis

3.1 Network Traffic Capturing

The data for this thesis gathered from ethio telecom access layer device. Wireshark used to capture network traffic. Which is a free and open-source packet analyzer that may use for network troubleshooting, analysis, software and communication protocol creation, and education. The data set gathered from three locations: Bole, Leghar, and Sarbet, where the majority of enterprise customers of the fixed network are located. The total number of rows data packets is 54,

338, 92 in the pcapng format, which is compatible with extraction tool. Then, I used the tranalyzer2 tool in order to generate a flow feature from the captured packet. This tool is a lightweight flow generator and packet analyzer designed for simplicity, performance, and scalability [36] and depends on different enabled modules, denoted as plugins [36]. From these different enabled modules in this work, the nDPI plugin classifies flows according to their protocol/application by analyzing the payload content instead of the destination port. Using MS Excel to save the dataset for the Weka tool in a Comma Separated Values (CSV) file format.

3.2 Data Preprocessing

The collected packets frequently contain a large amount of raw data that is irrelevant to this investigation. As a result, cleaning the raw data before moving on to the next stage is critical. Understanding what the data is and what you want to achieve is critical before you begin data cleaning. Without that knowledge, you will not be able to decide what data is relevant while cleaning and preparing the required dataset. Since the goal of this effort is to classify network traffic into end user applications. Therefore, a packet containing network control information is regarded useless data, and a row with a missing attribute eliminated from the data collection.

Number of Raw Packets Collected		Number of flows	Number of Filtered flows
Raw packet	54,338,92	882916	25000

Table 3. 1 Total number of row data

3.3 Feature Selection

After capturing network traffic and generating flow features from raw packet data, the attribute reduced using the feature selection approach. Using the correlation plus ranker feature selection approach on Weka 3.9.5, the first 16 features derived from raw packet data reduced to nine.

Table 4.2 shows the attribute before the feature selection method is applied

NO	Attribute Name(In short hand)	Description
1	Duration	Duration of flow
2	Flow stat	Flow states

3	Flow ind	Flow index
4	Std_pkt size	Standard deviation layer 3 packet
5	Ave_pkt size	Average packet layer 3 size
6	Max _pkt size	Maximum layer 3 packet size
7	Min_pkt size	Minimum layer 3 packet size
8	Pyid_entropy	payload entropy
9	Pyid_chratio	Payload character ratio
10	Pyid_binratio	Payload binary ratio
11	Numbytesrcvd	Number received bytes
12	Numbytesnt	Number of sent bytes
13	Src ip	Source ip
14	Scrport	Source port
15	Dst Ip	Destination ip
16	Dst port	Destination port

Table 3. 2 Network feature used for classification

By using feature selection method, the above attribute reduced into nine.

The feature selection is the process of reducing attribute used to train a machine-learning mode.

This may help as:

- Using small features in the dataset will reduce train time for models.
- Improve accuracy of the algorithm
- Better resource usage: example memory usage.

As a feature selection strategy, the correlation plus ranker method applied in this study. The degree of linear linkage between two or more quantitative variables is measured by correlation [37]. Its value ranges from one to negative one. One means there is a positive correlation, negative one means there is a negative correlation, and zero means there is no correlation. Using the Weka correlation plus ranker feature selection method, feature subsets with a greater correlation with the class label and a lower correlation with other feature subsets chosen and rated higher. This method is useful for removing redundant and irrelevant attributes from a data source.

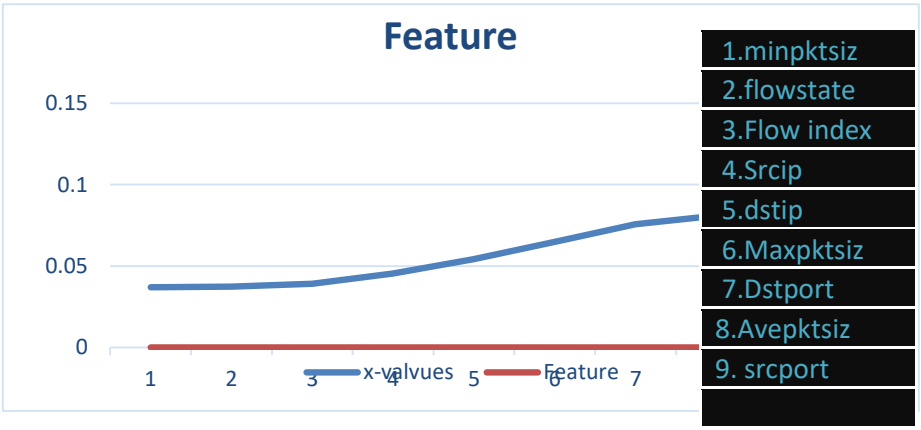

Figure 3. 2 Feature selected by correction plus ranker method

After selecting a feature, the network traffic labeled with eight end-user applications. The application kinds seen in traffic listed in the table below.

Class	**Application/protocol**
WEB	HTTP, and HTTP browsing
VoIP	Facebook, telegram, WeChat, Messenger

Streaming	MS Media Server, Real Player iTunes, Quick Time. YouTube QQ
P2P	BitTorrent, Filopodia
Mail	SMTP, POP3, IMAP, IMAPs POP3s, HTTP Mail
FTP	Ftp data, Ftp control, HTTP file transfer
DB	LDAP, Microsoft SQL, Oracle SQL, MySQL
Unknown	Unknown

Figure 3. 3 Feature selected by correction plus ranker method

After selecting a feature, the network traffic labeled with eight end-user applications. The application kinds seen in traffic listed in the table below.

3.4 Machine Learning Algorithm Selection

The dataset utilized in this work labeled manually, allowing supervised machine learning classification to be used. As shown in Table 4.8, two classifiers, Random Forest and C4.5, chosen. The selection criteria for those algorithms mostly based on the dataset's properties, as well as their application, computational complexity, and execution time, and their ability to deal with overfitting problems. Both algorithms are widely used in traffic classification [5] [9].

