

TIME SERIES ANALYSIS OF DISTRIBUTED DENIAL OF SERVICE ATTACKS
IN INTRUSION DETECTION SYSTEMS

by

Cemil Eren Kayataş

B.S., Electronics Engineering, İstanbul Kültür University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2018

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor, Prof. Emin ANARIM, for his great support and understanding in this long journey. He has that limitless kindness, encouragement and patience for me and this study. Without his supervision and great expertise this thesis would not have been possible.

I also acknowledge Prof. Mehmet AKAR and Prof. Tayfun AKGÜL for constructive criticism and spending their valuable time.

Ramin FADAEI FOULADI was with me on this tough road since day one. His positive approach and motivation keep me on going. I thank him for being not just a friend but more like a brother to me.

I would like to thank to our DDoS group members, Veronica BOLON and Mark HALL for their valuable support on my thesis. I also thank to all my friends in BUSIM Lab, Oğuzhan ÇEPNİ, Aytaç DEMİRKIRAN, Kadircan CEYHAN, Ozan MİNEZ and Sennur ALPKAYA for their support while finishing this thesis.

In my Bogazici life, Aynur İŞLER had the endless support and kindness to me. I take this as an opportunity to thank her and Leyla ÇEKEN for their patience and favour during this time.

Finally, I would like to thank to my father Yavuz KAYATAŞ and my sister Elif KAYATAŞ. They have always supported me in my whole life. I devoted this thesis to my beloved mother Sevim KAYATAŞ. She is the one who is always with me no matter what. I am always proud of being her son.

ABSTRACT

TIME SERIES ANALYSIS OF DISTRIBUTED DENIAL OF SERVICE ATTACKS IN INTRUSION DETECTION SYSTEMS

Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks cause million dollar losses to the companies, hence network security has become one of the most important topic in the modern world. DDoS attacks especially distinguish the availability where confidentiality and integrity are the other two main elements of Information security services. Data analysts try to prevent or detect DDoS attacks by applying different methods and using different data analysis tools. In this thesis, time series analysis of DDoS attacks are investigated. Arrival time of the incoming packets is recorded and windowed in order to obtain the network traffic data in terms of time series. In the statistical sense, the skewness and kurtosis features of time series data are chosen as dispersion parameters and classification features due to them having useful statistical information. After investigating the results with Shapiro-Wilk, Kolmogorov-Smirnoff and Chi-Square tests techniques; normal and attack traffics are modelled with different goodness-of-fit tests. Furthermore, Naive Bayes, k-nearest neighbor and thresholding methods are used in order to detect attacks from network traffic data. After heavy simulation studies, it has been shown that the detection rate of DDoS attacks depending on skewness feature is higher than the kurtosis. Naive Bayes outperforms with respect to the thresholding method by considering the classification performance. Consequently, the proposed method is a simple, low cost and highly accurate technique based on time series analysis while using different data analysis techniques which resulted in higher classification performance for attack detection.

ÖZET

SIZMA SAPTAMA SİSTEMİNDEKİ HİZMET ENGELLEME SALDIRILARININ ZAMAN SERİSİ ÇÖZÜMLEMESİ

Hizmet Engelleme (DoS) ve Dağıtılmış Hizmet Engelleme (DDoS) saldırıları şirketlere milyon dolarlık zarar vermektedir, dolayısıyla ağ güvenliği modern dünyadaki en önemli konu haline gelmiştir. DDoS saldırıları özellikle Veri Yararlanırlığı özelliğini ortadan kaldırmaktadır ki; Veri Gizliliği ve Veri Bütünlüğü ile birlikte bilgi güvenliği servislerinin en temel üç unsurunu oluşturmaktadır. Veri çözümleyicileri farklı yöntem ve araçları kullanarak DDoS saldırılarını önleme ya da korunmaya çalışmaktadırlar. Bu tezde, DDoS saldırılarının zaman serisindeki çözümlemesi araştırılmıştır. Ağ trafik verisini zaman serisi olarak elde etmek için giriş paketlerinin geliş zamanı kaydedilip, pencereleme fonksiyonu uygulanmıştır. İstatistiksel kullanılabilir bilgiler içerdiklerinden, dağılımın çarpıklığı ve savrukluğu; zaman serilerinin dağılımı parametresi ve sınıflandırma özneliği olarak seçilmiştir. Sonuçları Shapiro-Wilk, Kolmogorov-Smirnoff ve Chi-Square testleri ile inceledikten sonra; normal ve saldırı trafikleri çeşitli dağılıma uyum dereceleri ile modellenmiştir. Ek olarak Naive Bayes ve k en yakın komşu ayrıntılı sınıflandırma ve eşik değeri yöntemleri, ağ trafiği verisinden saldırıları tespit etmede kullanılmıştır. Ağır benzetim çalışmalarından sonra, DDoS saldırılarının sezim oranı göstermektedir ki; çarpıklık özneliği sonuçları savrukluk özneliği sonuçlarından daha yüksektir. Naive Bayes, ayrıntılı sınıflandırma performansı olarak eşik değeri yöntemine göre daha iyi sonuç vermektedir. Sonuç itibarıyla bu tezde, saldırı sezimi için yüksek sınıflandırma performansı ile sonuçlanan, farklı veri çözümlemesi kullanılarak, zaman serileri çözümlemesine dayalı basit, düşük maliyetli ve yüksek doğrulukta bir yöntem önerilmiştir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	v
LIST OF FIGURES	viii
LIST OF TABLES	xi
LIST OF SYMBOLS	xiii
LIST OF ACRONYMS/ABBREVIATIONS	xiv
1. INTRODUCTION (IDS STATE OF ART)	1
1.1. Computer Security and DDoS	1
1.2. Intrusion Detection System-State of Art	3
1.2.1. IDS Types-HIDS vs NIDS	4
1.2.2. IDS Approaches	4
1.2.2.1. Signature-based / Misuse IDS	4
1.2.2.2. Anomaly-based IDS	5
1.2.2.3. Hybrid-based IDS	5
1.3. Literature Review	5
2. TIME SERIES PARAMETERS ANALYSIS	8
2.1. Time Series Parameters	8
2.1.1. Mean	10
2.1.2. Mode	11
2.1.3. Median	11
2.1.4. Variance and Standard Deviation	11
2.2. Skewness and Kurtosis	13
2.2.1. Skewness	13
2.2.2. Kurtosis	14
3. HYPOTHESIS TESTING AND DISTRIBUTION FITTING	17
3.1. Hypothesis Testing	17
3.2. Visualization of Data	18
3.2.1. Histogram of Signals	18

3.2.2.	Quantile-Quantile Plot	19
3.2.3.	Shapiro-Wilk Test	21
3.3.	Distribution Fitting	22
4.	STATISTICAL TESTS AND TOOLS	24
4.1.	Goodness-of-fit (GOF) Tests	24
4.1.1.	Kolmogorov-Smirnov Test	25
4.1.2.	Chi-Square Test	29
4.2.	Classification Methods	32
4.2.1.	Naive Bayes Classification	32
4.2.2.	k-Nearest Neighbour	33
4.3.	Box Cox Transformation	35
4.4.	WEKA	39
4.4.1.	ARFF (Attribute Relation File Format)	39
4.4.2.	WEKA Explorer	40
5.	EXPERIMENTS AND RESULTS	42
5.1.	Statistical Results	42
5.1.1.	Overview	42
5.1.2.	Histogram of the Datasets	43
5.1.3.	Normality Tests	48
5.1.4.	Distribution Fitting Test	49
5.2.	Classification Results	52
5.2.1.	Naive Bayes Classification	52
5.2.2.	Analysis After Box Cox Transformation	56
5.2.3.	Thresholding Method	58
6.	CONCLUSION	66
	REFERENCES	67

LIST OF FIGURES

Figure 1.1.	DDoS attack elements	3
Figure 2.1.	Block diagram.	9
Figure 2.2.	Skewness example.	14
Figure 2.3.	Kurtosis example.	16
Figure 3.1.	Histogram example.	19
Figure 3.2.	QQ plot example.	20
Figure 4.1.	ECDF graph.	27
Figure 4.2.	Naive Bayes structure.	33
Figure 4.3.	k-Nearest Neighbour	34
Figure 4.4.	Signal before transformation	36
Figure 4.5.	Signal histogram before transformation	37
Figure 4.7.	Signal after transformation	37
Figure 4.6.	Signal QQ plot before transformation	37
Figure 4.8.	Signal histogram after transformation	38

Figure 4.10.	Signal log likelihood	38
Figure 4.9.	Signal QQ plot after transformation	38
Figure 4.11.	CAIDA dataset in attribute relation file format.	40
Figure 4.12.	WEKA gui	41
Figure 4.13.	WEKA explorer interface	41
Figure 5.1.	Normalized histogram of arrival packet numbers	43
Figure 5.2.	Scatterplot	45
Figure 5.3.	CAIDA skewness normal and attack.	46
Figure 5.4.	CAIDA kurtosis normal and attack.	47
Figure 5.5.	QQPlot CAIDA skewness attack.	48
Figure 5.6.	QQPlot CAIDA skewness normal.	48
Figure 5.7.	CAIDA normal skewness log logistic distribution.	50
Figure 5.8.	CAIDA attack skewness log logistic distribution.	50
Figure 5.9.	CAIDA normal kurtosis log logistic distribution.	51
Figure 5.10.	CAIDA attack kurtosis log logistic distribution.	51
Figure 5.11.	Confusion matrix.	53

Figure 5.12. NB classification results.	55
Figure 5.13. Attack skewness (a) Before Box Cox (b) After Box Cox.	56
Figure 5.14. Normal skewness (a) Before Box Cox (b) After Box Cox.	56
Figure 5.15. Attack kurtosis (a) Before Box Cox (b) After Box Cox.	57
Figure 5.16. Normal kurtosis (a) Before Box Cox (b) After Box Cox.	57
Figure 5.17. Skewness thresholding test.	59
Figure 5.18. Skewness thresholding test ROC.	60
Figure 5.19. Kurtosis thresholding test.	60
Figure 5.20. Kurtosis thresholding test ROC.	61
Figure 5.21. Combined ROC	62
Figure 5.22. Thresholding and NB classification results.	63
Figure 5.23. Bar plot NB classification results skewness BF and AF BXCX. . .	64
Figure 5.24. Bar plot NB classification results kurtosis BF and AF BXCX. . .	65
Figure 5.25. Bar plot NB classification results accuracy BF and AF BXCX. . .	65

