Statistical Modeling of Internet Traffic Flow Length and Flow Size

By: Shemelis Tadesse
Advisor: Dr. -Ing Dereje Hailemariam

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 Abab, Ethiopia

December 20,2019

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.

Shemelis Tadesse

_____

Name

_____

Signature

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Signed by the Examining Committee:

Examiner _____ Signature _____ Date _____

Examiner _____ Signature _____ Date _____

Advisor _____ Signature _____ Date _____

_____

Dean, School of Electrical and Computer Engineering

Acknowledgment

I would like to express my sincere gratitude to my advisor Dr.–Ing. Dereje Hailemariam for the uninterrupted support of my research, for his patience, motivation, passion, and enormous knowledge. His 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. Last but not least, 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

The growth of internet traffic forces telecom service providers to invest in new infrastructure and/or expanding the network to achieve the desired Quality of Service (QoS) and cope up the network congestion. But several techniques available for granting QoS and network congestion. The first step is to understand the network performance. Network performance needs accurate traffic modeling that has the potential to improve desired QoS, allocating the network resources (i.e. bandwidth). This thesis presents a statistical model of the internet traffic applications by their random nature of flow-length and flow-size. Wireshark network monitoring and analysis tool is used to collect internet traffic data from the ethio telecom core switch and generating experimental internet traffic data in controlled environment. The experimental data are used to train and test the machine learning model that helps to identify the internet applications. Firstly, identifying internet traffic applications using machine learning classification techniques. Secondly, statistical methods are used to fit the Cumulative Distribution Function (CDF) and select the parameters that best fitted in both flow-length and flow-size for identified applications. Finally, deliver statistical model to each applications and corresponding parameters. Recently internet traffic modeling is applicable in capacity planning for traffic engineering, anomaly detection and performance analysis are some of them. Based on the result found the Log-normal distribution is best fitted to flow-length and flow size for three applications and Weibull distribution is for SSH application in both flow length and flow size.

KEYWORDS: Internet Traffic Modeling, Identification, Flow Length, and Flow Size, Distribution

Contents

List of Figures

List of Tables

Acronyms

ACK Acknowledgment

BGP Border Gateway Protocol

CAGR Compound Annual Growth Rate

CDF Cumulative Distribution Function

DNS Domain Name Server

DPI Deep Packet Inspection

FTP File Transfer Protocol

HTTP HyperText Transfer Protocol

HTTPS Hypertext Transfer Protocol Secure

IANA Internet Assigned Numbers Authority

IP Internet Protocol

KNN K-Nearest Neighbor

LSE Least-Square Estimation

ML Machine Learning

MLE Maximum Likelihood Estimation

OSPF Open Shortest Path First

OSI Open System Interconnection

PDF Probability Distribution Function

QoS Quality of Service

RIP Routing Information Protocol

ROC Receiver Operating Characteristic

SLA Service Level Agreement

SMTP Simple Mail Transfer Protocol

SSH Secure Socket Shell

SVM Support Vector Machine

SYN Sync

TCP Transmission Control Protocol

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunications System

WWW World Wide Web

3G Third Generation

4G Forth Generation

1 Introduction

1.1 Background

The Internet is a network of networks that is used to transfer enormous amount of data and services, such as streaming and web browsing, around the world. In the globalizing world of the 21st century, the internet is playing a vital role in increasing peoples inter-connectivity and promoting flow of data around the world at present.The number of users is expected to show significant increment in the coming years as a result of global population growth as well as the increasing popularity of internet experience of users. Such an increase in the number of internet subscribers will ultimately have incremental impact on the amount of data transfer, which will affect internet traffic as the existing limited network will be congested by the increasing number of service users. This will, in turn, have impact on network efficiency and effectiveness as well as service quality and customer experience.

According to some sources, the global data flow has shown rapid increase in the past decade. For example, The total amount of data traffic in year 2016 was 6.7(EB) and increased to 11.5(EB) in year 2017. According to Cisco prediction from 2017 - 2022 the data traffic will increase 396(EB) per month globally [1]. This is, in fact, a huge increase that initiates the need for telecom service providers to work harder so as to cope up with the growing need for data services. Being one of the major challenges for many telecom service providers,

addressing the issue by properly allocating existing network resources and/or expanding their networks to achieve the desired QoS, along with increasing their revenue, is mandatory. However, expanding the existing network needs an extra investment of money and time, making it resource consuming.

Service providers need to have a better understanding of the traffic that passes through their networks as well as how their network is being utilized. Internet traffic modeling gives a better understanding of the network behavior and also provides inputs to network optimization and resource allocation process. The output of the model gives us an understanding of the efficient management of network resources to accommodate the increasing demand.

There are many approaches to model internet traffic[2]. Internet traffic modeling can be done in two broad levels; namely, packet-level and flow level. Packet-level models are usually used to estimate quantities such as queue sizes at buffers and throughput. However, packet-level models are fixed so that they do not capture flow-level dynamics such as flow duration or the number of active flows. On the other hand, in[3] flow-level models, such as those in [5], [6], flows are treated as finite-sized units of data, whose sizes are random and specified by a probability distribution. Flows arrive at the network according to an arrival process with a given distribution. The models aim to capture flow-level dynamics such as flow transfer duration and the number of active flows. This makes flow level models more preferred to packet level models in internet traffic modeling

1.2 Problem Statement

The rapid increase in the world's population and the increasing popularity of the internet has become one of the challenges of telecom service providers worldwide. Such a case is becoming a bottleneck for the service providers as they have to work harder to meet the increasing demand for their services. Internet traffic has been growing exponentially for the past decades and its growth is expected to continue globally. The growth has been observed in several aspects such as, traffic volume, number of subscribers connected to the internet and an average number of devices and connections. Pertinent to the above mentioned causes, global Internet Protocol (IP) traffic is expected to grow at a Compound Annual Growth Rate (CAGR) of 26 percent per year from 2017 to 2022[1].

The case is also the same in Ethiopia. The country has been experiencing rapid population growth in the past decades and the growth is expected to continue in the future. This situation is exerting pressure on ethio telecom as the increase in population will ultimately lead towards an increased number of mobile subscribers in the country. Such an increase in the number of subscribers has led to a growth in internet traffic.The total number of mobile

subscribers in ethio telecom is 41.9 million as of 2019 and most of these subscribers are expected to use the company's internet service[4]. ethio telecom is being challenged by a data traffic growth.For instance in Figure 1.1 shows the data growth for Huawie 3G network in Addis Ababa for more than two months. This data growth may lead to network congestion.

To properly manage the incurred network congestion, ethio telecom's mitigation measures are mainly focused on network expansion and/or deployment of new projects. Despite the importance of such measures in improving service quality and customers experience, they incur high cost, both in terms of finance and execution time, on the company, causing pressure on the country's limited resources. The company also performs some optimization activities to improve its traffic. However, it essential to have a proper optimization technique and resource allocation and know the parameter characteristics. A better ways to study parameters can, however, be attained by using the results acquired from modeling internet traffic variation.

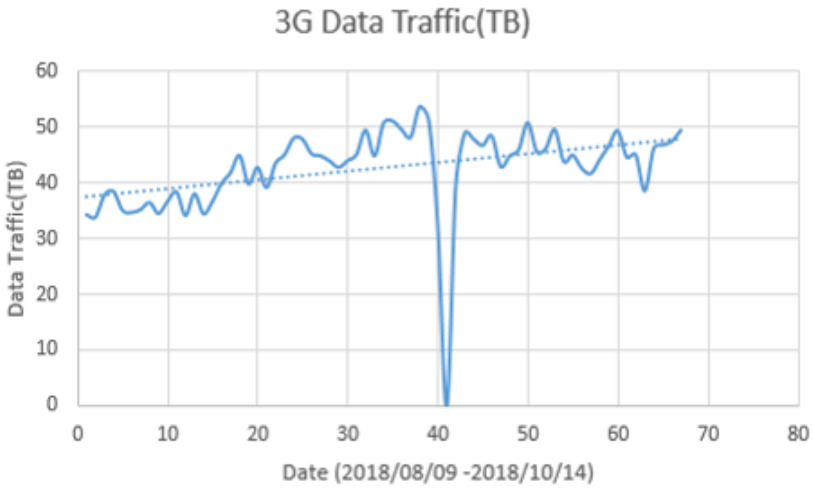

Figure 1.1: 3G Data Traffic Huawei Site in Addis Ababa

1.3 Objective

1.3.1 General objective

The main objective of this thesis is to study and model the flow length and flow size of different applications that is DNS, HTTP, SSH and HTTPS. For that, statistical modeling is used.

1.3.2 Specific Objectives

To achieve the general objective, this research accomplishes the below specific objectives.

- To analyze the internet traffic data and extract flow length and flow size data.
- To study the applications of statistical modeling techniques.
- Developing a statistical model that determine the flow length and flow size of applications stated in general objective
- Based on the result obtained from the model forward recommendations to the service provider to have an efficient network optimization.

1.4 Methodology

The methodology used in this research is show in Figure 1.2.

- Reviewing related literature papers extensively to identify and select for different IP traffics. Search for different machine learning algorithms and internet traffic modeling.
- Prepare the data set, generating and capturing for identification and modeling the internet traffic using Wireshark [5], which is a well-known packet capture and traffic analysis tool is used to capture internet traffic data in a laboratory environment and real-time traffic at 3G network
- Identify various internet traffic applications into HyperText Transfer Protocol (HTTP), Hypertext Transfer Protocol Secure (HTTPS), Secure Socket Shell (SSH), and Domain Name Server (DNS) with decision tree machine learning logarithms using MATLAB tool. Classification of applications accurately is fundamental in several network activities [6].
- In this research work MATLAB and Ms-excel tool is used for visualization and Machine learning.