Algorithm	Advantage	Disadvantage	Applications

C4.5	➢ Perform well in a variety of multi-label categorization situations. ➢ The algorithm aims to minimize information entropy by finding a single attribute that best separates different classes from each other. ➢ Constructed with information gain as splitting criterion and reduced error. ➢ Easy classification and data interpretation.	➢ In appropriate for excessive data ➢ Impact of variance ➢ Sensitive to wards biasness	Traffic and text classification
Random Forest	➢ Good performance on the training data, poor generalization to other data. ➢ The bootstrapping and ensemble scheme makes Random Forest strong enough to overcome the problems of over fitting	➢ Complexity on data interpretation. ➢ It requires more computational resources	Application identification in live network

Table 3. 3 Selected supervised machine learning algorithm

3.5 Model Training

To train and validate a classification model using a machine learning method, the first step is to divide the data set into train and test sets. The 10-fold cross validation method employed as a validation approach in this study. The data set partitioned into ten data subsets of about similar size in the 10-fold cross validation approach, which done by randomly picking cases from the

learning set without replacement. The model then applied to the remainder of the subset. This thesis employed 25000 raw data for the experiment, which partitioned into 10 data subsets using a 10-fold cross validation approach, with each subset having a 2500 row data set. The model trained using a k-1 subset of the data, which represents the entire 22500-piece train dataset, and it tested using a 2500-piece subset of the data.

3.6 Evaluation Metrics

There are different types of metrics to test the accuracy of supervised machine learning algorithm. Some of them are confusion matric, accuracy, precision, recall and f-score.

Confusion Matrix

The confusion matrix is widely used in machine learning supervised classification or determination of the behavior of classification models[38]. The square structure of confusion matrix is represented through row and columns, where rows are the actual class of instances and columns are predicted classes [39]. The confusion matrix has some basic terminology, which applied to distinguish the predicted values and real values of the model in machine learning algorithm. This basic terminology is:

True positive (TP): The actual class positive and the predicted class positive.

False Negative (FN): The actual class is positive, but the predicted class negative.

True Negative (TN): The actual class is negative, and the predicted class to be negative.

False Positive (FP): The actual class is negative, but the predicted class positive.

Classification accuracy: measures the performance of the model how often the classifier is correct in detecting the classes of newly observed data[28]. Accuracy calculated as:

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

Precision: measures the performance model how precise the classifiers when predicting positive instance.

$$Precision = \frac{TP}{TP \ + \ FP}$$

Recall: measures the performance model in how the given labeled class of traffic correctly identify by classifier.

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

F-measure: is a combined metric (weighted harmonic mean) that evaluates the trade-off between precision and recall [2].

$$F_1 = \frac{2.precision\ recall}{precision+recall}$$

4. Result and Discussion

The goal of this chapter is to compare the performance of the two-algorithm based on a new feature in fixed network traffic classification at the access layer, as well as an analysis of the dominant

application based on the classification result. The classification model trained using a dataset containing nine features in this experiment. To reduce the cost of experiment time, 10-fold cross validation used, in which nine folds randomly extracted as training datasets and the remaining data used as validation datasets. Using the Random Forest and C4.5, the suggested classification model trained with two additional new features: flow index and flow state, which are not included in the state-of-the-art work, and it achieved a good classification accuracy for most traffic flows.

The experiment carried out first using features that are available in the state of the art work, then by adding two additional features that are not available in the literature, the experiment carried out again, and the results compared and analyzed in the part below.

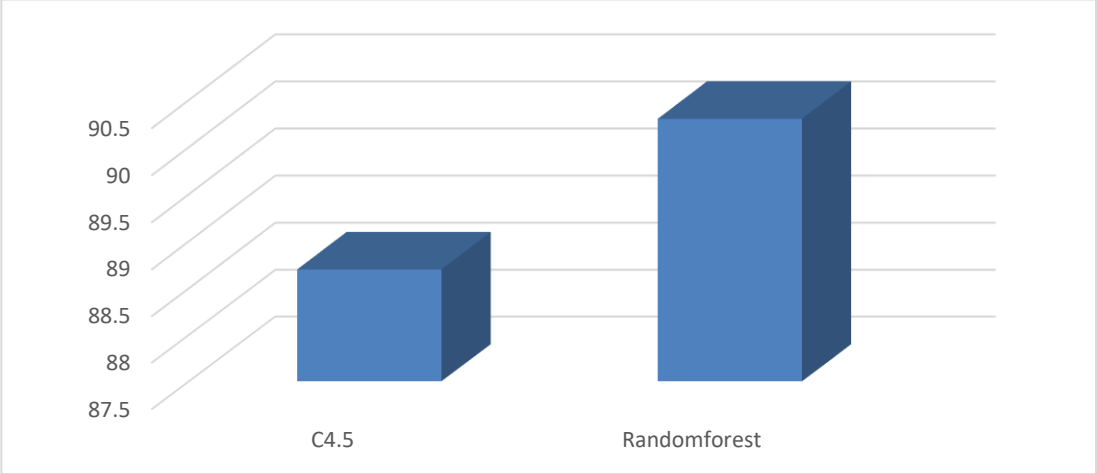

Figure 4. 1 The accuracy results of Random Forest, C4.5

Figure 4.1 shows the average accuracy results of all eight categories of applications using two machine-learning techniques, Random Forest and C4.5. The model achieves overall classification accuracy ranging from 88 percent to 90.8 percent, as shown in the figure. The Random Forest algorithm has the highest overall classification accuracy, which is 90.8 percent