LIST OF TABLES

Table 2.1.	Statistical parameters	10
Table 2.2.	Variance and standard deviation example	12
Table 4.1.	KS table.	29
Table 4.2.	Chi Square table.	31
Table 5.1.	Statistic measures of CAIDA	44
Table 5.2.	Scatterplot measures	45
Table 5.3.	Results of CAIDA skewness normal and attack	46
Table 5.4.	Results of CAIDA kurtosis normal and attack	47
Table 5.5.	Results of CAIDA and BOUN Shapiro Wilk test	49
Table 5.6.	Results of CAIDA attack skewness log logistic distribution	52
Table 5.7.	Results of CAIDA attack skewness log logistic distribution	52
Table 5.8.	Dataset	53
Table 5.9.	Naive bayes classification results	55
Table 5.10.	Thresholding skewness	58

Table 5.11.	Thresholding kurtosis	61
Table 5.12.	Thresholding and NB classification results after Box Cox	62
Table 5.13.	Thresholding and NB classification results before and after Box Cox	64

LIST OF SYMBOLS

a	Shapiro-Wilk test constant
cv	Critical value
C	Class
D	Kolmogorov-Smirnoff test statistic
F	Theoretical cumulative distribution function
H_A	Alternate hypothesis
H_0	Null hypothesis
H_1	Alternate hypothesis
k	Maximum difference between two cumulative distribution functions
N	Sample size
p	Precision value
std	Standard deviation
W	Shapiro-Wilk test statistic
λ	Box Cox transformation exponent
$\bar{x}$	Mean of a sample
μ	Mean of a population
α	Significance level
σ	Standard deviation
σ^2	Variance

LIST OF ACRONYMS/ABBREVIATIONS

AF	After
AUC	Area Under Curve
ARFF	Attribute Relation File Format
BF	Before
BXCX	Box Cox
CAIDA	Center for Applied Internet Data Analysis
CDF	Cumulative Distribution Function
DDoS	Distributed Denial of Service
DoS	Denial of Service
ECDF	Empirical Cumulative Distribution Function
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GUI	Graphical User Interface
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
kNN	k-Nearest Neighbour
KS	Kolmogorov-Smirnov
QQ	Quantile Quantile
PDF	Probability Density Function
ROC	Receiver Operating Characteristic
TCP	Transfer Control Protocol
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WEKA	Waikato Environment for Knowledge Analysis
Var	Variance

1. INTRODUCTION (IDS STATE OF ART)

Cybercrime has become a part of our life since the internet's and technology's massive evolution. Attackers try to give damage to companies and countries for different purposes. According to Symantec Internet Security Threat Report (ISTR), 430 million new unique pieces of malware appeared in 2015, increased up to 36% from the previous year [1].

Attacks are generally; smartphone threats, targeted attacks, Internet of Things (IoT) and social media scams. Distributed Denial of Service Attack (DDoS) is a very popular attack type and it is proved to be disruptive against businesses in the last couple of years. DDoS attacks employ compromised computers as agents for attacking the victim. These attacks are very effective, their set is simple and it is hard to stop. In order to make tracing back as difficult as possible source IP addresses are spoofed.

The main threat of DDoS attacks is against the availability of the system. Symantec surveys show that ordering a DDoS attack generally costs between US\$10 to US\$1,000 per day. Some of these attacks may even cost to victim organizations around US\$40,000 per hour [1].

1.1. Computer Security and DDoS

Computer security is a technological process for protecting systems against threats. Availability, integrity and confidentiality are the three major security services for computing systems. Availability means accessing the information or resources whenever it is deemed necessary. It guarantees that information is not corrupted, monopolised or destructed. Integrity is the reliability of information or resources. It ensures that the information is not changed or modified by unauthorised parties. Confidentiality is defined as keeping the information or resources private to its users. It guarantees that the information is not accessed, copied or seen by the unauthorised parties.

DDoS attacks generally hit the availability service of security systems. In this type of DDoS attacks, attacker tries to stop authorised traffic between clients and a server using spurious packets and large number of compromised agents [2]. Source IP address is also spoofed in order to conceal the identity of the attacker during the attack which makes it difficult to trace back.

DoS attacks can damage web-based applications, networks and services by flooding with excessive data or destroying them in different ways. The Denial-of-service (DoS) attacks are generally started by one computer or one connection, whereas for a DDoS attack, multiple sources are involved in the attacking process.

Some of the DDoS attacks are listed as follows;

- (i) UDP flood
- (ii) Peer-to-peer attacks
- (iii) SYN flood
- (iv) Reflected attack
- (v) Ping of death
- (vi) Slowloris
- (vii) Nuke
- (viii) Multi-vector attacks
- (ix) Degredation of service attacks
- (x) Zero Day DDoS
- (xi) Unintentional DDoS and Application level attacks

DDoS attacks consist of four elements [3];

- (i) Attacker: It controls the attack. All handlers are organized by the attacker.
- (ii) Handlers or Masters: Handlers controls the agents in the system. Handlers are the compromised hosts and attackers can run their attack program on handlers.
- (iii) Zombie hosts or Attack daemon agents or Slaves: Zombies are responsible for producing a flood of packets against the predetermined victim.

(iv) Victim: Target host.

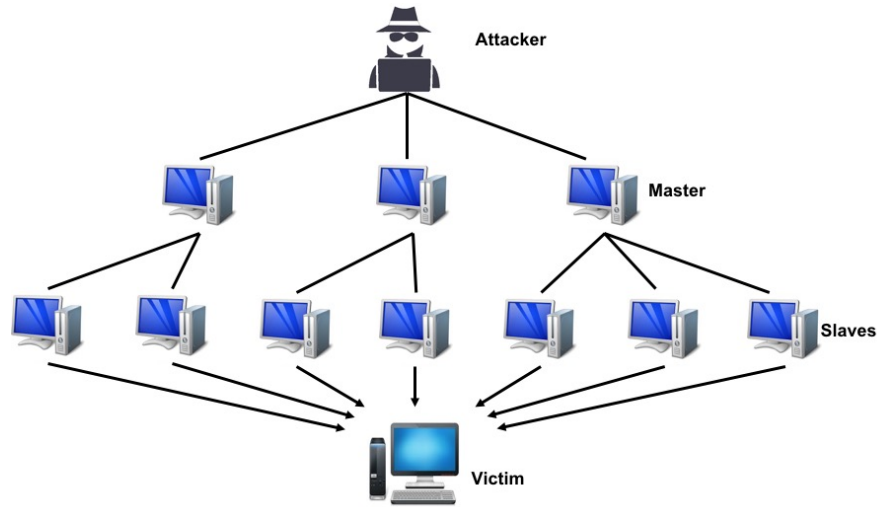


Figure 1.1. DDoS attack elements [4].

1.2. Intrusion Detection System-State of Art

Network security systems try to protect information systems against unknown vulnerabilities. There are two types of systems which are considered for handling intrusions; first one is Intrusion Detection Systems (IDS) and the second one is Intrusion Prevention Systems (IPS). An intrusion is defined as “ a malicious, externally-induced fault resulting from a successful attack ” [5]. General information and usage of IDS; packets are monitored via ports etc. and traffic is compared according to specified rules. If a threat is detected IDS sets a warning or an alarm. The main purpose in the detection process is generally to maximize the accuracy and recognize the real threats while minimizing the false positives. IPS can stop malicious attacks before invading the user. It observes traffic flow on network and terminates connections via blocking the user account and IP address of the attacker. Sometimes IPS can use other security controls like firewall or router for blocking the attacks. Our main purpose is on the Intrusion Detection Systems in order to detect attacks on the system.

Most organizations' main purpose is to secure their information systems from network hazards. Security systems can be violated by unidentified exposures and new hacks. It is classified as any activity of the user when he is not allowed to. An intrusion attack is described as the intruder tries to gain control of the system by doing repetitive actions. IDS's purpose is to detect and decide if an intrusion has been taken place or not. After detecting the intrusion, system should respond to this malicious activity. Currently there are two levels of combat against intrusions of cyber-attacks, these are the host level and the network level intrusion detection systems.

1.2.1. IDS Types-HIDS vs NIDS

The network-based IDSs (NIDSs) monitor the specifications of the network connection informations to differentiate normal and attack traffic. Systems monitor the network, analyze the traffic and report any irregularity to the admin. Whereas host-based IDSs (HIDSs) uses specific detection methods and locates the attacks at host level by monitoring the condition of workstations. If suspicious traffic is detected, system warns the system admin. There are also different IDS depending on; feature evaluation in the system (network data or log files data) [6], distribution mode in the analysis process (distributed, centralized) [7] and the time of analysis (offline or on-line) [8]. Anomaly-based, signature-based and hybrid are the three main classes of IDS [9].

1.2.2. IDS Approaches

1.2.2.1. Signature-based / Misuse IDS. Signature-based system has all the previous attack traffic model and specifications; system compares the new traffic with previously learned model. Previously known attack patterns and events under test similarities are compared. If the new traffic specifications match with the previously learned traffic; it is classified as an attack. Snort is a commonly used misuse IDSs example [10]. Snort has different capabilities. It does traffic analysis in real time, searches and matches the content by using previously identified attack signatures to discover attacks. This system is useful for capturing the known attacks whereas unknown attacks can not be

identified. This model is weak to classify zero-day attacks. New attacks which are not in the database can be classified as normal [11].