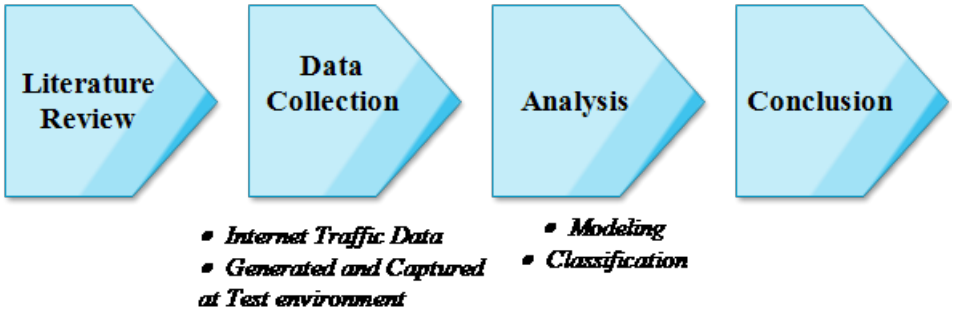

Figure 1.2: Summery of Methodology

1.5 Related Work

To model internet traffic in to different application, literature used different approaches. In [7] model flow length, flow size, flow inter arrival time and duration based on the collected data. The output of the performed analyses are empirical CDFs of flow length, size and duration of selected applications that is P2P and web. Fit particular distributions to the data and offer model parameters. It examine by using Transport Layer Protocol Transmission Control Protocol (TCP) port numbers. However, TCP port number is inefficient to identify applications.

In [8] flow characteristics of an empirical aggregated traffic data collected from The Cooperative Association for Internet Data Analysis (CAIDA) and Budapest University of Technology and Economics (BME) analyzed. Flow duration, number of flows (Packets),flow sizes (Bytes) and parameters are given along with the empirical CDF, but no models are derived.

In [9] focuses on modeling traffic class into multimodal (i.e. the variation of size) of different parameters. The technique used in the paper is clustering based on vector quantization (VQ) and Gaussian Mixture Model (GMM). The goodness of a model is related to the parameter used to model and classify. This paper used two parameters. Those are packet train length and packet train size.

Due to the fast growing of the Internet there is an increase in traffic and application protocol. Newly emerging applications using a high volume of data. Many different approaches have been presented to solve the traffic classification problem. The existing classification approaches still depend on the a well-known host port numbers, Internet Protocol (IP) and signatures for classification.

1.6 Scope and limitation

1.6.1 Scope of the Thesis

This study is focusing on statistical modeling of four applications and fit with theoretical distributions.

1.6.2 Limitation of the Thesis

This thesis is limited to capturing, and classifying HTTP, HTTPS,SSH and DNS packets on a test environment. In this thesis comparison of port-based classification and other classification algorithms is not done. Flow greater than one packet at both forward and backward directions are used.

1.7 Contribution of the Research

Statistical modeling of internet traffic using Machine Learning (ML) classification algorithms. To identify the applications usually is done by port-numbers. In this thesis identification of application is done by ML. Network traffic data generated in a controlled laboratory environment.

The output of this research helps to:

- Network administrators to manage the network and improve QoS by prioritizing applications running on the network.
- Service providers understand network traffic and use input for optimizing the data traffic and also gives clue about applications and there characteristics.
- This research helps a better understanding to the service provider in use of resource allocation(like bandwidth).
- Researchers who want to engage in modeling different applications on network traffic.

1.8 Thesis Organization

The rest of this paper is organized as follows. Chapter 2 provides the background of the TCP/IP model and the IP network traffic relevant to this research. Chapter 3 provides background information about ML and how it can be applied in IP traffic classification. This section also discussed the review of several statistical distributions applied throughout this research. The analytic models to fit the traffic data distribution are also discussed in the

detail. Chapter 4 discussed the details of the experimental analysis used in this research; network traffic generation and capturing, data pre-processing and feature selection, training, and evaluating the algorithms and modeling internet traffic and evaluation of the model are tasks discussed in this chapter. Chapter 5 the results obtained from Machine learning and statistical modeling. Finally, it concludes the paper with some final remarks and suggestions for possible future work.

2

TCP/IP MODEL

2.1 TCP/IP MODEL

TCP/IP specifies how data is transferred from source host to the destination host over in the internet. It provides end-to-end communications that identify how it should be layered into packets handled by TCP and addressing is handled by Internet Protocol (IP). The features used in this research is Flows of a packet. Flow is a combination of source IP, source port, destination IP and destination port) whiten the same application. TCP/IP is suited for internet.

IP is applicable in several types of communication networks including intranet, military network, home network, Third Generation (3G) and Forth Generation (4G) cellular networks. Most of the applications in a networking environment use IP as a transition medium protocol. Several IP protocols have been developed to group the communication protocol into a hierarchical structure. The Open System Interconnection (OSI) model (7 Layer) and TCP/IP model (5 Layer) are widely used in the communication network.

The TCP/IP protocol is named for the combination of the two most important protocols the TCP and IP[3]. TCP provides reliable, connection-oriented stream service over IP. It is a full-duplex connection between two end-hosts across a datagram network. It also provides

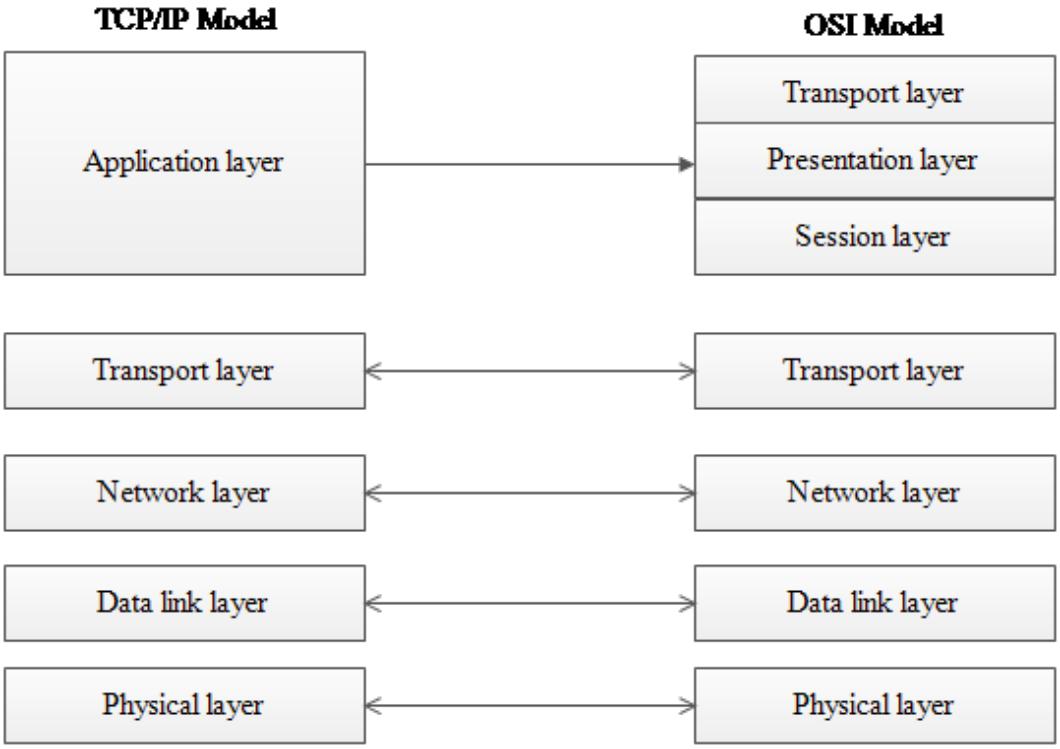

Figure 2.1: TCP/IP layers Mapping to the OSI layers [10]

flow and congestion control. source IP, destination IP, source port and destination port are the four tuple attributes that help TCP in uniquely identifying each connection.

TCP/IP is modeled in layer format [10]. The TCP/IP model contains four layers; Application layer, Network Access layer and Transport layer[11]. Those modeled are representing different protocols. Dividing into layers the protocols share the workload, easy for implementing and testing codes separately without affecting other layers. Layers communicating with above and below layers with interfaces. IP layer provides transfer of data from one host to another host without granting reliability but TCP delivers with reliability.

The TCP/IP protocol is considered that each device in a network has a unique IP Address (Internet Protocol Address) and each IP address can open and conversation over up to 65535 different ports for both ways communicate data among network devices. The IP Address uniquely detects the device on the network and a Port number recognizes a specific connection between one computer to another and is used to detect a unique connection between two devices. A TCP/IP port can be assumed as a two-way connection between source and destination.

2.2 Application Layer

The application layer is provided by the program that uses TCP/IP model for provides services and communicating to the end-users. The application layer is the topmost layer of the TCP/IP Model that provides the interface between the applications and the transport layer [10]. This layer is the final layer in the OSI model and is actually the user interface used to send or receive the data. Application layer gives several functions:

- The application layer is used to exchange messages.
- It provides access to global information about various services.
- It helps users to access files and manages it.

Below listed some commonly used application layer protocols.

Hypertext Transfer Protocol (HTTP)

The Hypertext Transfer Protocol is the basis of the World Wide Web (WWW). The main function of HTTP is to transfer web pages resources from the Web Server or HTTP server to the HTTP client. While you use a web browser such as Internet Explorer or Firefox, you are using a web client and also using the HTTP to transfer web pages to request from the remote servers.

Domain Name Server (DNS)

Every host in a network has a uniquely identified logical address that is IP address. These addresses are a group of numbers. If you visited a web site like www.google.com you are actually hosting IP adders. In reality, it is difficult to remember the IP address of every web site. The DNS helps to map web site like www.google.com to the IP adders. When you write the name of the web site into your browser first the system sends the DNS query to the DNS server to resolve the name of the IP address. After the name resolved HTTP or HTTPS session established.

Secure Socket Shell (SSH)

SSH is a network protocol that allows one computer securely connects to another computer and to access, transfer, and management of file over in an insecure network. SSH encrypts the data and transmitting to the internet. SSH is implemented as a client-server model.

Hypertext transfer protocol secure (HTTPS)

Like HTTP the HTTPS also used to exchange information from a web server and web browser. The main difference is that HTTPS is secure. The additional 's represent the connection is secure. Another difference is HTTP uses port 80 and HTTPS uses port 443.