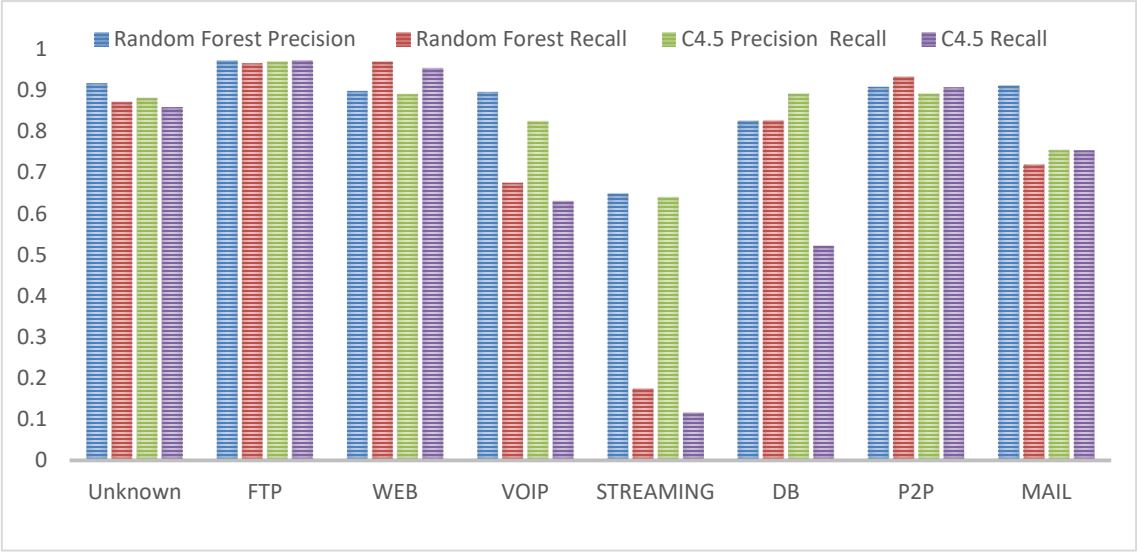

Figure 4. 2 The Recall and precision Results of Random Forest and C4.5

Figure 4.2 shows recall and precision results for Random Forest and C4.5. The Random Forest algorithms have higher precision in the mail, P2P, ftp, unknown, VoIP, DB, and web. Streaming's classification accuracy has dropped to 64.9 percent. Almost all applications expect streaming to be higher, yet streaming is relatively low. This is because the labeled dataset for streaming applications relatively small compared to the other application sample datasets.

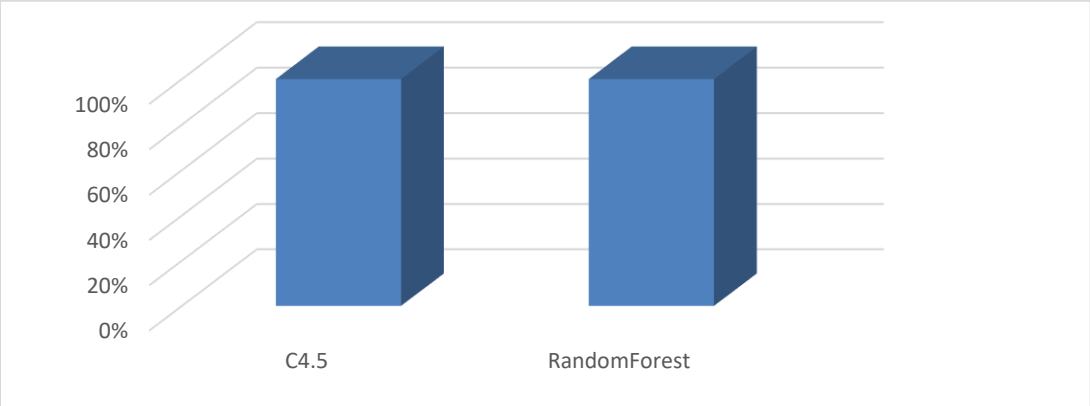

Figure 4. 3 Indicates the accuracy result with new feature set

Figure 4.3 depicts the classification of all eight-application types by the addition of two new features. Figure 4.1 depicts the classification of applications using seven features found in the literature. This feature includes srcip, srcport, dstip, dstport, avepktsiz, maxpktsize, and

minpktsize, but figure 4.2 illustrates classification with two additional features: flow index and flow state, which improve the accuracy of both algorithms. This is because the two feature sets have a higher correlation with the application class.

As illustrated in Fig. 4.3, the classification accuracy with new selected feature improves when compared to the previous classification accuracy. The average classification accuracy of the Random Forest and C4.5 with new selected features increased by 5.1 and 6.1 percent respectively.

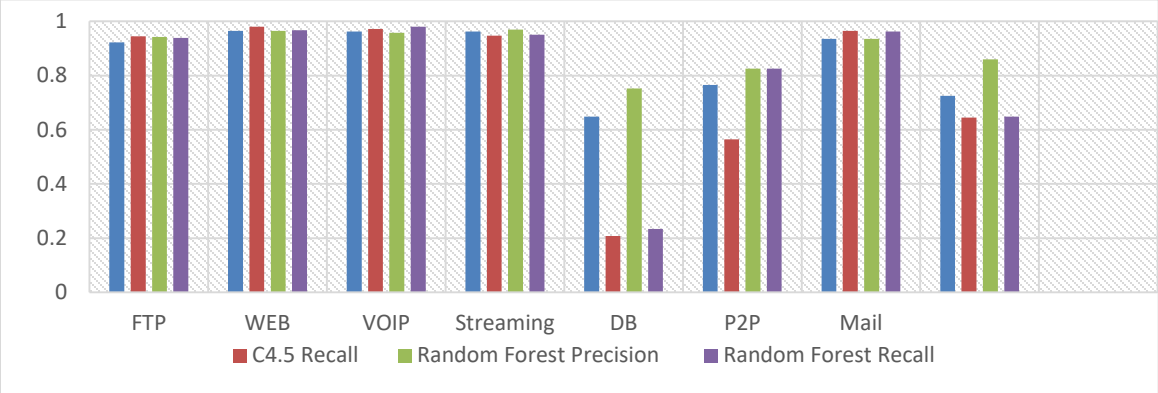

Figure 4. 4 The Classification Results with New Selected Features

Fig 4.4 shows that the model achieved a high average classification precision when compared to the previous model's accuracy of 95.2 percent for random forest and 94.8 percent for C4.5. The average recall for random forest and 95.1 percent of C4.5 are higher than the previous results, and the recall for almost all traffic classes is higher than the previous results.