1.2.2.2. Anomaly-based IDS. Anomaly-based IDS models the normal traffic and learns its properties. If the incoming traffic is different than what has been learnt as normal; it is classified as an attack. Anomaly-based IDSs' main advantage is, its ability to detect new attacks which do not have existing signatures. Normally, incoming traffic that is different from normal does not have to be an attack in some cases. Sometimes model can be misleading to classify normal traffic as attack. That is why this method suffers high false positive rate and has high operational costs [12].

1.2.2.3. Hybrid-based IDS. Hybrid-based IDS model is the combination of Signature-based IDS and Anomaly-based IDS. Anomaly detection is responsible for detecting anomalous behaviours based on the trained normal behaviour model whereas Misuse detection is responsible for detecting previously defined attacks based on their attack signatures [13]. Computational load of both signature matching and anomaly detection to evaluate the incoming traffic is the main drawback of this combined system.

1.3. Literature Review

Examination of payload and analysis at the packet level are the two major methods implemented in most intrusion detection systems [2]. Traditional methods rely on packet level scrutinizing that is why flow level analysis might be an alternative. IDS examines the arrived packets carefully in order to find any unusual traffic activity. Generally the new DDoS attacks mimic the legitimate web traffic and attack's payload characteristics are very similar to normal packets, which leave traditional methods ineffective [14]. That is why statistical measures of time series are used to find abnormality.

In [15], authors proposed a time domain analysis model for anomaly detection technique. Number of packets in time domain are transformed into complex gaussian

coefficients to set a threshold for detecting flooding attacks which are unknown.

In [16], authors proposed both time and frequency domain analysis of network-wide traffic by having a network traffic matrix and applying K-L transformation. After the transformation correlation coefficients are computed, Hilbert transform is used for frequency domain analysis. For periodic attacks, time domain and for non-periodic noisy data attacks, frequency domain resulted in better detection rate. As for computational complexity, time domain analysis is simpler and faster than frequency domain analysis.

In [17], authors proposed a detection approach of Stealthy DDoS and Slowly Increasing-Intensity DDoS Attacks (SIDA) based on time-series decomposition. It consists of random and trend components. Double autocorrelation and improved cumulative sum techniques are applied to the random and trend components. Flow connection entropy series is adopted to decrease dimensionality of high dimensional traffic series.

In [18], authors make an empirical study on forecasting accuracy. After understanding the nature of the data; attention is given to the rule induction for forecasting method selection. A new approach has been proposed for selecting a forecasting method for univariate time series. Combination of data-mining, meta-learning, clustering, classification and statistical measurement are used to find out the characteristics of the data. Trend, periodicity, seasonality, skewness, serial correlation, non linearity, kurtosis, chaos and self-similarity have been investigated. The investigation of time-series analysis in this work helped our study.

In [3], authors compared the Flood DDoS type attack and Low-Rate DDoS type attack in NS-2 simulation platform for time and frequency analysis. Attack generation principles, mechanism utilization, behaviour, signature and attack performance are studied. It is found that FDDoS is easier to detect than LDDoS attacks.

In [19], authors proposed a spectral analysis method to identify normal TCP traffic. Arriving packet numbers of a flow in fixed-length time intervals is used for spectral analysis. Power spectral density (PSD) of the signal which has the periodicity information is estimated. In both flow directions, TCP has the strong periodicity around its round-trip time. Periodicity is not shown for attack flow. Consequently this approach reduces the false positives of attacker identification pattern. Legitimate traffic will not have unnecessary delay or damage. Multi-sourced DDoS attacks PSD shows that; data distribution is in the lower frequency when compared to single-sourced DoS attacks. [20, 21].

As a classification feature, entropy is commonly used in DDoS attack detection. In [22] authors suggested a work method detecting DDoS attacks in the network using entropy based anomaly detection algorithm. They calculated the entropy value with respect to time and packet windows. It is found out that the entropy values lie in a narrow range in case of normal traffic whereas increases or decreases considerably in the case of attack.

In [23], authors analyzed an entropy-based approach to detect DDoS attacks. Effect on the entropy of all useful packet attributes during DDoS attack are analyzed and their usefulness were tested with different types of DDoS attacks. It was shown that, packet attribute's entropy values can deviate during the attack so; for better thresholding results, proper choice of attribute selection is also explained.

In this thesis, we extend the previous studies by introducing a new time series analysis method and give detailed analysis by using Skewness, Kurtosis as features for network traffic and Goodness-of-fit tests for data analysis. Box Cox transformation is applied for statistical improvement. As dataset; CAIDA 2007 and CAIDA 2008 are used for attack and normal traffic respectively. MATLAB R2016a and WEKA 3.6 is used for analysis and experiments. The rest of this thesis is organized as: chapter 2 is devoted to Time Series Parameters Analysis. Hypothesis Testing and Distribution Fitting is given in chapter 3. Chapter 4 belongs to Statistical Tests and Tools. Chapter 5 presents the Experiments and Results and the Conclusion is given in chapter 6.

2. TIME SERIES PARAMETERS ANALYSIS

CAIDA dataset which is evaluated in this thesis is an offline dataset that is why passive monitoring is used. CAIDA has traffic traces which shows timestamps of each arriving packet. Total number of packets are counted in every 1 ms of the timestamps. Thus a time series analysis of DDoS attacks was investigated by sampling the incoming traffic in each 1 ms. The obtained time series were further divided into 128-length window to create the dataset. For each row in the dataset the skewness and kurtosis parameters are found, which are chosen as basic features for network traffic because of the reason that they have useful information. Histogram and Quantile-Quantile plot are used for graphical evaluation and Shapiro-Wilk test is used for numerical evaluation in order to find if this new dataset has normal distribution or not. It has been shown that data do not have a normal distribution so, different distribution fitting methods (Logistic, Log logistic, Log normal, Weibull etc.) are used as goodness-of-fit tests and evaluated by Kolmogorov-Smirnoff and Chi-Square tests. It has been confirmed that the data have Log logistic distribution. As the next phase, Box Cox transformation is applied for improving statistical results. This new transformed data is used for thresholding, kNN and Naive Bayes classification performance in order to detect attack from normal traffic. Block diagram of our thesis is stated in Figure 2.1.

2.1. Time Series Parameters

In Section 2.1; mean, mode, median, variance, standard deviation and in Section 2.2, skewness and kurtosis parameters which are commonly used in time series analysis will be investigated.

In statistics there are two terms which have real importance for understanding the basic concepts. These are sample and population. A sample is a subset of group, items or events from a larger population. By looking the sample, information about the population can be learnt. Estimation on the population is quite accurate if the sample is random and large enough.

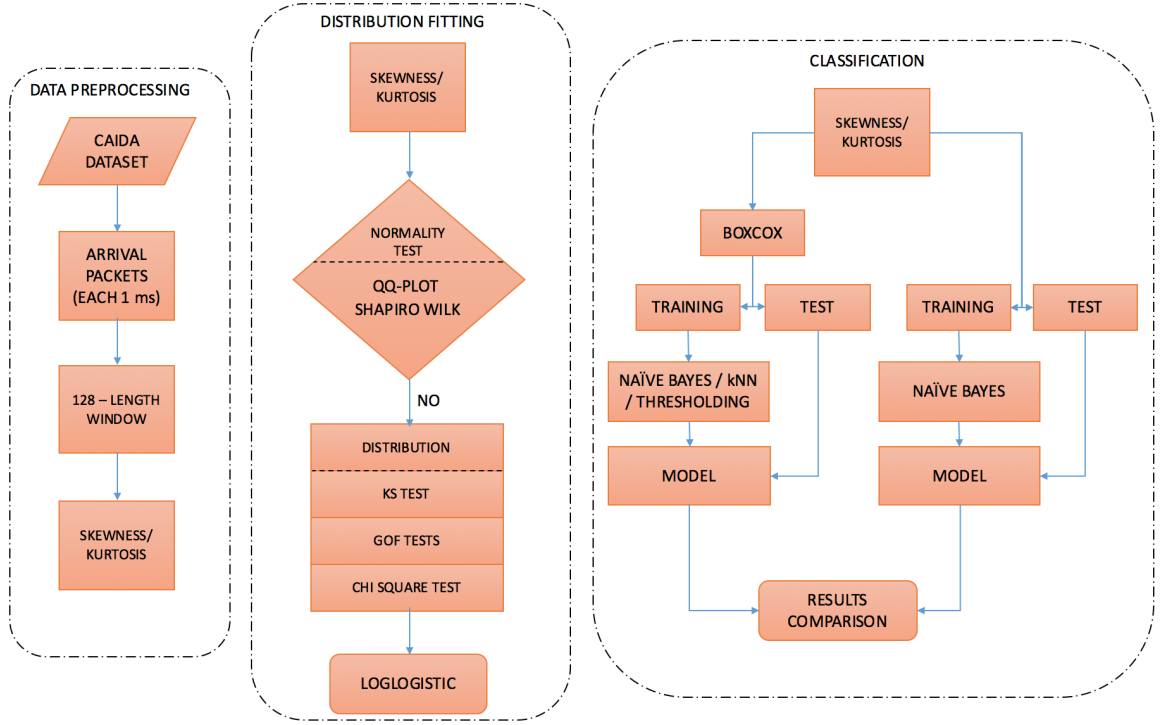


Figure 2.1. Block diagram.

Parameters are descriptive measures of an entire population. In order to produce distribution curves, parameters are used to determine probability density function (PDF). In order to differentiate parameters from sample statistics, parameters are usually represented by Greek letters. Greek letter mu (μ) represents population mean and Greek letter sigma (σ) represents the population standard deviation. Parameters do not change like variables, they are constant. It is not practical to calculate the whole population because of the reason that parameter's values are generally unknown.

Several parameters are used to represent distributions. Three distributions' parameters are given as an example in the Table 2.1. Values help to decide the location and shape of the curve on distributions. If combination of the parameters are used, they will result in different distribution curves.

Table 2.1. Parameter table.