2.3 Transport Layer

The transport layer delivers end-to-end transition from application to remote hosts. In the transport layer, multiple applications can communicate with hosts simultaneously. The most used TCP, which provides a reliable connection and User Datagram Protocol (UDP) is the basic transport layer protocol delivering unreliable data services[12]. This layer decides how data is sent and has the ability to perform error detection and validation of the process data.

2.3.1 Transmission Control Protocol (TCP)

TCP is designed in a TCP/IP suite. If the application layer wants to send data, it sends in to the lower layer(transport layer) for UDP or TCP to transport through the network. TCP established a virtual connection between source and destination the process to establish this connection is called three-ways-handshake then break the data into a segment and add a header to each segment send to the network layer.

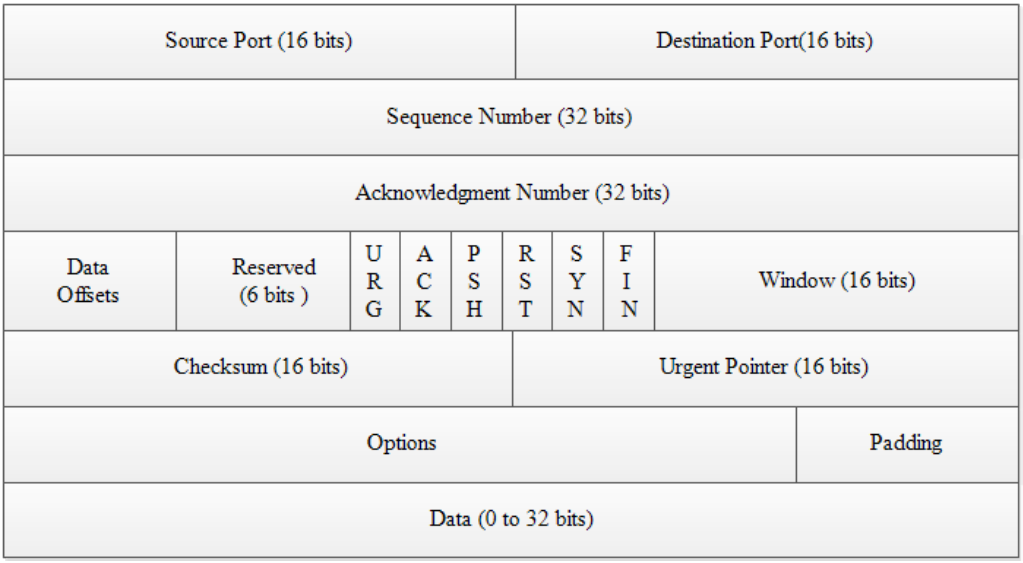

Figure 2.2: TCP Header Format [13]

The three-ways handshake uses Acknowledgment (ACK) and Sync (SYN) flags from the header section. These flags are the basic for TCP and discussed below.

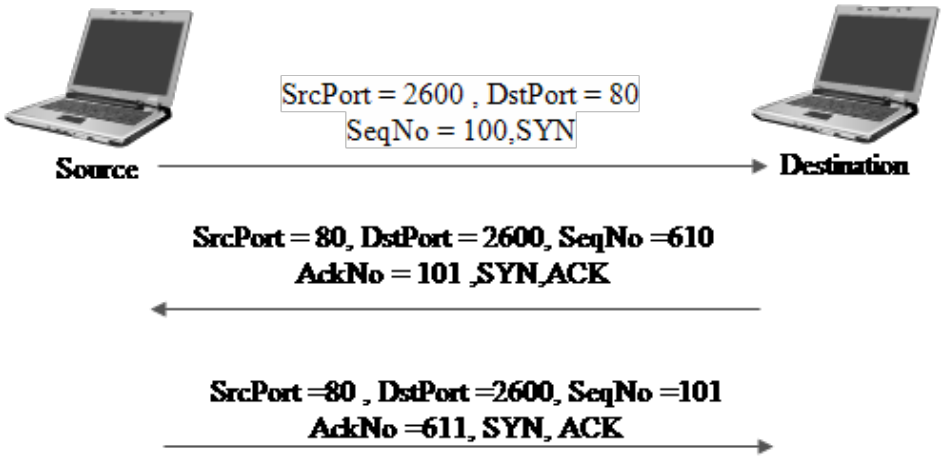

Figure 2.3: TCP Three Ways Handshake

In Figure 2-3, the source starts by sending a TCP header to the destination with the SYN flag set. The destination replies back with the SYN and ACK flag sent. the destination uses the received sequence number plus 1 as the acknowledgment number. This is because it is assumed that 1 byte of data was contained in the exchange. Finally the source answers back with only the ACK bit set. After this, the data flow can start.

2.3.2 User Datagram Protocol (UDP)

The common point between TCP and UDP is both use port numbers to transport traffic. UDP unlike TCP there is no Three Ways Handshake. UDP is connection less and unreliable protocols it provides data without overhead compared to TCP. The UDP header has five fields, They are:

- source port number, which is the number of the sender.
- destination port number, the port the datagram is addressed.
- length, the length in bytes of the UDP header and any encapsulated data
- checksum, which is used in error checking.
- Data, the actual data which holds the information

2.4 Network Layer

This layer also called the inter network layer provides the "virtual network" it is an image of the Internet. IP is the most important protocol in this layer IP provides a routing function

Source Port (16 bits)	Destination Port(16 bits)
Length (16 bits)	Checksum (16 bits)
Data (0 to 32 bits)	

Figure 2.4: TCP Header Format [14]

that transmitted message to there destination. This layer is tasked with finding the best route for the data to take in order to reach its destination. Protocols such as Routing Information Protocol (RIP), Border Gateway Protocol (BGP) and Open Shortest Path First (OSPF) are common at this stage.

3

Machine Learning and Review of Statistics

3.1 Machine Learning Algorithms

Machine Learning (ML) algorithms are used by mapping of instances of internet traffic flows into another or different internet traffic classes. flows by a set of statistical features correspond to feature values.Flow is a continuous packet chain associated with a single instances of application. In the internet traffic analysis flow is a conversation of four tuples (source IP, source port, destination IP and destination port) whiten the same application. A feature is a statistical value that can be calculated from more than one packets, such as the minimum value of packets or minimum packet size or the mean of inter-arrival times or the standard deviation of inter-arrival times. Every internet traffic flow is considered by a similar set of features, each flow exhibits different internet traffic classes where it belongs.

ML algorithms have the ability to learn from past experiences or historical data. There are several applications of machine learning ML Financial, Marketing, Health, Technology and Science it can be used to identify Geographical problems etc. Data maiming is widely used application [7]. Broadly there three types of machine learning algorithms. ML is needed in order to perform complex tasks for a human being to program or computing and which is not changed often [15].ML enables the organization to analyze the computational complexity to perform by manual methods with low accuracy. The organization may need to decide about

there network operations [**?**].

3.1.1 Supervised Machine Learning

Supervised Machine Learning algorithm contains a target (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these sets of variables, Generating a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Logistic Regression, etc.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) algorithm is a method for classifying instances based on the closest training instances in the feature space. KNN is a kind of instance-based learning where the function is only approximated locally and all computation is postponed until classification [16]. The KNN is the important and simplest classification method when it is little or no preceding knowledge approximate the distribution of the data [16]. This rule simply preserves the all-inclusive training set during learning and assigns to each request a class represented by the popular label of its k-nearest neighbors in the training set.

Support Vector Machine(SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm. When it uses as a classification algorithm it separates a given labeled training dataset with a hyper-plane the is the maximum distance from them (maximal margin hyperplane) [17]. The algorithm outputs an optimum hyperplane which categorizes new examples. In two dimensional space, this hyperplane is a line dividing a plane into two parts where each class lay on either side.

Decision Tree

Trees that classify occurrences by arranging them based on feature values referred to as decision trees. Each node in a decision tree represents a feature in an occurrence to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.[7].

- Check all cases belongs to the same class.
- Calculate information and information gain.
- Find the best split of the attribute.

Entropy:A decision tree is organized from a root node and involves partitioning the data into leaf nodes that contain instances with similar values. bits, nats or bans are the measurements of Entropy [17].

Let's define entropy on given collection S, entropy(S) is defined as:-

$$-\sum_{i} P(I)\log_2 P(I) \tag{3.1}$$

where, $p(I)$ is proportion of S belonging to class I.
Now, let's define $Gain(S,A)$ where S is example set and A is attribute:-

$$Entropy(S) - \sum_{v}(|S_v|/|S|) \times Entropy(S_v) \tag{3.2}$$

where, S_v is subset of S for which attribute A has value v.

3.1.2 Unsupervised Machine Learning

Unsupervised ML does not have any target or outcome variable to predict / estimate. It is used for the clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: K-means.

K-Means Clustering

K-means clustering algorithm is a commonly used clustering algorithm. For dividing into k-groups(i.e. K-closures). Where K is the value specified before clustering. It tries to cluster groups observed in the same relation to each other. Whereas observation with different groups is unlike. In K-means clustering, each group is represented by a centroid that is the mean of the observation in a group.

3.1.3 Reinforcement Machine Learning

Reinforcement Machine Learning the machine is trained to make specific decisions. It works this way: the machine is open to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible

knowledge to make accurate decisions. Example of Reinforcement Learning: Markov Decision Process

3.2 Probability Distributions

A probability distribution is a function that describes the likelihood of obtaining the possible values. This means, the variable changes based on the underlining probability distribution[18].

3.2.1 Poisson Distribution

Poisson distribution is appropriate in situations where events occur at random points of time. A distribution is called Poisson distribution when the following assumptions are fulfilled:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.

Any distribution fulfills the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

λ is the rate at which an event occurs, t is the length of a time interval, and X is the number of events in that time interval. Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution. Let μ denote the mean number of events in an interval of length t. Then, $\mu = \lambda$*t.