At the application level, Web, p2p, ftp, unknown, and VoIP applications have higher accuracy (over 95%) when compared to previous models for both algorithms, while dB, streaming, and mail applications have accuracy above 50%. This model improved precision and recall by 1% and 5%, respectively, especially for streaming applications

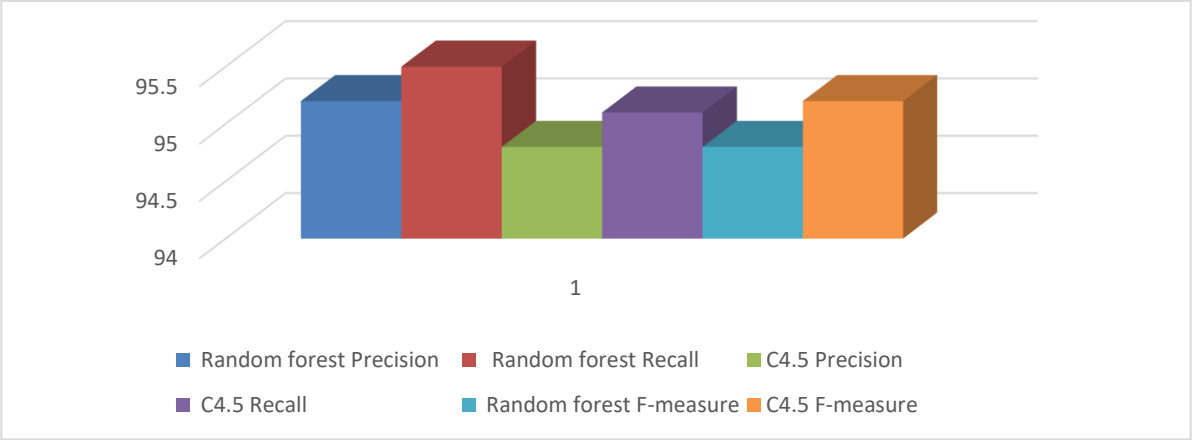

Figure 4. 5 Shows the average classification result by using Random Forest and C4.5

The average classification result using Random Forest and C4.5 shown in Figure 4.5. The Random Forest based model is more stable, with very high precision, recall, and f1-measure, though the classification accuracy of a few traffics is lower

4.1 Dominance Traffic Analysis

4.1.1 Application Composition in Traffic

Application profiling is a crucial stage in the development of a QOS solution. To properly aggregate applications into QOS classes, one must first understand the application's fundamental characteristics as well as the percentage of application composition within the traffic. This helps especially for ethio telecom to achieve the desired QOS targets for all applications.

This section describes the application composition in the fixed network traffic in term of bytes, packet and flow.

Figure 4. 6 indicate percentage value of byte/packet/flow in data set

The above graph represents the proportion of traffic composition by application. The internal ring indicates the percentage of bytes for eight applications; the middle ring shows the percentages of packet while outer ring shows the percentage of flow these eight applications in the dataset.

As figure 4.6 shows there is different composition of application in term of flow, packet and bytes.

- Web applications consume approximately 35.5 percent of bytes, 21.5 percent of packets, and 56.6 percent of flow.
- P2P accounts for 16.2 percent of bytes, 11.9 percent of packets, and 7.5 percent of total flow volume.
- Streaming accounts for 13.9 percent of bytes, 9.92 percent of packets, and 10 percent of flow.
- FTP accounts for 8.20 percent of bytes, 5.18 percent of packets, and 23.10 percent of traffic.
- VoIP accounts for 4.29 percent of total bytes, 3.95 percent of total packets, and 10.63 percent of total flow volume.
- Mail accounts for 0.15 percent of bytes, 0.1 percent of packets, and 0.3 percent of total flow volume.
- DB accounts for 0.1 percent of bytes, 0.4 percent of packets, and 0.34 percent of total flow volume.
- Unknown accounts for 14.3 percent of bytes, 12.50 percent of packets, and 18.63 percent of total flow volume.

4.2 CDF of Average Packet Size Distribution

This section describes the cumulative average packet size distribution of nine applications in the flow.

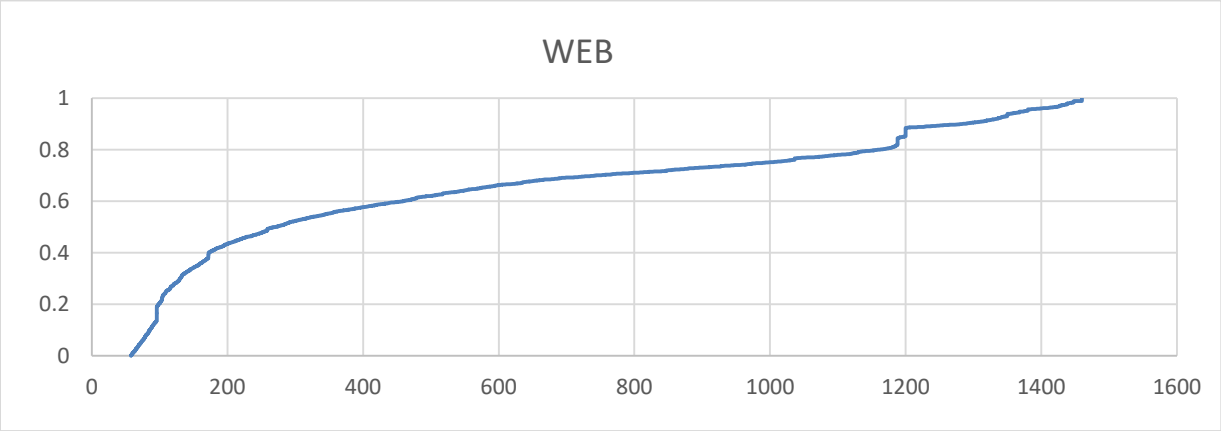

Figure 4. 7 Cumulative distribution function for web application

According to the graph above, approximately 65 percent of the flow has an average packet size for web applications of less than 600 bytes, 24.85 percent has an average packet size greater than 1000 bytes, and the overall average packet size for web applications is 513 bytes