Distribution	Parameter 1	Parameter 2	Parameter 3
Normal	Mean	Standard Deviation	
3 - Parameter Gamma	Shape	Scale	Threshold
Chi - Square	Degrees of Freedom		

For instance; standard normal distribution is defined by mean and standard deviation. If these two parameters are specified, the entire distribution is precisely known.

2.1.1. Mean

Mean is the average of all observed outcomes. It is calculated as; summation of all observed values are divided by the total number of events. It is represented as μ . In mathematical terms;

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (2.1)$$

where N is sample size and x is the observed value.

For instance, suppose that we have randomly distributed six observations like; 36, 45, 83, 108, 108 and 118. According to the mean formula; $\mu = 83$.

As it is seen from Equation 2.1, mean of the six observation is 83. In statistics, there are two types of means; sample mean and population mean. Sample is a subset of population which is represented by $\bar{x}$ and μ is used for population mean.

2.1.2. Mode

Mode is the number which has the highest frequency in a set of data. In the example given above; 108 is the mode, because it is repeated two times whereas all the remaining numbers are repeated once.

2.1.3. Median

Median is basically the middle score after putting the data in ascending or descending order. Above example in ascending order is stated as 36, 45, 83, 108, 108, 118. Because the dataset has even number of samples, median will be the average of 83 and 108 which results in 95.5. If the data has odd number of samples, the median would be the middle number in the sequence.

Median is very useful when the data have outliers. When the data have outliers, its mean is affected a lot if the outliers are too far from the sample. In that case the median gives better information about the data distribution.

2.1.4. Variance and Standard Deviation

Variance and standard deviation gives information about how spread out the data is. Standard deviation will be small if all the data is close to the mean. If data is spread out a lot and have large outlier values from the mean, standard deviation will be large. Consequently having outliers will increase the standard deviation. Variance is defined in Equation 2.2.

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2 \quad (2.2)$$

Standard deviation is defined in Equation 2.3.

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2} \quad (2.3)$$

From the example above; calculated mean is 83.

Table 2.2. Variance and standard deviation calculation

x	μ	$(x - \mu)$	$(x - \mu)^2$
36	83	-47	2209
45	83	-38	1444
83	83	0	0
108	83	25	625
108	83	25	625
118	83	35	1225
			6128

As it is seen from the Table 2.2, calculated mean 83 is subtracted from the samples and squared the difference. According to variance formula 6128 is divided by 5 (which is $N - 1$) and result is 1225,6. This number is the variance.

When it is squared root, result will be approximately 35. 35 is the standard deviation which shows how spread out the data is from the mean. Namely, $83 - 35 = 48$ and $83 + 35 = 118$; most of the data is spread out between 48 and 118.

In normal distribution, 68% of the data is inside 1 *std*, 95.4% of the data is inside 2 *std* and 99.7% of the data is inside 3 *std* area.

2.2. Skewness and Kurtosis

In this section we will give brief information about our selected features for classification.

2.2.1. Skewness

Skewness and Kurtosis are two numerical measures that give more precise information on the shape of a distribution. Mean of x is the measure of center point of a distribution. Variance is the second moment of x about the mean and gives information about how spread the data is around the mean. Third moment about the mean of x gives information about the Skewness of the data and lastly fourth moment about the mean of x which gives information about the Kurtosis values.

Skewness is a degree of symmetry, literally lack of symmetry. If a distribution has equal values around the center (left and right), it can be concluded that the distribution is symmetric. Skewness formulation is given in Equation 2.4;

$$Skewness = \frac{1}{N\sigma^3} \sum_{i=1}^N (x_i - \mu)^3 \quad (2.4)$$

where σ is standard deviation.

After the calculation; given example distribution's skewness value is -0.4318 . If the skewness result is a positive number, data is positively skewed and if it is negative, data is negatively skewed. In statistical aspect it means; if the data is skewed to the right that means right tail is longer than left tail (positively skewed). Longer tail has more information about the dataset than the shorter tail. Basic interpretation of skewness is given below;

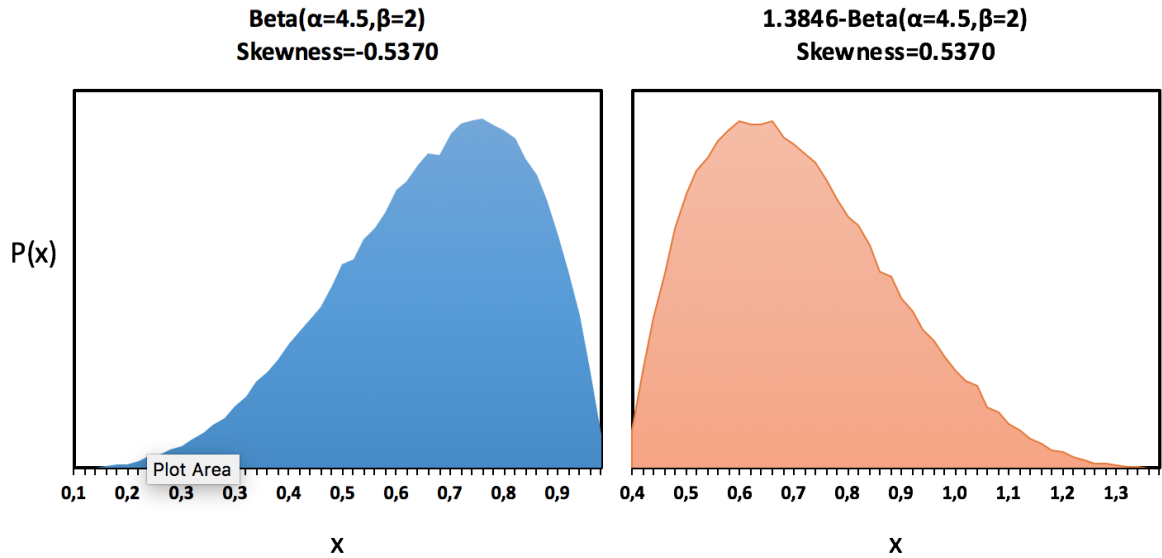


Figure 2.2. Skewness example.

- (i) Highly skewed distribution if: $\text{skewness} < -1$ or $\text{skewness} > +1$,
- (ii) Moderately skewed distribution if: $-1 < \text{skewness} < -1/2$ or $+1/2 < \text{skewness} < +1$,
- (iii) Approximately symmetric if $-1/2 < \text{skewness} < +1/2$.

Positively skewed data means that; $\text{Mean} > \text{Median} > \text{Mode}$ whereas in negatively skewed data's $\text{Mean} < \text{Median} < \text{Mode}$.

In Figure 2.2, graph on the left has negative skewness or skewed to the left, left tail of the distribution is longer; whereas graph on the right has positive skewness and skewed to the right, right tail of the distribution is longer.

2.2.2. Kurtosis

Kurtosis is a degree that whether the data is heavy tailed or light tailed compared to the normal distribution. It demonstrates how sharp and tall the central peak is, compared to standard distribution. High kurtosis represents heavy tails or outliers whereas low kurtosis represents light tails and no outliers. Histogram can be used to visualize the distribution.

Kurtosis formulation is given in Equation 2.5;

$$Kurtosis = \frac{1}{N\sigma^4} \sum_{i=1}^N (x_i - \mu)^4 \quad (2.5)$$

Depending on the graphical interpretation of the histogram; if it is flatter distribution than normal, it means negative kurtosis. If distribution has more peaks than gaussian distribution, it can be said that data has positive kurtosis. Gaussian distribution kurtosis value is 3.

$$Excess \ Kurtosis = Kurtosis - 3 \quad (2.6)$$

When number 3 is subtracted from Kurtosis, it gives the result of Excess Kurtosis. By looking the Excess Kurtosis result, we can directly compare our data against the standard normal distribution.

- (i) If kurtosis ≈ 3 and excess kurtosis is ≈ 0 it is called Mesokurtic,
- (ii) If kurtosis > 3 and excess kurtosis is > 0 it is called Leptokurtic; fatter and longer tails, sharper and higher central peak compared to standard normal distribution. More probability than expected in the tails of the distribution due to extreme values away from the mean. Probability values are pushed away from the mean towards the tails.
- (iii) If kurtosis < 3 and excess kurtosis is < 0 it is called Platykurtic; shorter and thinner tails, broader and lower central peak compared to standard normal distribution.

In Figure 2.3, all three distributions'; mean is zero, standard deviation is 1 and skewness value is zero. They are plotted on the same vertical and horizontal scale. From left to right, kurtosis increases. Moving from uniform to normal distribution, the intermediate values in the distribution have become less likely whereas central and extreme values have become more likely. Moving from normal distribution to logistic

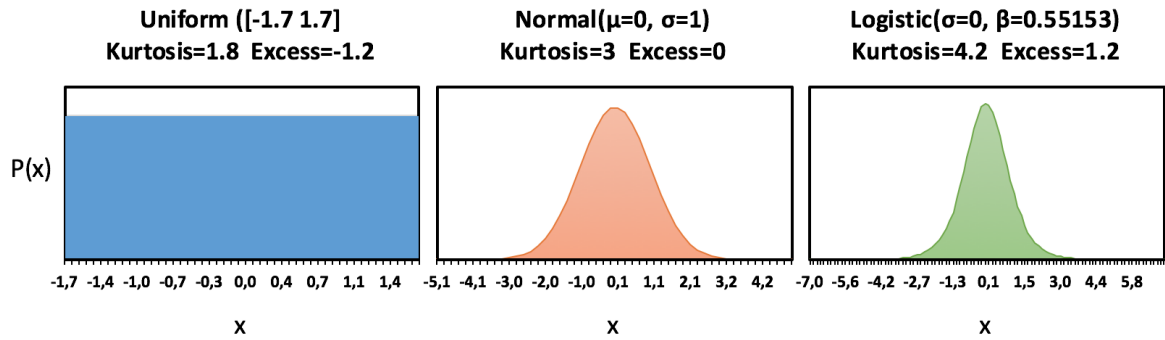


Figure 2.3. Kurtosis example.

distribution, there are more data in the tails and less data in shoulders. Central peak is higher and narrower.