A Poisson random variable X with scale parameter μ has probability mass function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad x = 0, 1, 2, \ldots . \tag{3.3}$$

The cumulative distribution function on the support of X is

$$F(x) = P(X \leq x) = \frac{\Gamma(x+1,\ \mu)}{\Gamma(x+1)} \qquad x = 0, 1, 2, \ldots . \tag{3.4}$$

The mean and variance of X are

$$E[X] = \mu \tag{3.5}$$

$$V[X] = \mu \tag{3.6}$$

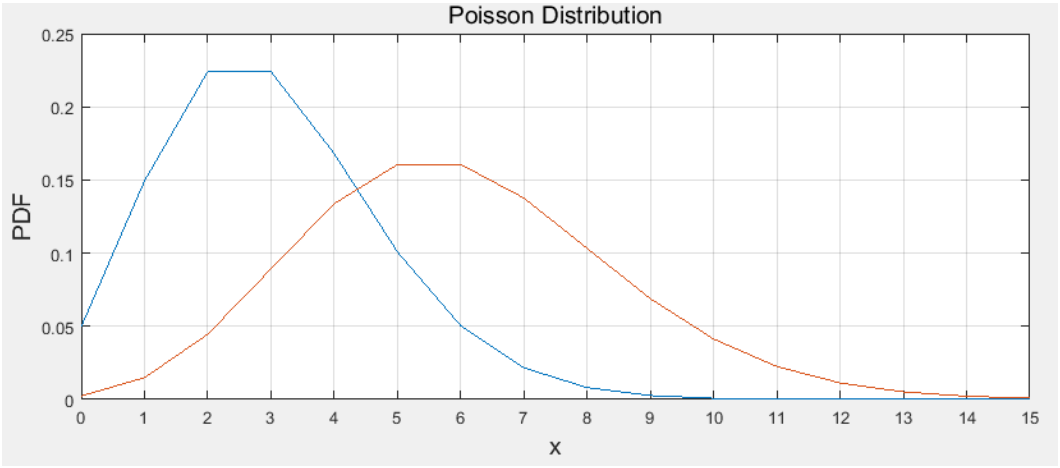

Figure 3.1: Poisson Distribution.

Figure 3.1 shows a Poisson distribution with μ =3 and μ = 6.

3.2.2 Log-normal Distribution

The shorthand $X \sim \log \text{normal}(\alpha, \beta)$ is used to indicate that the random variable X has the log normal distribution with parameters α and β. A log normal random variable X with parameters α and β has probability density function

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} e^{-\frac{1}{2}(\ln(x/\alpha)/\beta)^2} \qquad x > 0 \tag{3.7}$$

for α and $\beta > 0$. The Log-normal distribution can be used to model the lifetime of an object, the weight of a person, or service time. The central limit theorem indicates that the Log-normal distribution is useful for modeling random variables that can be thought of as a product of several independent random variables.

The cumulative distribution function on the support of X is

$$F(x) = P(X \leq x) = \frac{1}{2} + \frac{1}{2}\,\text{erf}\left(\frac{\sqrt{2}\,(\ln(x) - \alpha)}{2\beta}\right) \qquad x > 0, \tag{3.8}$$

where (erf error function)

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt. \tag{3.9}$$

The mean and variance of X are:

$$E[X] = \alpha e^{\beta^2/2} \tag{3.10}$$

$$V[X] = \alpha^2 e^{\beta^2}(e^{\beta^2} - 1) \tag{3.11}$$

3.2.3 Normal Distribution

The shorthand $X \sim \mathrm{N}(\mu, \sigma^2)$ is used to indicate that the random variable X has the normal distribution with parameters μ and σ^2. A normal random variable X with mean μ and variance σ^2 has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{})^2}{2\sigma^2}} \qquad -\infty < x < \infty,$$

for $-\infty < \mu < \infty$ and $\sigma > 0$.
The population mean and variance of X are:

$$E[X] = \mu \tag{3.12}$$

$$V[X] = \sigma^2 \tag{3.13}$$

3.2.4 Weibull Distribution

The Weibull distribution with scale parameter $\alpha > 0$ and shape parameter $\beta > 0$. A Weibull random variable X has probability density function

$$f(x) = \frac{\beta}{\alpha} x^{\beta-1} e^{-(1/\alpha)x^{\beta}} \qquad x > 0. \tag{3.14}$$

The probability density function is plotted below for $\alpha = 1$ and $\beta = 1/2, 1, 2, 3$.

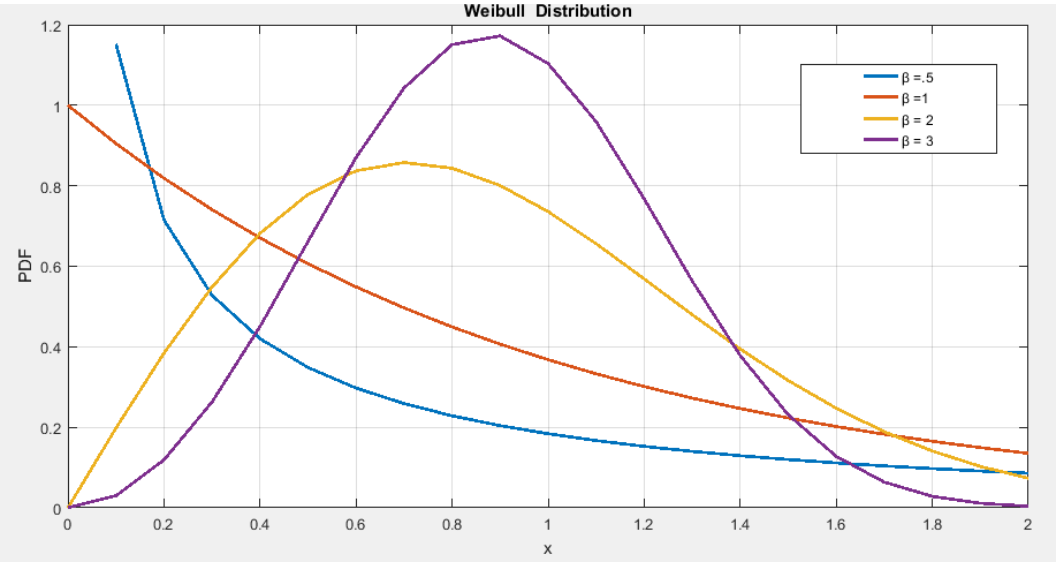

Figure 3.2: Weibull Distribution

The cumulative distribution function on the support of X is

$$F(x) = P(X \leq x) = 1 - e^{-(1/\alpha)x^{\beta}} \qquad x > 0. \tag{3.15}$$

The population mean and variance of X are:

$$E[X] = \frac{\alpha}{\beta}\Gamma\left(\frac{1}{\beta}\right) \tag{3.16}$$

$$V[X] = \alpha^2 \left\{ \frac{2}{\beta}\Gamma\left(\frac{2}{\beta}\right) - \left[\frac{1}{\beta}\Gamma\left(\frac{1}{\beta}\right)\right]^2 \right\} \tag{3.17}$$

3.2.5 Rayleigh Distribution

The random variable X has the Rayleigh distribution with parameter α.A Rayleigh random variable X with positive parameter α has probability density function

$$f(x) = \frac{2xe^{-x^2/\alpha}}{\alpha} \qquad x > 0. \tag{3.18}$$

The probability density function with three different parameter settings is illustrated below.

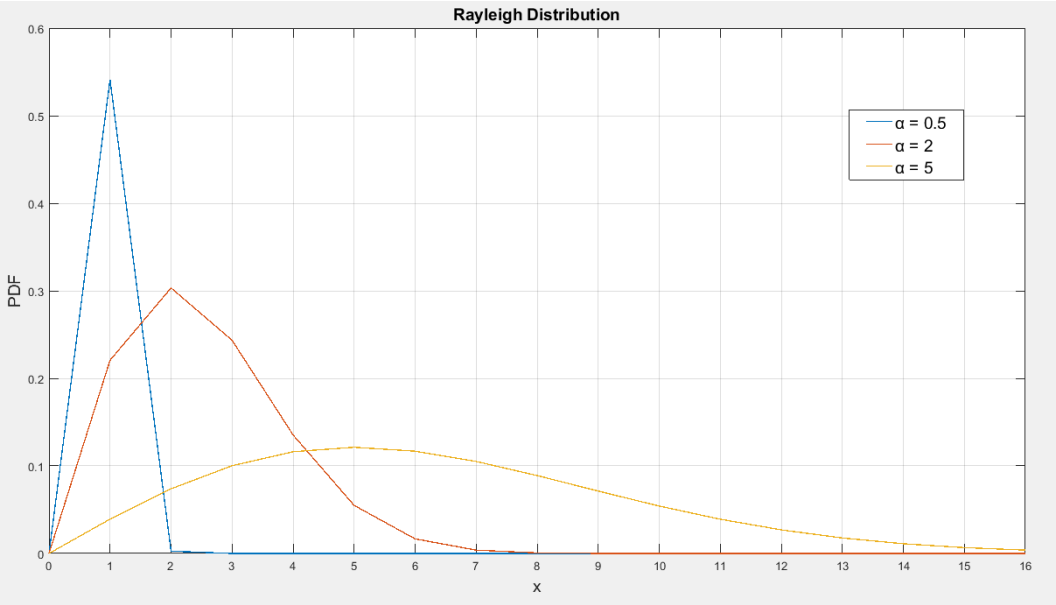

Figure 3.3: Rayleigh distribution

The cumulative distribution function on the support of X is

$$F(x) = P(X \leq x) = 1 - e^{-x^2/\alpha} \qquad x > 0. \tag{3.19}$$

The mean and variance of X are:

$$E[X] = \frac{\sqrt{\alpha\pi}}{2} \tag{3.20}$$

$$V[X] = \frac{\alpha(4-\pi)}{4} \tag{3.21}$$

3.3 Best Fitted Distribution Parameter Selection

3.3.1 Maximum Likelihood Estimation(MLE)

Maximum Likelihood Estimation (MLE) could be a favored strategy of parameter estimation in insights instrument for numerous statistical modeling technique. Once information have been collected and the probability work to show a given information is determined, one is in a position to form measurable deductions almost the populace, that's , the likelihood dissemination that underlies the information [19].