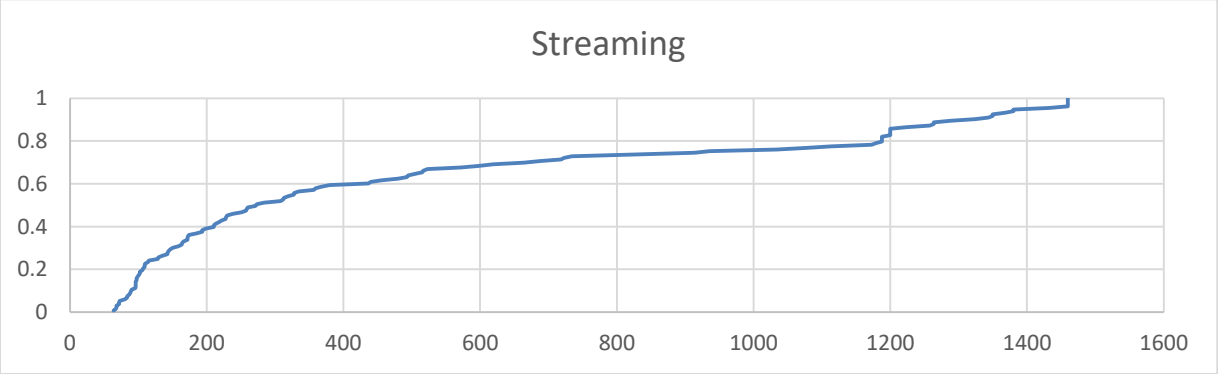

Figure 4. 8 Cumulative distribution function streaming application

Figure 4.8 shows the average packet size for streaming is around 642 bytes, with 59 percent of the flow having a packet size of less than 400 bytes, 23.5 percent having a packet size of more than 1000 bytes, and the overall average packet size for streaming being around 642 bytes.

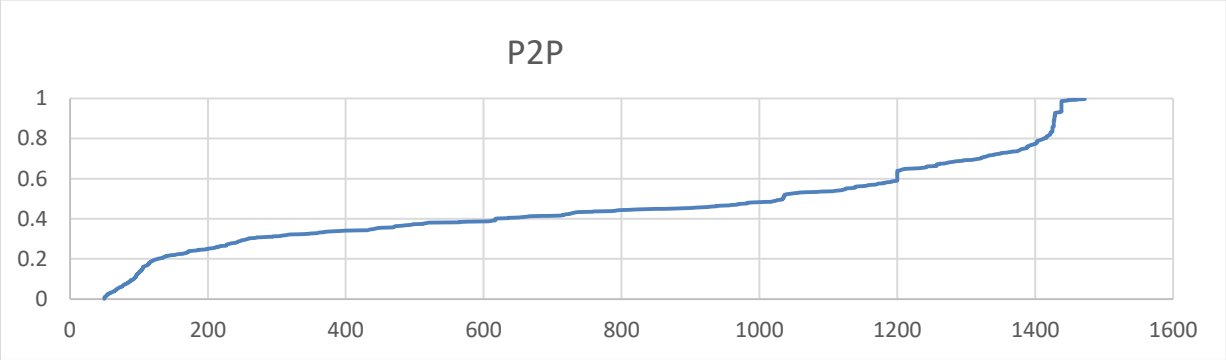

Figure 4. 9 Cumulative distribution function p2p application

Figure 4.0. Approximately 48.4% of the flows have an average packet size of p2p less than 1000 bytes, and 51.6% of the flows have an average packet size greater than 1000 bytes, with the overall average packet size for p2p being around 820.2 bytes

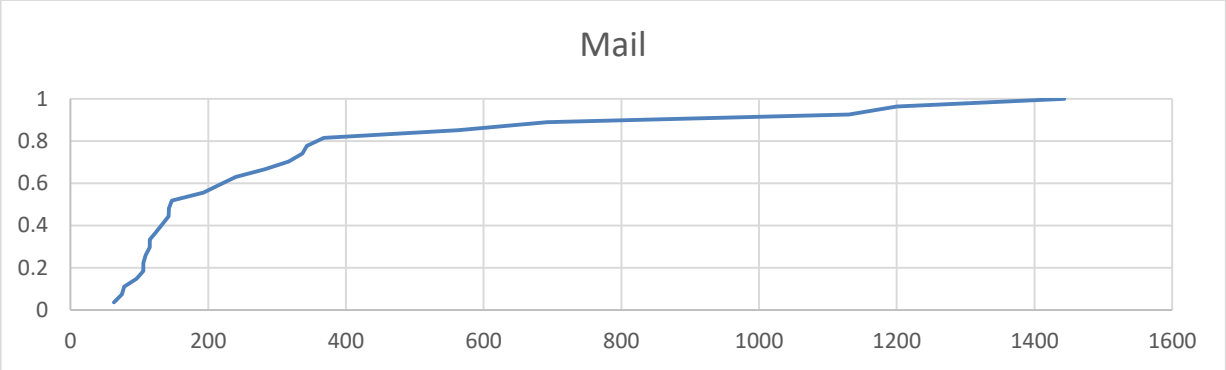

Figure 4. 10 Cumulative distribution function for mail application

Figure 4.10 shows that approximately 78.5 percent of the flow has an average packet size for mail application of less than 400 bytes, while approximately 10.7% of the flow has an average packet size of more than 1000 bytes, for an overall average packet size of 328.86 bytes.