Basically by looking to the histogram of skewness and kurtosis of a dataset, statisticians can get an idea about the shape of the distribution and they can conclude if the data is normally distributed or not. Depending on the result some transformation can then be applied to make the data similar to normal distribution (e.g. Box Cox transformation) for statistical improvement.

3. HYPOTHESIS TESTING AND DISTRIBUTION FITTING

3.1. Hypothesis Testing

Hypothesis testing is making decision with available resources on a specific event. First an initial assumption is being made, then the evidence is collected and finally decision is given whether to accept or reject the initial assumption based on the evidence that was collected.

There are basically two methods for Hypothesis testing; p-value approach and critical value approach. A hypothesis test is a kind of statistical test which determines whether sufficient evidence is available in a sample of data in order to figure out that a certain condition is true for the overall population. Hypothesis can be grouped into two; null hypothesis (H_o) and alternate hypothesis (H_1 or H_a). Tested statement, “No effect” or “No difference” statement is the null hypothesis generally.

Based on the sample, test rejects or accepts the null hypothesis. In order to make a decision; p-value is compared to the level of significance (threshold). Statistical significance means that the sample statistics are most likely not due to chance. If p-value is greater than the cut-off point, test accepts the null hypothesis otherwise test rejects the null hypothesis. p-value is a measure of the evidence against H_o , the smaller the p-value, the stronger the evidence against H_o [24].

In hypothesis testing, test statistic is calculated from the sample data. After the calculation, it is compared with the critical value under the null hypothesis to decide whether to reject or accept the null hypothesis. Critical values depend on the significance level α which selected by the statisticians before starting the test. Statistical significance increases beyond the test statistics in case the sample data is against the assumptions for null hypothesis and becomes statistically significant.

It means test statistic is greater than critical value. p-value calculation is done by using the test statistic. Statistically significance causes small p-value to reject the null hypothesis.

P-value gives information about the results whether they are statistically significant or not. They are used in hypothesis testing in order to reject or fail to reject the null hypothesis. Value range is between 0 and 1. p-value can be compared to level of significance α in order to decide whether the null hypothesis (H_0) should be rejected or not. It is a kind of probability which measures the evidence against the null hypothesis. Lower probabilities have stronger proof against the null hypothesis. If the calculated p-value is smaller than α , reject H_0 ; otherwise we fail to reject the null hypothesis. Generally 0.05 is used as level of significance α value. Thus, if the p-value is less than or equal to 0.05, null hypothesis H_0 is rejected.

3.2. Visualization of Data

3.2.1. Histogram of Signals

A histogram is a kind of plotting technique and graphical display which helps to show the frequency distributions of a continuous data. It shows the general distribution, outliers, skewness and kurtosis etc. of the data. Before getting into the details of the numbers and results of the data, it gives an overview and idea to the data analysts. Vertical axis represents independent variable whereas horizontal axis represents dependent variable.

- (i) y axis represents the frequency; it shows how many times each score counts in that specific range. Highest bin shows the mode of the signal.
- (ii) x axis represents the bins of continuous data. Depending on the statistics type, bin number can be set manually or automatically.

In Figure 3.1, there are 306 values changing between 0.01 and 6.3580. These numbers are chunked into the bins changing between 0 and 7. There are 53 values changing

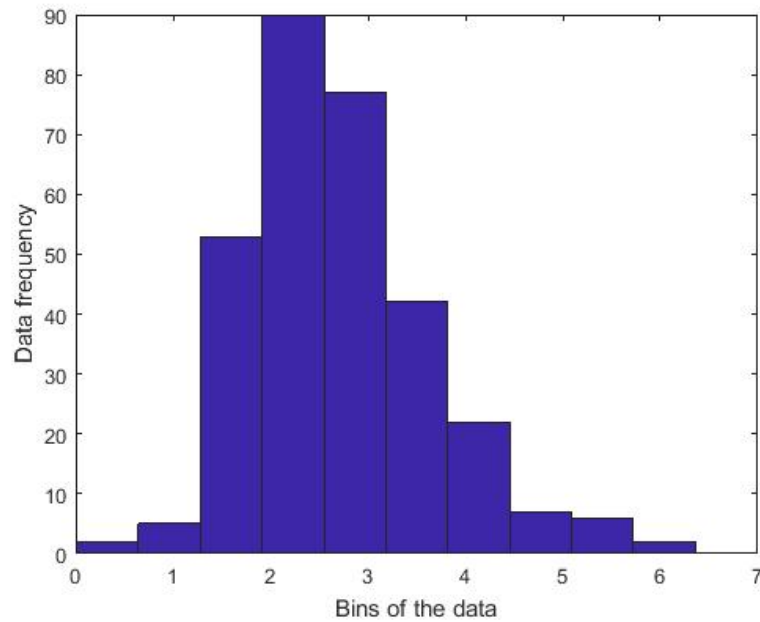


Figure 3.1. Histogram example.

between 1.28 and 1.91 in the third bin from the left.

3.2.2. Quantile-Quantile Plot

QQ plot is one of the normality tests in statistics which gives graphical information about the data. They are created by plotting the sample order statistics of a dataset against the expected quantiles of a normal distribution [25]. This technique is one of the graphical representation techniques for finding if a dataset belongs to the some specified distribution or not. QQ plot can also be used to compare two different dataset by plotting the quantiles of the first dataset against the quantiles of the second dataset. Quantile is the percentage of the points below the given value. For instance, 0.2 or 20% quantile is the point where 20% of the data is below and 80% is above that value.

45-degree reference line is used for decision making. Points fall along the 45-degree line, if the two datasets belong to a population with the same distribution.

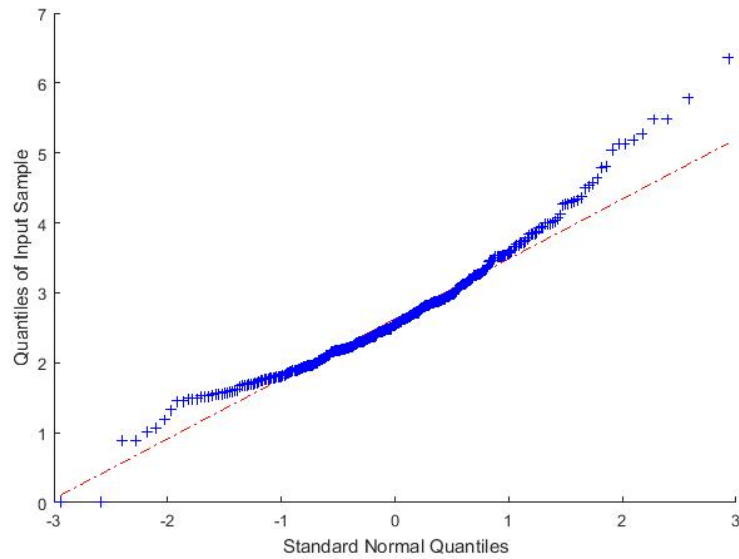


Figure 3.2. QQ plot example.

If the distribution is above or below the the reference line, it means that compared datasets have different distributions.

Here is the list of advantages of using QQ-Plot;

- (i) Sample size does not have to be equal. Quantiles are chosen according to the smaller data, larger dataset is interpolated.
- (ii) Statistical aspects like location shifts, scale shifts, symmetry changes and the outlier presence can be detected.

In Figure 3.2, there are 306 sample values which are between 0.01 and 6.358. Vertical axis represents estimated quantiles from the given dataset and horizontal axis represents theoretical quantiles from the standard normal distribution. Blue data points are close to the red line in the center, whereas distance from the red line increases for the smaller and larger values which shows that they have different distribution than normal. Apart from graphical normality test like QQ plot, there are also some other statistical normality tests like Kolmogorov-Smirnoff and Shapiro-Wilk which are going to be the next topic.

3.2.3. Shapiro-Wilk Test

Shapiro-Wilk test is one of the normality tests that is used in advance statistics. It calculates W statistic which tests a random sample $x_1, x_2, \dots, x_N$ belongs to a normal distribution or not. Formulation for W statistic is given in Equation 3.1.

$$W = \frac{\left(\sum_{i=1}^N a_i x_{(i)} \right)^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (3.1)$$

where $x_{(i)}$'s are ordered sample values and a_i 's are constants which are calculated from the variances and covariances of the order statistics of a sample of size N from a normal distribution.

In Equation 3.1, the nominator is the square of an appropriate linear combination of the sample order statistics where the denominator is an estimate of variance [26]. Under normality hypothesis, squared slope of the probability plot regression line is an estimate of the population variance multiplied by a constant. It can be said that it is the slope of observed data vs. expected normal values, normalized to a constant number.

Shapiro-Wilk tests' null hypothesis is that; the sample is drawn from a normally distributed population. W values change between 0 and 1. Small values of W means that samples may be significantly different from normal distribution [27]. Nominator is an estimate of σ^2 and the denominator is an estimate of σ^2 . If the dataset is normally distributed, W becomes 1 because both the nominator and the denominator will be the same. Generally univariate, continuous data is used. Some important notes regarding to Shapiro-Wilk test is given below;

- (i) There may be some problems if the data has too many equal values.
- (ii) For small N values test rarely reject H_0 .

- (iii) Test rejects the null hypothesis for tiny differences from normality when N is large.

Shapiro-Wilk test and QQ plot are most commonly used normality tests in statistics. They can be both applied to a data to cross check the obtained results.

3.3. Distribution Fitting

Dataset created by random process is compared with known statistical distributions in order to find out the obtained dataset properties. By looking to the fitted distribution, it can be concluded that what type of distribution suits or represents this introduced new random data.

In order to make inferences and deal with the uncertainty problems in a mathematical aspect, probability distributions are used. Distributions are used to perform calculations and results are applied to business decisions. Using proper tools helps the statistician to get accurate results. Results and decisions will be incorrect if the chosen distribution does not fit the data properly.