4

Experimental Analysis

The experimental process provided in this research is discussed in detail. Figure 4.1 shows the experimental process steps. the experimental process is categorized into four different sections. Those sections are discussed separately; internet traffic Generation and Capturing, Data pre-processing, Classification algorithm, and Classify unseen data set.In this research Flow is considered an a continuous packet chain associated with a single instances of application. In the internet traffic analysis flow is a conversation of four tuples (source IP, source port, destination IP and destination port) within the same application.

4.1 Internet Traffic Generation and Capturing

The classification algorithm needs a labeled traffic data set. Generating and capturing of internet traffic consists of four applications and data set is similar to that of publicly available data set. Generating and capturing of the internet traffic is done by Wireshark which is open-source traffic analysis software. For training and testing purposes prepare two separate data sets with the different applications. Those applications are; DNS, HTTPS, SSH, and HTTP. Table 4.1 shows the list of applications ready for this research. There is also an unlabeled data set captured from the ethio telecom core switch the need for this raw data set is to classify based on the training algorithm for internet traffic modeling. For generating and

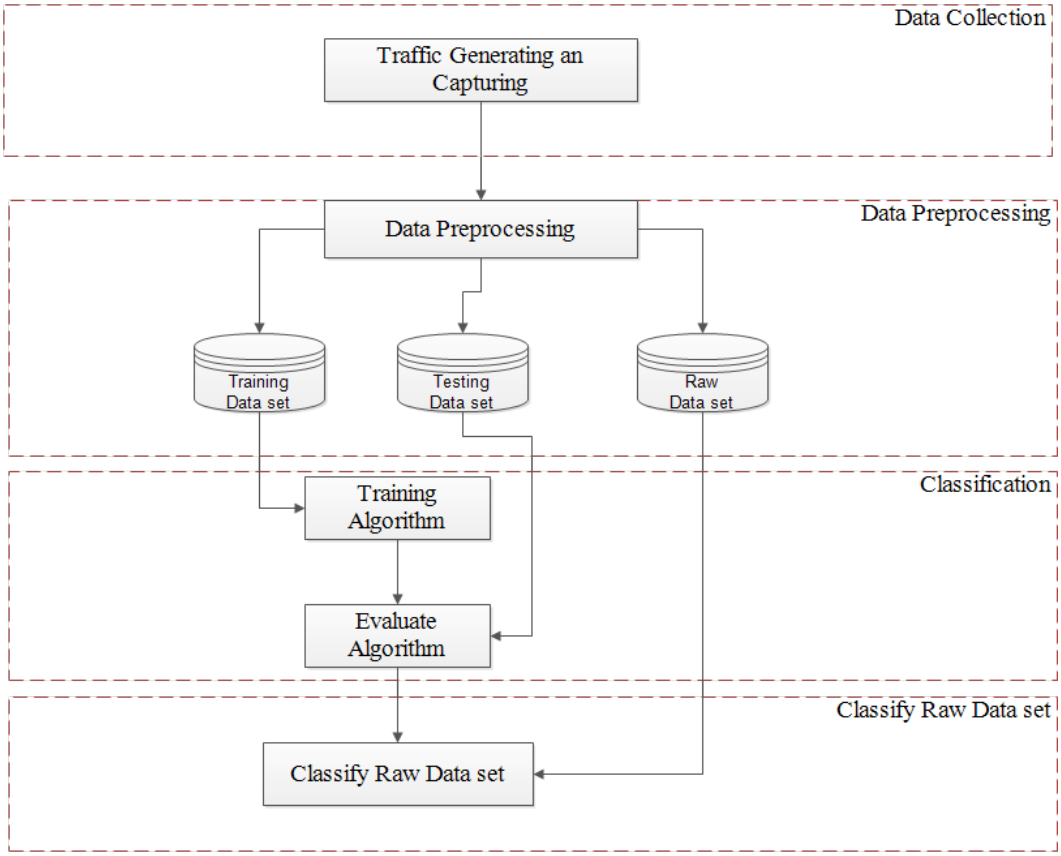

Figure 4.1: Experimental Processes

capturing traffic data set Dell LATITUDE Laptop 4 GB RAM and 4 CPUs with 2.9 GHz clock rate is used and Huawie 3G Wi-Fi dongle is used.

Table 4.1: Experimental data set

App.	Data Set 1 (No of Flows)	Data Set 2 (No of Flows)	Raw Data set (No of flows)
DNS	546	548	16927
HTTP	1972	512	
SSH	256	151	
HTTPS	2330	509	

Figure 4.2 shows the percentage representation of training and testing data sets 75% for training and 25% for testing the classification model.

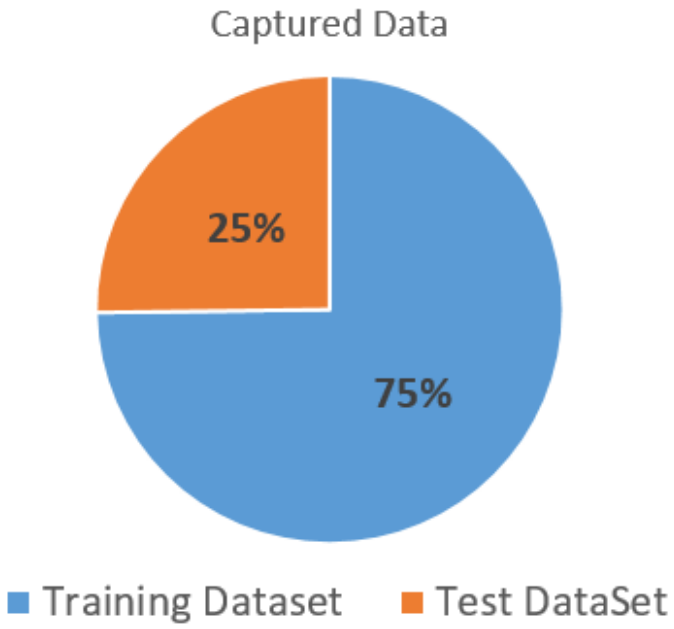

Figure 4.2: Data Set Representation

4.2 Data Pre-processing

The data set used in this research is generated using wireshark tool. So, some tasks should be done in the captured data set. Such tasks are; attribute selection and correlation and handling of outliers and missed data. Figure 4.3 shows take handled in this section and discussed in detail.

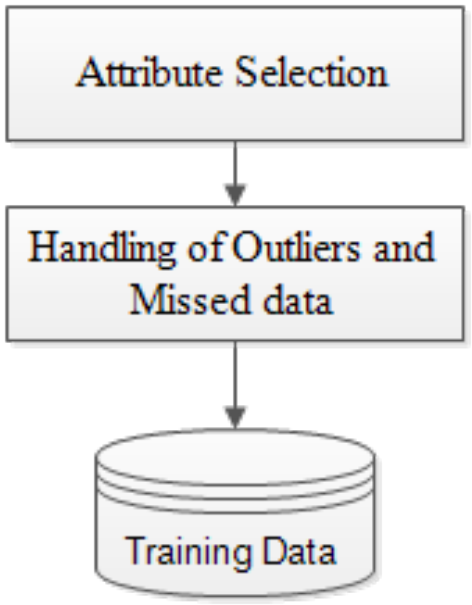

Figure 4.3: Data Pre-processing [20]

Attribute Selection

After captured the network traffic data set, the next step is attribute selection and groping of packets in to flows. In this section the flows are extracted from the packet data such as packet length, packet size, protocol, duration source, and destination IP and source and destination port number, etc. the extracted feature is used for train ML classifier. For grouping flows, we use the MS-excel pivot and extracting 22 features and one class. Those 22 features are reduced to 14 using person correlation coefficient methods. Table 4.2: shows attributes befor selection and selected attributes

Table 4.2: Feature

No	Short hand	Description	Selected
1	min_fpktl	minimum forward packet length	x
2	mean_fpktl	mean forward packet length	x
3	max_fpktl	maximum forward packet length	x
4	std_fpktl	standard deviation forward packet length	
5	min_bpktl	minimum backward packet length	
6	mean_bpktl	mean backward packet length	x
7	max_bpktl	maximum backward packet length	x
8	std_bpktl	standard deviation backward packet length	
9	min_fiat	minimum forward inter-arrival time	
10	mean_fiat	mean forward inter-arrival time	x
11	max_fiat	maximum forward inter-arrival time	x
12	std_fiat	standard deviation forward inter-arrival time	
13	min_biat	minimum backward inter-arrival time	
14	mean_biat	mean backward inter-arrival time	x
15	max_biat	maximum backward inter-arrival time	x
16	std_biat	standard deviation backward inter-arrival time	
17	duration	duration	x
18	proto	protocol	
19	total_fpackets	total forward packets	x
20	total_fvolume	total forward volume	x
21	total_bpackets	total backward packets	x
22	total_bvolume	total backward volume	x

Handling of Outlier and Missed Data

After extracting all the features listed above, the Ms-excel saves the output in simple ".csv" file format.

Flow Length

Flow length was measured in number of packets in a corresponding flow.

Flow Size

Flow size measured in number of bytes in a corresponding flow.

Handling of Outlier and Missed Data

The outlier is value that did not have consistent values with the rest of the data set. In this case, since we are selecting the attributes based on flows, so flows whose packet length is less than two packets in both forward and backward direction is considered as an outlier. Flows who did not have one of the values considered as a tuple is missed data after removing both the outlier and missed data shown in Table 4.3

Table 4.3: Outliers and missed Data

Data sets	**No of Flows**
Data set-1	1228
Data set-2	26
Raw Data set	0

4.3 Training Algorithm

After the pre-processing step is completed the next step is training the selected algorithm which is the decision tree classification algorithm and building the classification model. The first step is bringing two data set; data set-1 and data set-2 for training and testing respectively. After training the data set-1 build an ML model and evaluate the model with data set 2 and finally predict previously unseen data set-2 with the model building previously. The training is done by the Matlab classification learner app which is built-in MATLAB tool.