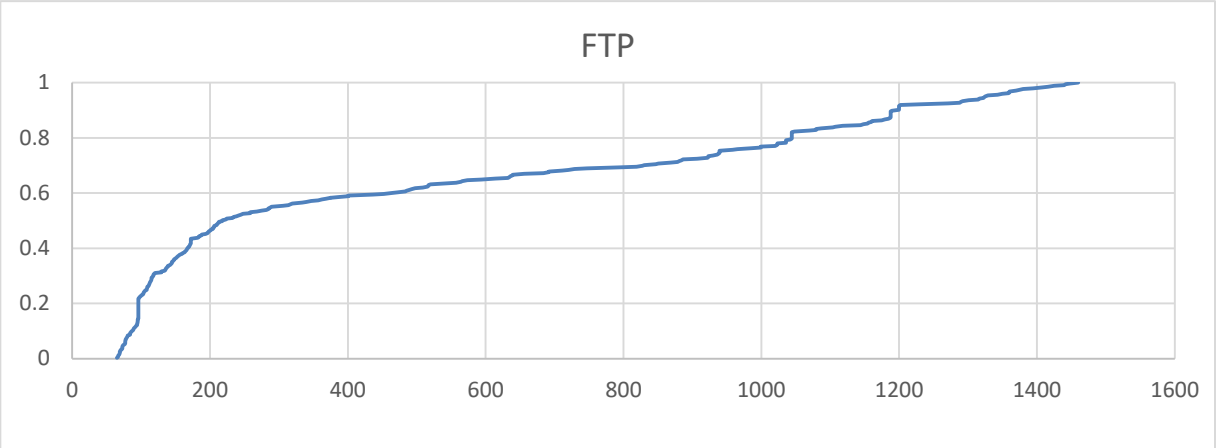

Figure 4. 11 Cumulative distribution function FTP application

Figure 4.11. The average packet size for FTP is 492.6 bytes, with approximately 60% of the flow having an average packet size of less than 400 bytes and approximately 23.1 percent having an average packet size of more than 1000 bytes.

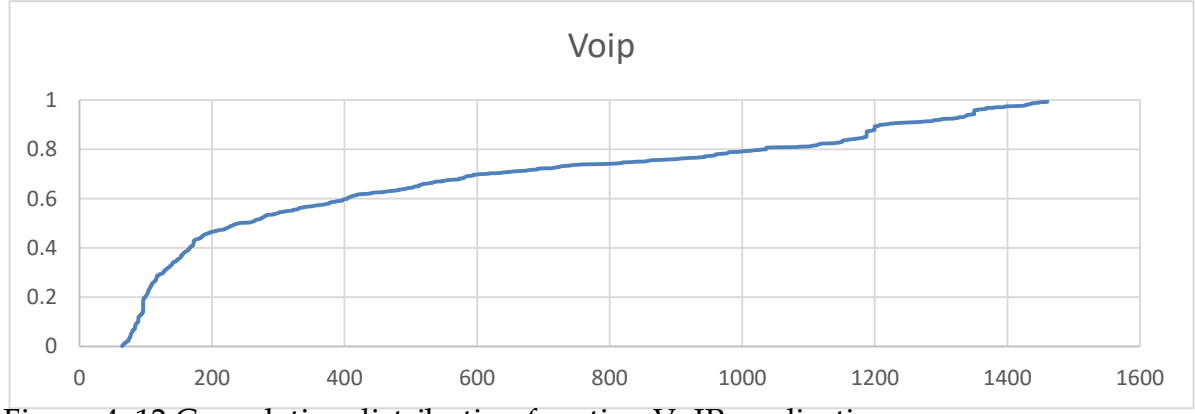

Figure 4. 12 Cumulative distribution function VoIP application

Figure 4.12. Around 60% of the flow has an average packet size for VoIP of less than 400 bytes, and around 20% of the flow has an average packet size of more than 1000 bytes, with an overall average packet size for VoIP of 479 byte

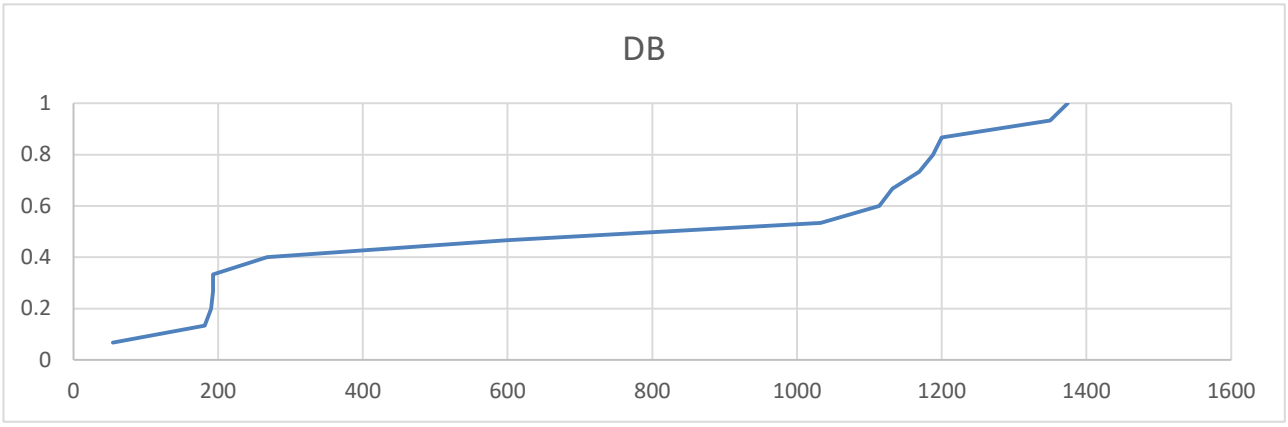

Figure 4. 13 Cumulative distribution function for DB application

Figure 4.13 shows around 43.75 percent of the flow has an average packet size of less than 400 bytes, and about half of the flow has an average packet size of more than 1000 bytes, with an overall average packet size of 749.2 bytes for DB.