Distribution fitting helps to improve random process models. Normal distribution is the oldest distribution which is generally used in statistics. It is symmetric about its mean, unbounded on the entire real axis ($-\infty$ to $+\infty$) and its shape is constant. Normal distribution has location and scale parameters which are represented by mean and the variance respectively and it does not have the shape parameter.

Generally more than one distribution may fit to a dataset. After results are compared, most valid one is selected. By looking the histogram and QQ-plot (graphical representation of the data), general idea about the data is obtained and related distribution can be selected accordingly after statistical calculations.

Depending on the sample, statistical methods are used to estimate distribution parameters. Special software programmes can be used for this purpose. It is important

to find out how the distribution is fitted to the data. Some specific goodness-of-fit tests are available for distribution fitting. Distribution accuracy can be found out after comparing the empirical and theoretical distributions. It is basically the comparison of sample data and the fitted data. Depending on the test results; the model which describes the data better will be chosen. There are different kinds of distribution types, commonly used ones and their area of use are given below.

- (i) Normal Distribution; used in biology to measure the size of living tissue (length, height, weight.). Also it is used for measurement errors in physical experiments and sometimes in finance and hydrology.
- (ii) Exponential Distribution; used in queuing theory, reliability engineering, physics and hydrology.
- (iii) Logistic Distribution; used in chess ratings and in hydrology for distribution of long duration river discharge and rainfall.
- (iv) Gamma Distribution; used in wireless communication to model multi-path fading of signal power; in neuroscience to describe the distribution of inter-spike intervals and also for modeling the size of insurance claims.
- (v) Log logistic Distribution; used in survival analysis, networking, hydrology for modeling precipitation and economics to model distribution of wealth or income.
- (vi) Inverse Gaussian Distribution; used in numeric computation and software like R.
- (vii) Log normal Distribution; used in human behavior analysis on the online internet articles and discussion forums, biology and medicine, social sciences and demographics and in reliability analysis.
- (viii) Poisson Distribution; used in telecommunications for telephone calls arriving in a system, astronomy for photons arriving at a telescope, biology and management for customers arriving at a call center.
- (ix) Rayleigh Distribution; used in magnetic resonance imaging (MRI).
- (x) Rician Distribution; used in target shooting for effect of sighting.
- (xi) Weibull Distribution; generally used in reliability engineering, survival analysis and weather forecasting for wind speed.

4. STATISTICAL TESTS AND TOOLS

4.1. Goodness-of-fit (GOF) Tests

Sample data properties are obtained and compared against the population in order to assign the sample to a correct or approximate population. Goodness-of-fit tests are used for this purpose basically. Kolmogorov-Smirnov, Chi-Square, Shapiro-Wilk, Cramer-von-Mises and Anderson-Darling are some of the commonly used examples of goodness-of-fit tests [28].

In order to find if a sample comes from a population with specific properties or not, parametric and nonparametric tests are used. Parametric tests assume that data follow a specific distribution, generally normal distribution. Whereas in nonparametric tests data does not have to follow any distributions, they are generally known as distribution free tests. Generally t-tests, f-tests, ANOVA are the examples of parametric tests where as Kolmogorov-Smirnov (KS), Chi-Square tests are the examples of non-parametric tests. Information about the KS test and Chi-Square tests will be given in the following sections. Brief information on some of the parametric tests are given below as an example.

T-tests are used when the population parameter is not known. It is used to estimate population parameter (mean) and finding the differences between two groups' mean. One sample t-test tests if sample comes from a population with a particular mean or median whereas two sample t-test tests two or more sample sets have same location parameters. z-tests are used when the population parameter is known. Test assumptions; x and y should be random samples from two independent populations and populations are approximately normal. Sample data can estimate population variances and population variances are roughly equal.

F-tests are used for finding if there is any variance or not within samples. This test is basis for ANOVA. f statistics is never negative and it is non-directional. F

distribution is positively skewed whereas z-test and t-test distribution is symmetrical.

ANOVA is the abbreviation of Analysis of Variance. It is used when more than two group mean is compared. t-test is a special kind of ANOVA where two group means are compared.

Obtained dataset is not normally distributed, that is why Kolmogorov-Smirnoff and Chi-Square non-parametric tests are used.

4.1.1. Kolmogorov-Smirnov Test

Kolmogorov-Smirnov (KS) test is a non parametric test which is used to decide if a sample belongs to a population with specific distribution or not [29]. One sample Kolmogorov-Smirnov (KS test) tests if sample comes from a population with a particular distribution or not whereas two-sample KS test checks if two sample sets have the same distribution or not [30]. This test is based on empirical cumulative distribution function (ECDF).

Formulation is given in Equation 4.1;

N ordered data points $X_1, X_2, \dots, X_N$ and empirical cumulative distribution function is defined as,

$$E_N = \frac{n(j)}{N} \quad (4.1)$$

where $n(j)$ is the number of points less than X_j . X_j are ordered from smallest to largest.

KS test formulation is given in Equation 4.2;

$$D = \max_{1 \leq j \leq N} \left(F(X_j) - \frac{j-1}{N}, \frac{j}{N} - F(X_j) \right) \quad (4.2)$$

where D is the test statistic and F is theoretical CDF. F must be continuous.

Hypothesis in KS test are stated below;

- (i) H_0 : Two samples drawn from populations with same distribution.
- (ii) H_1 : Two samples are not from the same distribution.

Critical value is determined from the KS table which is calculated and published for specific values. In order to do the hypothesis testing; critical value of D is compared against the test statistic D . D value is the maximum absolute difference between the two CDFs. If two samples have the same distributions D value should be zero in theory. Samples are taken randomly from the population that is why D value is not zero even the two distributions are the same. ECDF plot is a graph for evaluating the fit of a distribution to the data, estimating percentiles and comparing different sample distributions. Once the test statistic is calculated, it is compared against the critical value and the decision to reject the Null Hypothesis or fail to Reject Null Hypothesis is made depending on the p-value.

KS test's purpose is to test for differences in the shape of two sample distributions or to compare the expected statistical distribution. It compares overall shape of distributions, not specifically central tendency, dispersion or other parameters. Generally univariate, continuous data is used. P-value is obtained by finding the area under the expected distribution for values at least as extreme as the D statistic.

- (i) Each value is plotted against the percentage of values in the sample which are less than or equal to. Stepped line is used to connect the points.
- (ii) Fits a cumulative distribution function (CDF) for the selected distribution in order to find out how well the distribution fits the data.
- (iii) Displays a table with the distribution parameter estimates and the number of observation of the data.

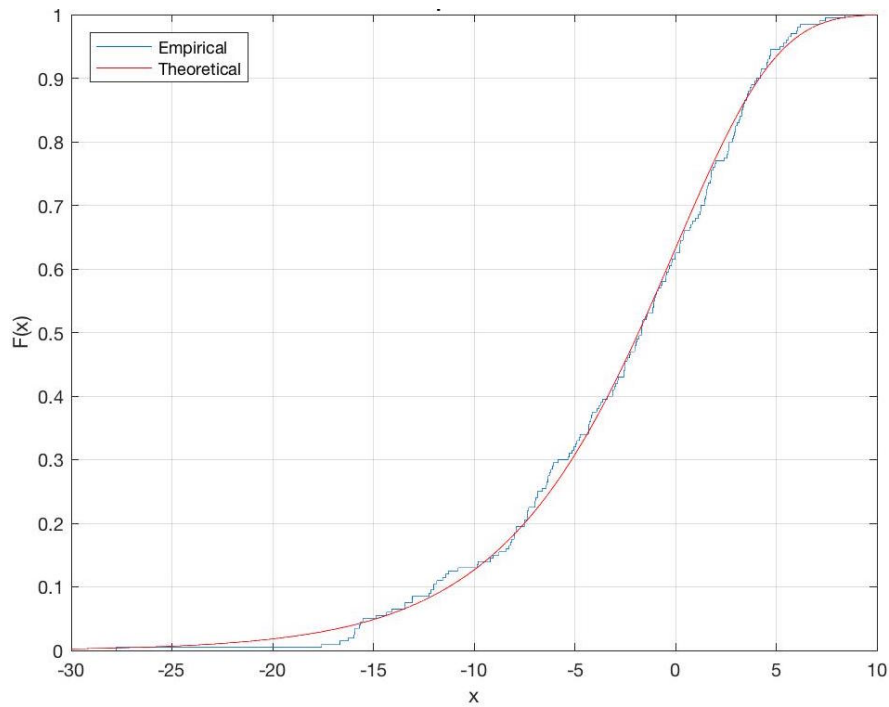


Figure 4.1. ECDF graph.

In Figure 4.1; maximum absolute difference between the empirical cumulative distribution graph and the theoretical cumulative distribution graph gives the distance value D . If D is greater than the test statistic D critical, test will reject the null hypothesis. Advantages and limitations of KS test;

Advantages;

- (i) KS test statistic does not rely on the fundamental Cumulative Distribution Function being tested.
- (ii) KS test does not have to have enough sample size for the approximations to be valid.

Limitations;

- (i) KS test is only appropriate for continuous distributions.
- (ii) KS test is more sensitive closer to the center of the distribution.
- (iii) Distribution should be fully specified. By simulation; shape, scale and location parameters should be determined.

Kolmogorov-Smirnov test is applied both to the normal and attack datasets using Matlab. Parameters used in Matlab are stated below;

k : maximum difference between the two cumulative distribution functions (which is known as D).

cv : critical value (obtained by Table 4.1) or calculated as;

$$cv = \frac{1.36}{\sqrt{n}} \quad (4.3)$$

n : sample size.

1.36 is given as an example for 95% significance level which is obtained from KS table. KS table contains the critical values for different significance levels. For $n > 35$;

$$\frac{\sqrt{-0.5 \cdot \ln(\frac{\alpha}{2})}}{\sqrt{n}} \quad (4.4)$$

According to Hypothesis Testing;

If $k < cv$; accept null hypothesis,

If $k > cv$; reject null hypothesis.

Decision to accept or reject null hypothesis depends on p-value and significance level α not by comparing the k and cv . Detailed calculation and results will be given in Experiment and Results chapter.