4.4 Evaluation Method and Performance Metric

As an evaluation metrics, confusion matrix and classification accuracy which are widely used evaluation metrics in classification tasks are used in this research.

Accuracy

The important condition distinguishing the performance of the classification model is predictive accuracy (i.e., how accurately a classification model makes decisions when predicting with previously unseen data). A number of metrics can express the predictive accuracy. A traffic classifier is being used to identify (classify) flows of packets belonging to the class. A

common method to distinguish a classifier's accuracy is through metrics known as Accuracy [21]. Accuracy is the percentage of correctly classified occurrences among the total number of occurrences. Table 4.4 shows the

Table 4.4: Classification Accuracy

\	*Class 1* *Predicted*	*Class 2* *Predicted*
Class 1 *Actual*	TP	FN
Class 2 *Actual*	FP	TN

Here,

- Class 1 : Positive
- Class 2 : Negative

Definition of the Terms:

- Positive (P) : Observation is positive
- Negative (N) : Observation is not positive
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Accuracy is one metric for evaluating classification models. Accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Classification Rate or Accuracy is given by the relation:

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

Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken

down by each class. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

5

Result and Discussion

5.1 Classification

Separate test data validation techniques have been used for traffic classification and decision tree classification algorithm is selected. Size of the separate test data and description of the evaluation metrics; confusion matrix and overall accuracy.Table 5.1 shows the confusion matrix of the data.

Table 5.1: Confusion Matrix

		Predicted Class				
		dns	https	ssh	http	
Actual Class	dns	548	0	0	0	548
	https	0	397	0	112	509
	ssh	0	7	140	4	151
	https	15	21	0	477	512
		563	425	140	593	1720

An accuracy of 90.8% is observed is a separate test data. Based on this accuracy classified the second an seen data set namely raw data set. These raw data set is classified based on the model build previously in the chapter 4. The next section is used this classified raw data set.

5.2 Fitting Internet Traffic and Parameter Estimation

The results obtained best fits distribution for applications are discussed in this section. Distributions are defined by parameters the maximum log likelihood estimation method is used to estimate the distribution's parameters from the data. Figure 5.6shows the summery of flow length and Figure 5.10 shows flow size with there corresponding fitted distribution. Flow length of the three applications, DNS, HTTP and HTTPS are approximated by log-normal distribution with different shape and scale parameters whereas the flow length of SSH is approximated by Weibull distribution. On the other hand flow size of DNS, HTTP and HTTPS are also approximated log-normal and SSH is approximated by Weibull. For description purpose the plot is visualized in log scale. for detail result analysis and estimated parameters for each applications the distribution of the measured data is discuses in next section 5.3 and 5.4.

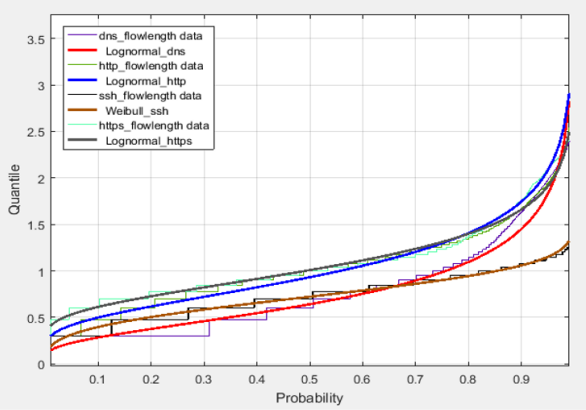

Figure 5.1: Distribution of Flow Length

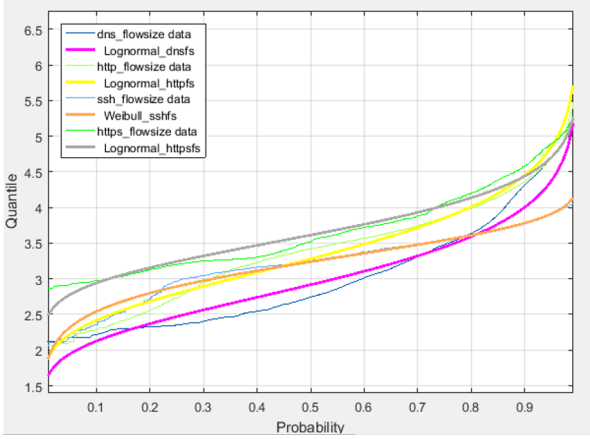

Figure 5.2: Distribution of Flow Size

5.3 Internet Traffic Flow Length

5.3.1 Mean, Variance and Log likelihood

Based on the log likelihood estimation compering four theoretical models the value which is maximum log likelihood value is the bast fitted distribution to the application. Table 5.2 shows the mean,Variance and log likelihood values of four applications. For DNS application best fitted distribution is Log-Normal which has the maximum log-likelihood value compered to the the other three distribution.similarly HTTP and HTTPS application best fitted distribution is Log-Normal which has the maximum log-likelihood value compered to the the other three distribution. On the other hand SSH application best fitted with Weibull distribution which has the maximum log-likelihood value compered to the the other three distribution.

Table 5.2: Mean,Variance and Log likelihood of Flow Length

App	Distribution	Mean	Variance	Log likelihood
DNS	Normal	0.790135	0.292351	-8946.84
	Log-normal	0.786174	0.310262	-5844.32
	Weibull	0.797194	0.262615	-6788.68
	Rayleigh	0.848485	0.196712	-7386.57
HTTP	Normal	1.04125	0.230792	-2198.92
	Log-normal	1.05088	0.299954	-2037.41
	Weibull	1.04265	0.232976	-2038.45
	Rayleigh	1.01623	0.282183	-2088.22
SSH	Normal	0.725071	0.066504	-60.2583
	Log-normal	0.73064	0.094564	-110.953
	Weibull	0.726461	0.064492	-46.3234
	Rayleigh	0.68197	0.127079	-170.056
HTTPS	Normal	1.08916	0.202933	-1017.51
	Log-normal	1.08802	0.194207	-794.112
	Weibull	1.09049	0.213443	-950.675
	Rayleigh	1.0445	0.298101	-1027.18

The average packet numbers in a flow is 7 packets for DNS,10 Packets for HTTP , 7 Packets for SSH and 10 Packets for HTTPS.

5.3.2 CDF

In Figure 5.3 CDF of flow length (packets) for DNS application are shown. The DNS traffic category were best approximated by the Log-normal distribution. The majority of the packets around 90% is less than 30 packets and only a few amount that is 10% is grater than 30 Packets.

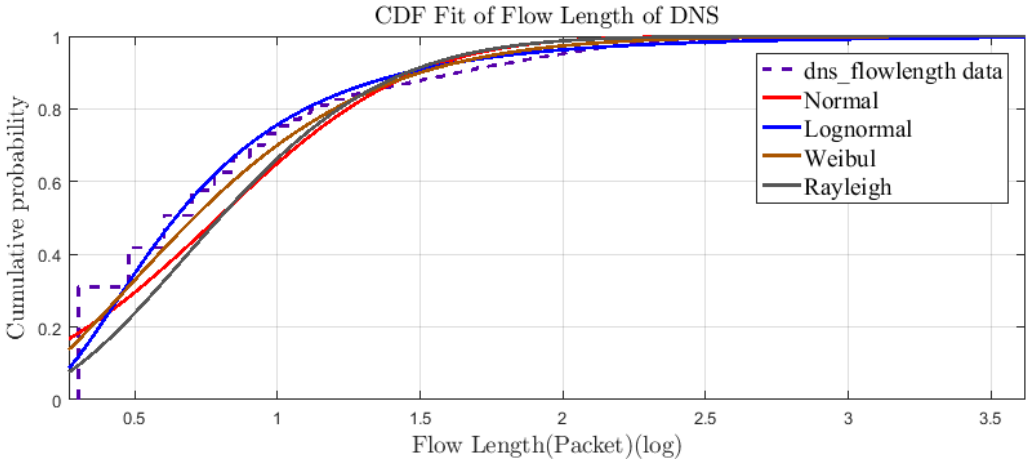

Figure 5.3: CDF Fit of Flow Length of DNS

In Figure 5.4 CDF of flow length (packets) for HTTP application are shown. The DNS traffic category were best approximated by the Log-normal distribution. The majority of the packets around 80% is less than 30 packets and only a few amount that is 20% is grater than 30 Packets.

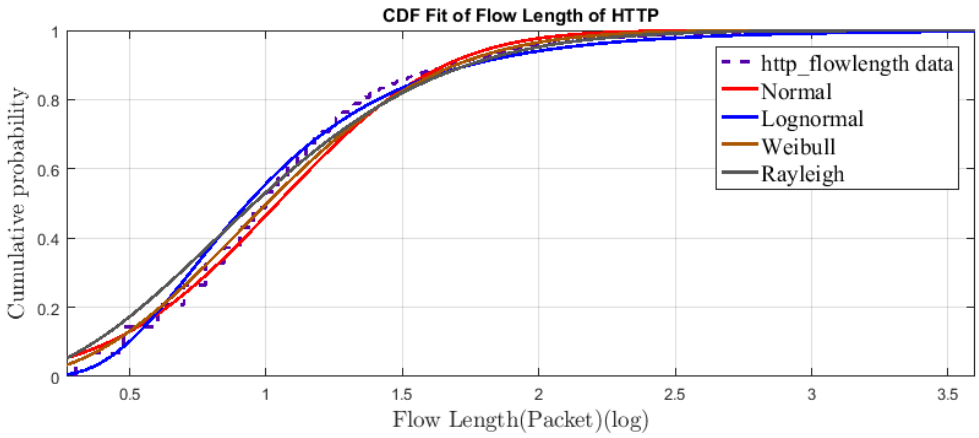

Figure 5.4: CDF Fit of Flow Length of HTTP

In Figure 5.5 CDF of flow length (packets) for SSH application are shown. The SSH traffic category were best approximated by the Weibull distribution. The majority of the packets around 80% is less than 10 packets and only a few amount that is 20% is grater than 10 Packets.