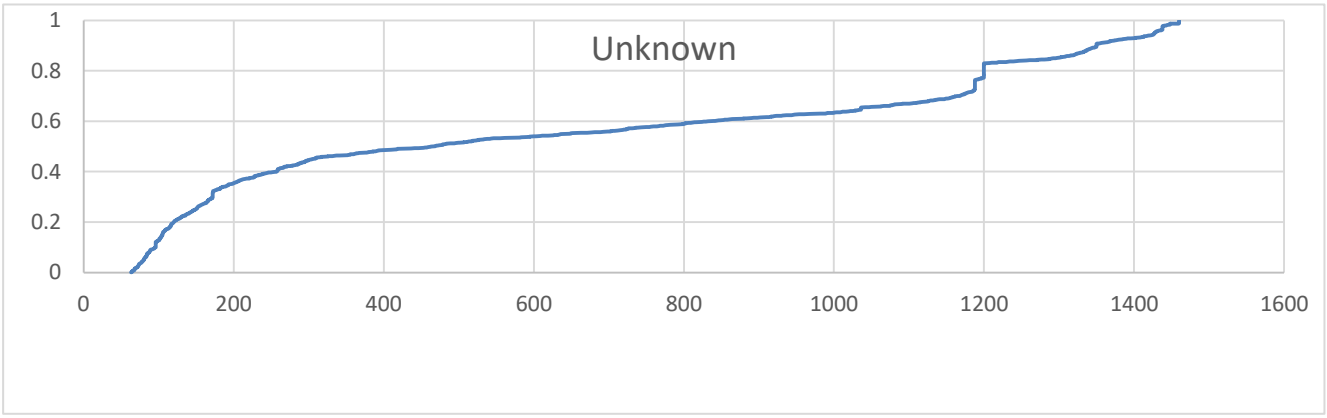

Figure 4. 14 Cumulative distribution function for Unknown application

Figure 4.14.Shows around 50% of the flow has an average packet size of less than 478 bytes, 36.7 percent has an average packet size of more than 1000 bytes, and the overall average packet size for unknown applications is around 641.5 percent bytes.

5. Conclusion and Future work

5.1 Conclusion

In this thesis, fixed network traffic collected from ethio telecom access layer devices to study mainly which applications dominate in the fixed network traffic.

The size of the data is limited to 25000 flows, which captured from access layer devices on three different sites. In order to collect the data, the main challenge is the size of the data, which is very high above the capacity of the captured device and privacy issue.

In general, this thesis shows the classification of applications running on fixed networks into eight application categories by using machine learning techniques.

- The overall accuracy of the model using C4.5 is 95.1 percent, and the random forest is 95.6 percent.
- The accuracy of random forest is slightly higher than C4.5
- Recall and precision of web, p2p, ftp, and unknown classes are higher than 90%.
- The precision of streaming, DB, and mail is almost higher than 50% but lower recall. This means a larger fraction of flow in this class classified in other classes.
- The statistical distribution of applications in the traffic shows web applications make up about 35.5 percent of bytes, 21 percent of packets, and 56.6 percent of the total flow of applications in the traffic. This makes the web application become the dominating application.
- P2P accounts for 16.2% of bytes, 11.9% of packets, and 7.55% of total flow. P2P, on the other hand, has a large overall average packet size when compared to other applications. This shows that it has a smaller percentage of flow with larger packets that make up most of the bytes consumed by the application. However, DB, streaming, and unknown next to p2p have a large overall average packet size of 742.9%, 642%, and 641.5%, respectively.

It has a smaller percentage of flow, with larger packets that make up most of the bytes consumed by the application.

- VoIP and FTP have relatively longer flow lengths, accounting for 10.63% and 23.10% of total flow length, respectively, but smaller overall average packet size. This indicates that with longer flow, smaller packets of bytes consumed by these two applications. On the other hand, mail has a smaller flow with a smaller overall average packet size. This indicates that fewer bytes consumed by this application.

5.2 Future work

It is impossible to collect large enough packets due to a combination of time constraints, the sensitivity of the data that the problem deals with, and the fact that the study conducted on a real fixed network. However, the thesis presented a method for analyzing application dominance in fixed networks and classifying network traffic using machine-learning algorithms.

Based on knowledge gathered from this research, the following future works recommended for further investigation.

- Using additional new futures not found in the state of the art based on similar traffic types, analyze the close relationship between network traffic and application characteristics, and implement an application classification system based on machine learning to investigate the close relationship between network traffic and application characteristics.
- Future research, whose focus is on QOS design mechanisms for fixed network traffic management, should take the findings of this study as an input
- Application profiling is a crucial stage in the development of a QOS solution. To efficiently aggregate applications into QOS classes, one must first understand the fundamental properties of the traffic. This study helps as starting point for further research on the application's characteristics.

References

[1] T. T. T. Nguyen and G. Armitage, “A Survey of Techniques for Internet Traffic Classification using ML,” Ieee Comst, vol. 10, no. 4, pp. 56–76, 2008.

[2] Z. Fan and R. Liu, “Investigation of machine learning based network traffic classification,” Proc. Int. Symp. Wirel. Commun. Syst., vol. 2017-Augus, pp. 1–6, 2017, doi: 10.1109/ISWCS.2017.8108090.

[3] B. Yamansavascilar, M. A. Guvensan, A. G. Yavuz, and M. E. Karsligil, “Application identification via network traffic classification,” 2017 Int. Conf. Comput. Netw. Commun. ICNC 2017, pp. 843–848, 2017, doi: 10.1109/ICCNC.2017.7876241.

[4] B. Analysis,“Network Traffic Analysis Using tcpdump,” vol. 118, no. 24, pp. 1–59, 2003.

[5] M. Pietrzyk, “Methods and algorithms for network traffic classification,” no. April, 2011, [Online].Available:http://www.eurecom.fr/publication/3366%5Cnhttp://www.eurecom.fr/en/publication/3366/detail/methods-and-algorithms-for-network-traffic-classification.

[6] S. Zhao, K. Ye, and C. Z. Xu, “Traffic classification and application identification based on machine learning in large-scale supercomputing center,” Proc. - 21st IEEE Int. Conf. High Perform. Comput. Commun. 17th IEEE Int. Conf. Smart City 5th IEEE Int. Conf. DataSci.Syst.HPCC/SmartCity/DSS2019,pp.2299–2304,2019,doi: 10.1109/HPCC/SmartCity/DSS.2019.00319.

[7] G. Alotibi, N. Clarke, F. Li, and S. Furnell, “Identifying Users by Network Traffic Metadata,” Int. J. Chaotic Comput., vol. 4, no. 2, pp. 103–112, 2016, doi: 10.20533/ijcc.2046.3359.2016.0013.

[8] F. Pacheco et al., “An autonomic traffic analysis proposal using Machine Learning techniques To cite this version : HAL Id : hal-02423382 An autonomic traffic analysis proposal using Machine Learning techniques,” 2020.