Table 4.1. KS table.

$n \backslash 1 - \alpha$	0.1	0.05	0.01	$n \backslash 1 - \alpha$	0.1	0.05	0.01
1	0.950	0.975	0.995	19	0.271	0.301	0.361
2	0.776	0.842	0.929	20	0.265	0.294	0.352
3	0.636	0.708	0.829	21	0.259	0.287	0.344
4	0.565	0.624	0.734	22	0.253	0.281	0.337
5	0.510	0.563	0.669	23	0.247	0.275	0.330
6	0.468	0.520	0.617	24	0.242	0.269	0.323
7	0.436	0.483	0.576	25	0.238	0.264	0.317
8	0.410	0.454	0.542	26	0.233	0.259	0.311
9	0.387	0.430	0.513	27	0.229	0.254	0.305
10	0.369	0.409	0.489	28	0.225	0.250	0.300
11	0.352	0.391	0.468	29	0.221	0.246	0.295
12	0.338	0.375	0.450	30	0.218	0.242	0.290
13	0.325	0.361	0.432	31	0.214	0.238	0.285
14	0.314	0.349	0.418	32	0.211	0.234	0.281
15	0.304	0.338	0.404	33	0.208	0.231	0.277
16	0.295	0.327	0.392	34	0.205	0.227	0.273
17	0.286	0.318	0.381	35	0.202	0.224	0.269
18	0.279	0.309	0.371	> 35	$\frac{1.224}{\sqrt{n}}$	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$

4.1.2. Chi-Square Test

Chi-Square test is used to find if the sample data belongs to a specific distribution or not. It answers the question of how well do our observed frequencies fit the expected frequencies. It is also a non-parametric test. Chi-square test tests if a sample set comes from a population with a particular variance as dispersion test. Test uses Cumulative Distribution Function like Kolmogorov-Smirnov test. In this test, data is put into groups as bins whereas this is not a real limitation. Actually test statistics depend on how the data is binned before the test. Test needs enough samples for the chi-square

estimate to be accurate [31].

Apart from Anderson-Darling and Kolmogorov-Smirnov, Chi-Square can be used as an alternative test. Kolmogorov-Smirnov and Anderson-Darling tests are applicable only for continuous datasets whereas like Binomial and Poisson distributions; Chi-Square test can be used on discrete distributions.

For the null and alternate hypothesis representation;

H_0 : Data follows the specified distribution.

H_1 : Data does not follow the specified distribution.

$$\chi^2 = \sum_{j=1}^M \frac{(O_j - E_j)^2}{E_j} \quad (4.5)$$

where O_j is the observed, E_j is the expected frequency and M is the number of classes (normal, attack) and the expected frequency is given by;

$$E_j = N(F(X_u) - F(X_l)) \quad (4.6)$$

where $F(.)$ is Cumulative Distribution Function for the tested distribution, X_u is the upper limit and X_l is the lower limit for class j , additionally N is sample size.

Selection of bins is important for Chi-Square test. Depending on the distribution, optimal bin width can be found. Expected frequency should be minimum 5.

α , test's significance level, generally chosen as 0.05,

k , non-empty cell number,

c , estimated parameter number (For example shape, scale and location)+1. For instance; in Weibull distribution there are 3 parameters required and $c=3+1$ thus it is 4.

Table 4.2. Chi Square table [32].

Degrees of Freedom	0.05	0.01	0.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

“Degree of Freedom” term is used in Chi-Square distribution and it equals to $k-c$. Depending on the degree of freedom and significance level, table gives the critical value. After applying Chi-Square goodness-of-fit test; calculated statistic will be compared to the critical value given by the table. If calculated statistic is higher than the critical value, test rejects the null hypothesis. Otherwise statistic will be less than Chi-Square critical value and then null hypothesis is accepted. Detailed information about the calculation is stated in Section Five.

Generally, statistical tests have distributional assumptions like normality or in reliability modeling, assumption of Weibull distribution. Robust and non-parametric techniques do not have strong distributional assumptions. Under a broad area of distributional assumptions, robust tests give good results. Specific distributional assumption techniques are more effective than robust and non-parametric techniques. Parametric

methods are preferable when distributional assumption is confirmed.

4.2. Classification Methods

In this section, Naive Bayes classification and k-Nearest Neighbor classification methods are studied.

4.2.1. Naive Bayes Classification

Classification is one of the main tasks in data analysis which helps the statisticians to categorize the data for an efficient and effective use. A classifier is kind of a function which assigns a class (in this thesis aspect, normal or attack) to instances presented by a set of attributes (source port, destination port, ip addresses, skewness, kurtosis etc.). Naive Bayes (NB) classifiers are one of the most effective classifiers. Bayes theorem is the basis of Naive Bayes classification.

By using the training data, NB classifier learns the conditional probability of each attribute A_i given the class label C . In order to calculate the probability of C given the particular instance of $A_1, \dots, A_n$ Bayes rule is applied. Class is then predicted with the highest posterior probability [33]. There is a strong assumption in NB classification that all the attributes A_i 's are conditionally independent given the value of the class C . Probabilistic independence means, A_1 is independent of A_2 given C whenever $P_r(A_1|A_2, C) = P_r(A_1|C)$ for all possible values of A_1, A_2 and C whenever $P_r(C) > 0$. Bayes theorem calculates the posterior probability $P(C|Y)$, from $P(C)$, $P(Y)$ and $P(Y|C)$ which is likelihood.

$$P(C|Y) = \frac{P(Y|C)P(C)}{P(Y)} \quad (4.7)$$

Naive Bayes assumes conditional independence and the Equation 4.7 becomes;

$$P(C|Y) = P(C) \prod_j P(Y_j|C) \quad (4.8)$$

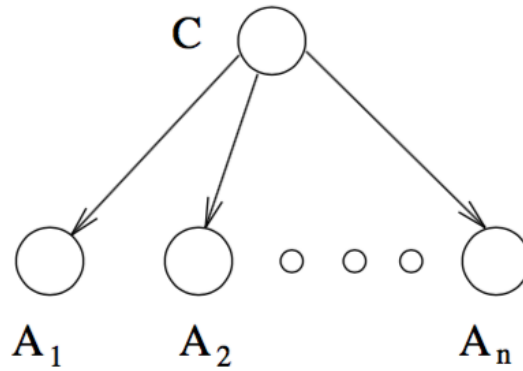


Figure 4.2. Naive Bayes structure.

Considering the Equation 4.8, the likelihood probability $\prod_j P(Y_j|C)$ can be used as a result of class C . This result can be used as a threshold for classification.

NB is simple and if the conditional independence assumption really valid, classifier converge in less time than other models like logistic regression etc. It will need less training data and also decrease the training time. Naive Bayes classification is very useful for large dataset because of the reason that it can be implemented easily without any complicated iterative parameters. Even though the method is simple, Naive Bayes classification outperforms other sophisticated classification methods.

%75 of the dataset is used for training and the remaining %25 is used for testing. WEKA tool is used for NB classification. Results are given in the following chapter.

4.2.2. k-Nearest Neighbour

The kNN is a non-parametric, instance-based, lazy learning method which is commonly used in pattern recognition and machine learning. If the samples are similar, they are placed close to each other in the distribution [34]. Classifiers which called “lazy” are instance based learners. “Lazy” means, model keeps all the training data and when a new test data comes; it builds a classifier and then classify the incoming test data. It does not build a classifier according to the training data that is why

classification times are higher [35]. Required time for training phase of kNN is less when compared with eager learning methods (which is the opposite of the lazy learning systems) like decision trees, bayesian classifiers, etc [36]. Eager methods prepare the classification model according to the train dataset.

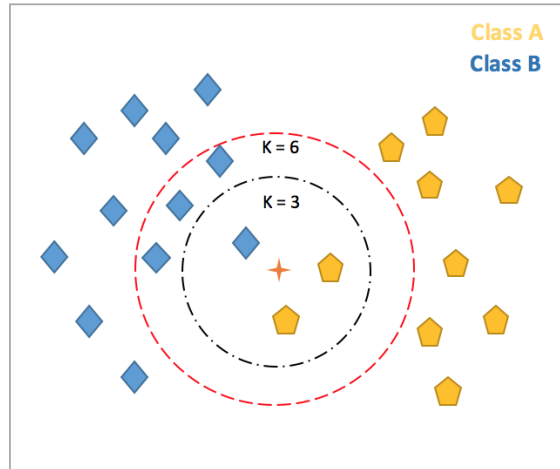


Figure 4.3. k-Nearest Neighbour

When a new test data is classified, its distance to the training samples are calculated. Depending on the closeness to the specified class groups; test data is assigned to a class. Test data's classification depends on its nearest neighbours classes'. k is the number of neighbours, if k is three; classification of the test data depends on the three neighbour's classes. Deciding the number of k directly effects the classification performance [37]. There are different distance measuring methods are used in literature. Euclidean distance and Hamming distance are the commonly used distance measuring techniques for kNN [38]. In Figure 4.3, it can be seen that selection of k affects the classification directly. When k is selected as 3, class A is chosen whereas when k increases to 6, class B is selected. Advantages and the disadvantages of the kNN stated below [39];

Advantages;

- (i) Implementation is easy.
- (ii) Training time is fast.
- (iii) Does not easily affected by the noise.
- (iv) Performs well on the multiple class labelled data.

Disadvantages;

- (i) Computational costs are expensive when the neighbours have large numbers.
- (ii) Needs more storage.
- (iii) Runs slowly due to supervised learning. It scans entire training data to make each prediction.
- (iv) Easily affected from the data structure.

In this thesis, kNN algorithm is used to compare the results which obtained from the Naive Bayes classification. Results are shown in experiment section.

4.3. Box Cox Transformation

Normally distributed data is preferred by the statisticians in order to apply some statistical tools like t-tests, analysis of variance (ANOVA) etc. Normality assumption is a gateway to apply different statistical techniques. When non-normally distributed data is investigated, the non-normality reason should be determined and then corrective actions should be taken into account.