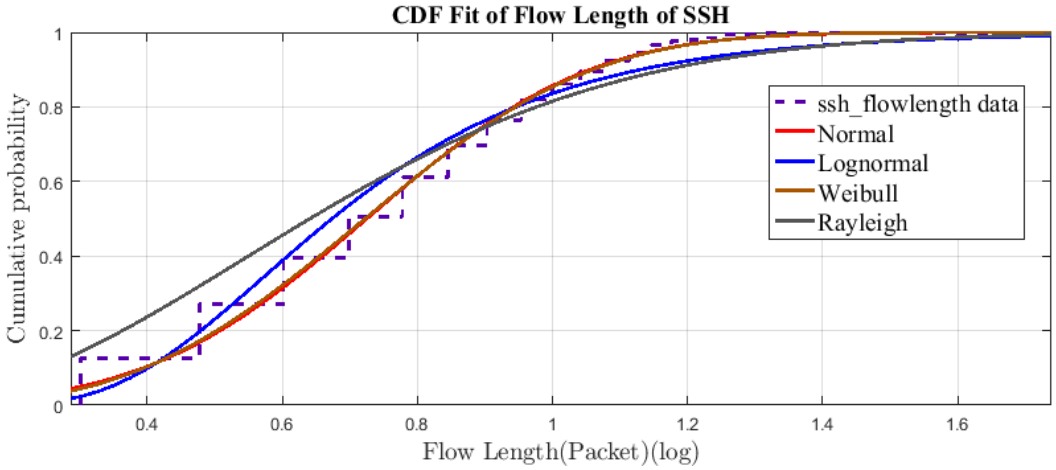

Figure 5.5: CDF Fit of Flow Length of SSH

In Figure 5.6 CDF of flow length (packets) for HTTPS application are shown. The HTTPS traffic category were best approximated by the Log-normal distribution. The majority of the packets around 90% is less than 30 packets and only a few amount that is 10% is grater than 30 Packets.

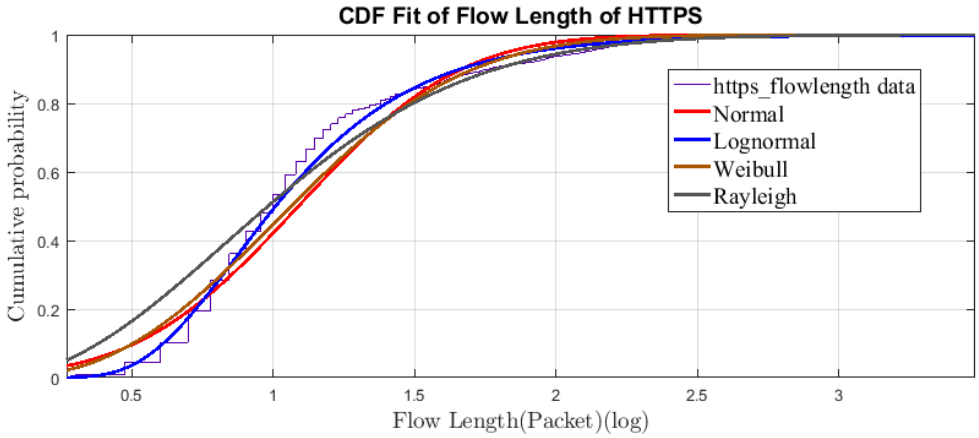

Figure 5.6: CDF Fit of Flow Length of HTTPS

5.3.3 Estimated Parameters

The distribution parameters are shows in Table 5.3.For DNS traffic with shape factor α=-0.443971 and β=0.637799. The distribution parameters are given in Table 5.3 Parameter HTTP traffic with shape factor α=-0.070519 and β=0.49019. The distribution parameters. For SSH traffic with shape factor α=0.811916 and β=3.13304. The distribution parameters. For HTTPS traffic with shape factor α=0.00840772 and β=0.389756.

most parameters are shape and scale parameters of a graph.

Table 5.3: Estimated Parameters flow length of DNS

App	Distribution	Flow Length
DNS	Mu ()	-
	Sigma ()	-
	Beta (β)	0.637799
	Alpha(α)	-0.443971
HTTP	Mu ()	-
	Sigma ()	-
	Beta (β)	0.49019
	Alpha(α)	-0.070519
SSH	Mu ()	-
	Sigma ()	-
	Beta (β)	3.74923
	Alpha(α)	3.32539
HTTPS	Mu	-
	Sigma	-
	Beta (β)	0.389756
	Alpha(α)	0.00840772

5.4 Internet Traffic Flow Size

5.4.1 Mean, Variance and Log likelihood

The flow Size (Byte) of application was analyzed. The mean,variance and Log likelihood values of four theoretical distributions are shown in Table 5.4. Based on the log likelihood estimation compering four theoretical models the value which is maximum log likelihood value is the bast fitted distribution to the application. For DNS application best fitted distribution is Log-Normal which has the maximum log-likelihood value is -12161.2 compered to the the other three distribution.similarly HTTP and HTTPS application best fitted distribution is Log-Normal which has the maximum log-likelihood value -3764.67 and -1436.64 respectively compered to the the other three distribution. On the other hand SSH application best fitted with Weibull distribution which has the maximum log-likelihood value -661.55 compered to the the other three distribution.

Table 5.4: Mean,Variance and Log likelihood of Flow size

App	Distribution	Mean	Variance	Log likelihood
DNS	Normal	3.01292	0.657246	-13454.3
	Log-normal	3.00796	0.571147	-12161.2
	Weibull	3.0032	0.797932	-13798.7
	Rayleigh	2.76509	2.08912	-16824.5
HTTP	Normal	3.37365	0.62168	-3787.85
	Log-normal	3.37558	0.667472	-3764.67
	Weibull	3.36555	0.70453	-3853.54
	Rayleigh	3.07036	2.57587	-5143.72
SSH	Normal	3.18874	0.253212	-697.994
	Log-normal	3.19101	0.296467	-752.67
	Weibull	3.1934	0.238719	-661.55
	Rayleigh	2.86088	2.23637	-1435.47
HTTPS	Normal	3.66172	0.381247	-1533.95
	Log-normal	3.66036	0.351973	-1436.64
	Weibull	3.64209	0.520949	-1656.17
	Rayleigh	3.2909	2.95919	-2696.38

5.4.2 CDF

In Figure 5.7 CDF of flow size (byte) for DNS application are shown. The DNS traffic category were best approximated by the Log-normal distribution. The majority of the flow size around 90% is less than 10 kbyte Bytes and only a few amount that is 10% is grater than 10 kbyte.

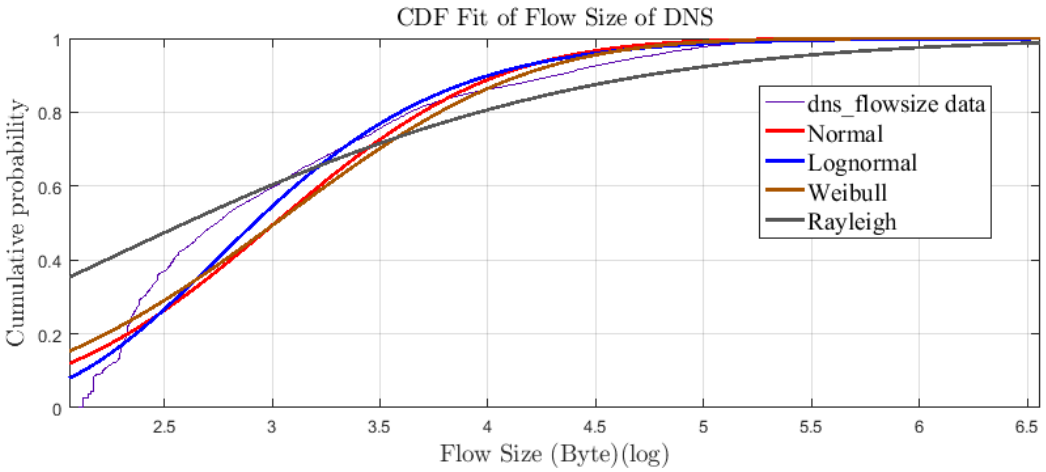

Figure 5.7: CDF Fit of Flow Size of DNS

In Figure 5.8 CDF of flow size (byte) for HTTP application are shown. The HTTP traffic category were best approximated by the Log-normal distribution. The majority of the flow

around 80% is less than 10 kbyte Bytes and only a few amount that is 10% is grater than 10 kbyte.

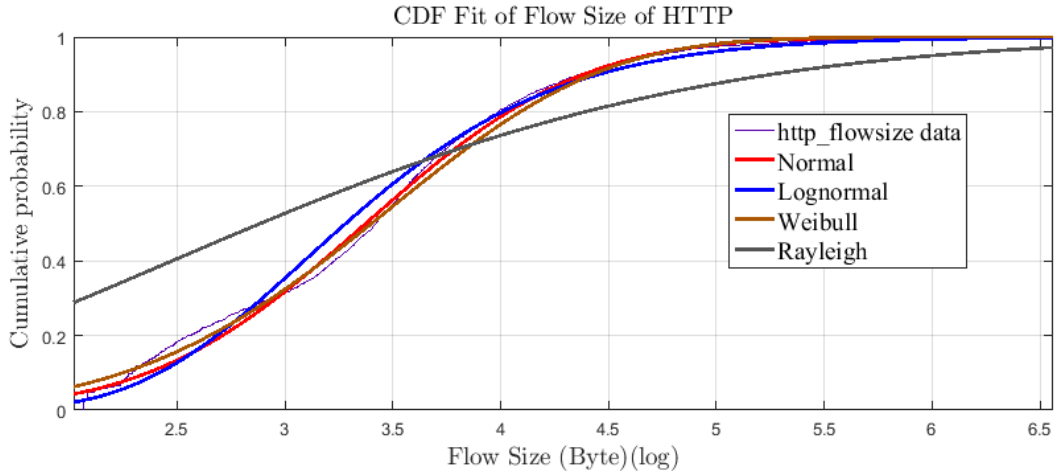

Figure 5.8: CDF Fit of Flow Size of HTTP

In Figure 5.9 CDF of flow size (byte) for SSH application are shown. The SSH traffic category were best approximated by the Weibull distribution. The majority of the flow around 98% is less than 10 kbytes and only a few amount that is 2% is grater than 10 kbyte.compered to the other application the flow size of SSH is around 10 kbyte.