[9] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, “Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms,” 2016 2nd IEEE Int. Conf. Comput. Commun. ICCC 2016 - Proc., no. October, pp. 2451–2455, 2017, doi: 10.1109/CompComm.2016.7925139.

[10] C. Alcantara, V. R. Dasari, C. Bumgardner, and M. P. McGarry, “Evaluating features for network application classification,” p. 30, 2020, doi: 10.1117/12.2558687.

[11] M. Joshi and T. H. Hadi, “A Review of Network Traffic Analysis and Prediction Techniques,” 2015, [Online]. Available: http://arxiv.org/abs/1507.05722.

[12] M. Kim, Y. J. Won, H. Lee, J. W. Hong, and R. Boutaba, “Flow-based Characteristic Analysis of Internet Application Traffic Flow-based Characteristic Analysis of Internet Application Traffic,” no. January 2004, 2014.

[13] A. H. Theyazn, “A review of network traffic analysis and prediction techniques,” no. January, 2021.

[14] A. (Karunya U. Jamuna and V. (Karunya U. Ewards S.E, “Efficient Flow based Network

Traffic Classification using Machine Learning," Int. J. Eng. Res. Appl., vol. 3, no. 2, pp. 13241328,2013,[Online].Available:https://pdfs.semanticscholar.org/d968/3577f48dd7d6d7b73565e567f778796d282c.pdf.

[15] J. Bakker, B. Ng, W. K. G. Seah, and A. Pekar, "Traffic classification with machine learning in a live network," 2019 IFIP/IEEE Symp. Integr. Netw. Serv. Manag. IM 2019, pp. 488–493, 2019.

[16] T. Furlong, "Tools, Data, and Flow Attributes for Understanding Network Traffic without Payload," Network, 2007.

[17] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, "Reviewing traffic classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7754, pp. 123–147, 2013, doi: 10.1007/978-3-642-36784-7_6.

[18] A. Dainotti, A. Pescapé, and G. Ventre, "A packet-level characterization of network traffic," 2006 11th Int. Work. Comput. Model. Anal. Des. Commun. Links Networks, vol. 2006, pp. 38–45, 2006, doi: 10.1109/CAMAD.2006.1649716.

[19] Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," AIP Conf. Proc., vol. 1820, no. March 2017, 2017, doi: 10.1063/1.4977376.

[20] J. Kelner, A. Callado, C. K. Member, G. Szabó, and B. Péter, "A Survey on Internet Traf fi c Identi fi cation," vol. 11, no. 3, pp. 37–52, 2009.

[21] D. J. Parish, K. Bharadia, A. Larkum, I. W. Phillips, and M. A. Oliver, "Using packet size distributions to identify real-time n etvvo r ked a pp I icat ions," Computer (Long. Beach. Calif)., vol. 152, no. 6, pp. 0–4, 2005, doi: 10.1049/ip-com.

[22] T. Seyed Tabatabaei, M. Adel, F. Karray, and M. Kamel, "Machine learning-based classification of encrypted internet traffic," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7376 LNAI, no. July, pp. 578–592, 2012, doi: 10.1007/978-3-642-31537-4_45.

[23] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Comput. Sci., vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[24] S. Jain and A. Saha, "Rank-based univariate feature selection methods on machine learning classifiers for code smell detection," Evol. Intell., no. 4, 2021, doi: 10.1007/s12065-020-00536-z.

[25] S. Bernard, L. Heutte, and S. Adam, "On the selection of decision trees in Random forests," Proc. Int. Jt. Conf. Neural Networks, pp. 302–307, 2009, doi: 10.1109/IJCNN.2009.5178693.

[26] B. Rajoub, “Supervised and unsupervised learning,” Biomed. Signal Process. Artif. Intell. Healthc., no. January, pp. 51–89, 2020, doi: 10.1016/b978-0-12-818946-7.00003-2.

[27] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” J. Phys. Conf. Ser., vol. 1142, no. 1, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[28] “M ASTER T HESIS Performance Evaluation of Supervised Machine Learning Algorithms to Detect IP Spoofing At- tack : The Case of Ethio telecom LTE Network,” 2020.

[29] E. Budiman, Haviluddin, N. Dengan, A. H. Kridalaksana, M. Wati, and Purnawansyah, “Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation,” Lect. Notes Electr. Eng., vol. 488, no. February, pp. 380–389, 2018, doi: 10.1007/978-981-10-8276-4_36.

[30] D. Berrar, “Cross-validation,” Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., vol. 1–3, no. April, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[31] Y. C A Padmanabha Reddy, P. Viswanath, and B. Eswara Reddy, “Semi-supervised learning: a brief review,” Int. J. Eng. Technol., vol. 7, no. 1.8, p. 81, 2018, doi: 10.14419/ijet.v7i1.8.9977.

[32] K. Szczypiorski, A. Janicki, and S. Wendzel, “‘The good, the bad and the ugly’: Evaluation of Wi-Fi steganography,” J. Commun., vol. 10, no. 10, pp. 747–752, 2015, doi: 10.12720/jcm.v.n.p-p.

[33] N. Duffield, J. Erman, P. Haffner, and S. Sen, “A Modular Machine Learning System for Flow-Level Traffic,” vol. 6, no. 1, 2012, doi: 10.1145/2133360.2133364.

[34] W. Li and A. W. Moore, “A Machine Learning Approach for Efficient Traffic Classification.”

[35] M. Breternitz, “REINFORCEMENT LEARNING : A LITERATURE REVIEW (September 2020) REINFORCEMENT LEARNING : A LITERATURE REVIEW (September 2020),” no. December, 2020, doi: 10.13140/RG.2.2.30323.76327.

[36] T. D. Team, “Version 0.7.1.”

[37] P. Ducange and G. Mannar, “A Novel Approach for Internet Traffic Classification based on Multi-Objective Evolutionary Fuzzy Classifiers,” no. July, 2017, doi: 10.1109/FUZZ-IEEE.2017.8015662.

[38] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking,” pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[39] O. Caelen, “A Bayesian Interpretation of the Confusion Matrix.”