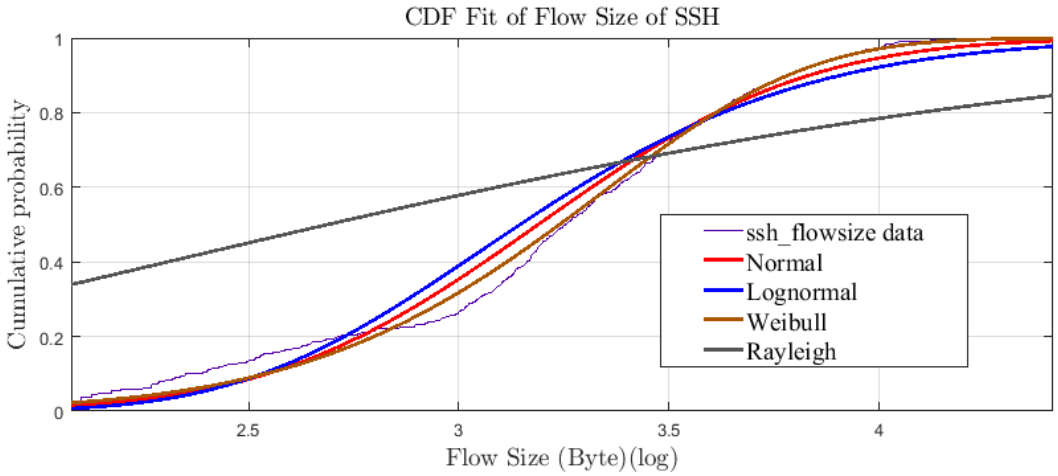

Figure 5.9: CDF Fit of Flow Size of SSH

In Figure 5.10 CDF of flow size (byte) for HTTPS application are shown. The HTTPS traffic category were best approximated by the Log-normal distribution. The majority of the flow around 90% is less than 25 kBytes and only a few amount that is 10% is grater than 25 kbyte. comperd to other applications HTTPS flow size is around 25 kbyte.

Applications flow size are difference in size compered to each other but HTTPS is double in in flow size than the other three applications.

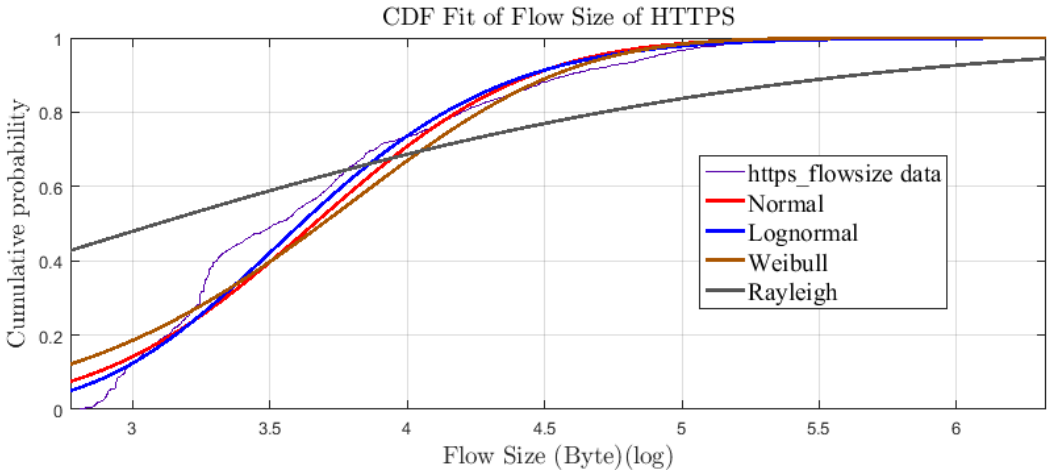

Figure 5.10: CDF Fit of Flow Size of HTTPS

5.4.3 Estimated Parameters

The distribution parameters are show in Table 5.5. For DNS application the parameters shape factor α= -0.443971 and β=0.637799. The HTTP application parameters has α= 1.18811 and β=0.238593.The SSH application parameters has α= 3.39642 and β=7.74099.The HTTPS application parameters has α= 1.2846 and β=1.2846.

Table 5.5: Estimated Parameters Flow Size

App	Distribution	Flow Size
DNS	Mu ()	-
	Sigma ()	-
	Beta (β)	0.637799
	Alpha(α)	-0.443971
HTTP	Mu ()	-
	Sigma ()	-
	Beta (β)	0.238593
	Alpha(α)	1.18811
SSH	Mu ()	-
	Sigma ()	-
	Beta (β)	7.74099
	Alpha(α)	3.39642
HTTPS	Mu ()	-
	Sigma ()	-
	Beta (β)	0.16103
	Alpha(α)	1.2846

6

Conclusion and Future Work

6.1 Conclusion

In this thesis, real traffic traces collected over ethio telecom core switch was analyzed and modeled to study the underlying traffic patterns of four different applications. One has to have a profound understanding of the packet header structure for various protocols to study the traffic pattern. For the given dataset, we investigated traffic flow length patterns and flow size pattern for most common protocols in use.

The size of the data is limited to 16927 flows around 174619 forward packets and 211549 backward packets in 3 to 5 minutes. To collect packet data is challenging mainly in the aspects of privacy issue of the users and the size of the captured packet data is very high in a specific time interval. Several research conduct using publicly available data set and limited to campus network.

- There is a variation in mean of different applications but closely related between HTTP and HTTPS in flow length but the variation of means in flow size is also observed but DNS have significant variation compered to those three applications.
- Related to the flow length and flow size applications has variations and this gives an overview (mean of flow length and flow size for different applications) to service

providers for strategy for optimization technique. This research intended to limited applications the major applications are studied but it is also included variations.

- This research also provides identification of applications using machine learning technique. Classification of applications using this technique is further studied with better data size and also improved the accuracy of the classification.

6.2 Future work

Modeling internet traffic applications in network with different parameters and several varieties of application impact the network and the QoS some of future works related are described below.

- The challenge will be to studied in more detail the relevance of ML classification than port-based classification for internet traffic modeling.ML training and testing data is generated in a controlled laboratory environment with limited data size and comparison of appropriate ML Algorithm is selected by comparing different ML.
- Future traffic characteristic could be done in detail to remove certain error in the identified Log-normal and Weibull parameters for different applications.

References

[1] "Cisco visual networking index: Forecast and trends, 2017–2022 white paper." https://www.cisco.com/, 2019. Acssesed November 30, 2019.

[2] H. Hassan, J. Garcia, and O. Brun, "Internet traffic modeling," in *2006 2nd International Conference on Information & Communication Technologies*, vol. 2, pp. 3169–3174, IEEE, 2006.

[3] S. B. Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *ACM SIGCOMM Computer Communication Review*, vol. 31, pp. 111–122, ACM, 2001.

[4] ethio telecom, "Mobie subscriber." https://www.ethiotelecom.et/, 2019.

[5] V. Ndatinya, Z. Xiao, V. R. Manepalli, K. Meng, and Y. Xiao, "Network forensics analysis using wireshark," *International Journal of Security and Networks*, vol. 10, no. 2, pp. 91–106, 2015.

[6] J. Cai, Z. Zhang, and X. Song, "An analysis of udp traffic classification," in *2010 IEEE 12th International Conference on Communication Technology*, pp. 116–119, IEEE, 2010.

[7] M. Pustisek, I. Humar, and J. Bester, "Empirical analysis and modeling of peer-to-peer traffic flows," in *MELECON 2008-The 14th IEEE Mediterranean Electrotechnical Conference*, pp. 169–175, IEEE, 2008.

[8] M.-S. Kim, Y. J. Won, and J. W. Hong, "Characteristic analysis of internet traffic from the perspective of flows," *Computer Communications*, vol. 29, no. 10, pp. 1639–1652, 2006.

[9] D. M. Divakaran, H. A. Murthy, and T. A. Gonsalves, "Traffic modeling and classification using packet train length and packet train size," in *International Workshop on IP Operations and Management*, pp. 1–12, Springer, 2006.

[10] S. A. Fedaghi, A. Alsaqa, and Z. Fadel, "Conceptual model for communication," *arXiv preprint arXiv:0912.0599*, 2009.

[11] M. M. Alani, "Tcp/ip model," in *Guide to OSI and TCP/IP models*, pp. 19–50, Springer, 2014.

[12] P. B. Nath and M. M. Uddin, "Tcp-ip model in data communication and networking," *American Journal of Engineering Research*, vol. 4, no. 10, pp. 102–107, 2015.

[13] J. Postel, "Rfc 793: Transmission control protocol," 1981.

[14] J. Postel, "User datagram protocol," *Isi*, 1980.

[15] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[16] Y. Liao and V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," *Computers & security*, vol. 21, no. 5, pp. 439–448, 2002.

[17] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[18] M. J. Evans and J. S. Rosenthal, *Probability and statistics: The science of uncertainty*. Macmillan, 2004.

[19] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[20] H. Tewodros, *Network Traffic Classification Using Machine Learning: A Step Towards Over-the-Top Bypass Fraud Detection*. PhD thesis, 2018.

[21] D. Qin, J. Yang, J. Wang, and B. Zhang, "Ip traffic classification based on machine learning," in *2011 IEEE 13th International Conference on Communication Technology*, pp. 882–886, IEEE, 2011.