Data transformation (Box Cox transformation) is one of the most used method to make the data normal. Linear transformations do not help to make non-normally distributed data into normally distributed data. In this transformation, main purpose is to decrease the standard deviation. After decreasing the standard deviation, even the transformed data is not normally distributed; it gives better classification result.

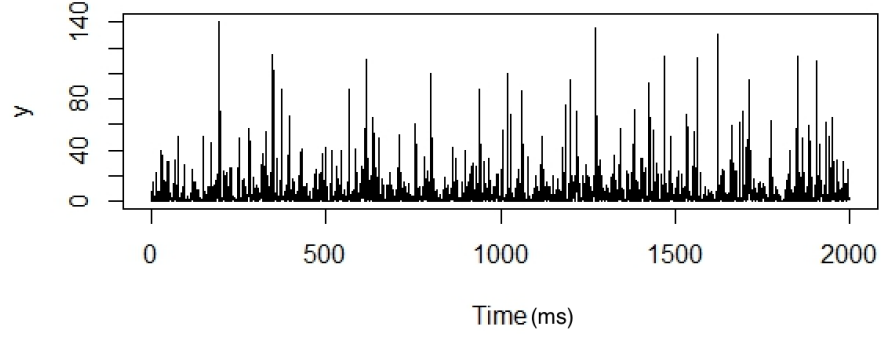


Figure 4.4. Signal before transformation

Lambda (λ) is an exponent which is used for Box Cox transformation. It is found by maximum likelihood estimation. The point where the log likelihood is maximum, gives a lambda (λ) value which can be seen in the following figures. Obtained lambda value is used in the transformation formulation below. Lambda values generally varies between -5 and 5 and they are treated as optimum value. Optimum value results in the best approximation of a normal distribution curve [40].

Transformation formula is given in Equation 4.9;

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log y, & \text{if } \lambda = 0. \end{cases} \quad (4.9)$$

It is important to note that Box Cox transformation works on positive data. It is easy to calculate the new transformed values by built-in functions on different statistical applications. Box Cox transformed data is given in Section Five.

In order to observe the effect of the transformation, positively skewed data is created and its transformation is given in Figure 4.4. Histogram of the signal before transformation is given in Figure 4.5. QQ Plot of the signal before transformation is given in Figure 4.6.

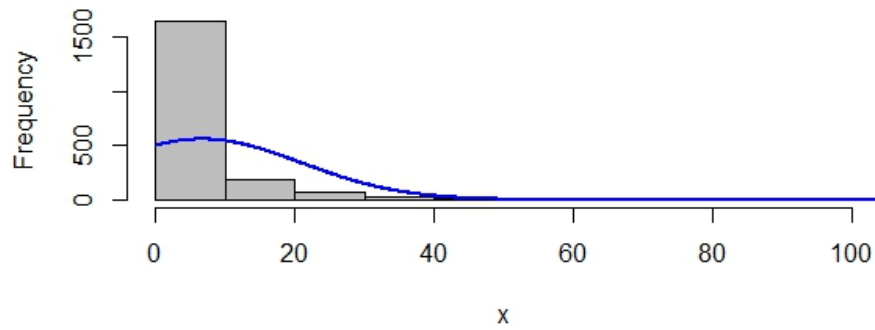


Figure 4.5. Signal histogram before transformation

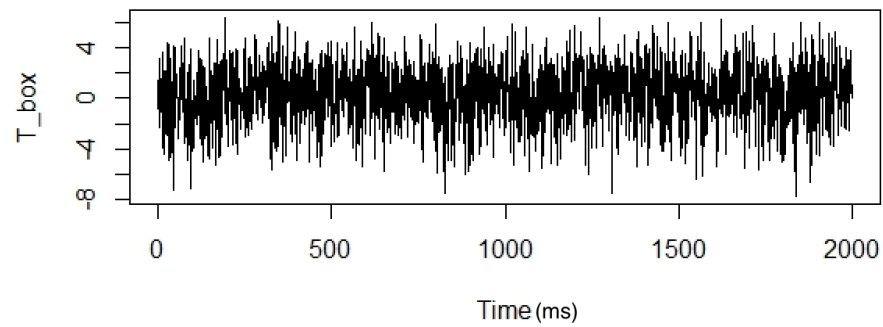


Figure 4.7. Signal after transformation

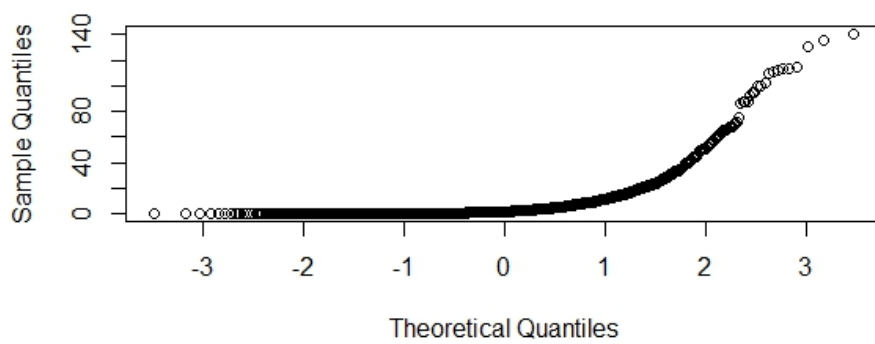


Figure 4.6. Signal QQ plot before transformation

After applying Box Cox, transformed data is given in Figure 4.7. Histogram of the transformed data is in Figure 4.8. QQ Plot of the transformed data is given in Figure 4.9.

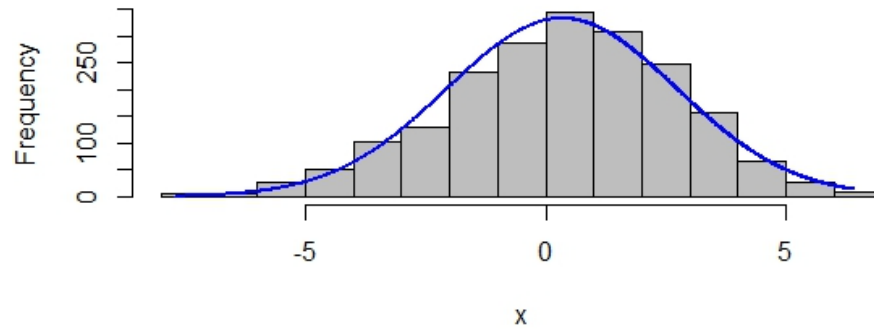


Figure 4.8. Signal histogram after transformation

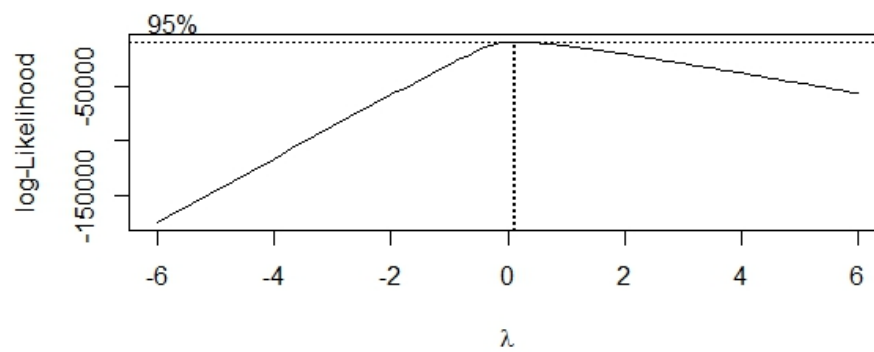


Figure 4.10. Signal log likelihood

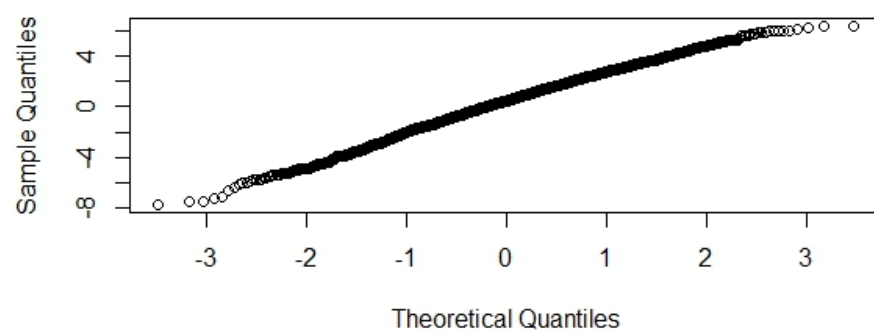


Figure 4.9. Signal QQ plot after transformation

λ log likelihood is stated in Figure 4.10. Transformation test statistics summary; before transformation standard deviation is 14.33632 whereas after the transformation

standard deviation becomes 2.399434. Standard deviation is decreased and the new data is more normally distributed then before the transformation.

4.4. WEKA

Waikato Environment for Knowledge Analysis (WEKA) is a statistical tool which is designed to be used in machine learning and data mining areas. It provides implementation of learning algorithms that can easily be applied to the dataset. There are variety of tools for transforming datasets. Datasets can be preprocessed before any implementation and sent to the learning scheme. Resulting classifier and it's performance can also be analyzed [41].

The WEKA workbench contains methods for the main data mining problems: classification, regression, association rule mining, clustering and attribute selection. Different data visualization facilities and data processing tools are available.

4.4.1. ARFF (Attribute Relation File Format)

WEKA uses ARFF (Attribute Relation File Format) for loading datasets. Basic explanation is given in Figure 4.11 with an example. Arff files have two different parts:

(i) Header Information

- Relation Declaration: This is the first line of the .arff file. The format is @relation "relation-name" where "relation name" is a string.
- Attribute Declaration: Every feature is represented by its own @attribute statement. The format is @attribute "attribute-name" "datatype" where "attribute-name" must start with an alphabetic character and "datatype" can be any of numeric, "nominal-specification", "string" or "date".

Real or integer numbers could be the numeric attributes. Nominal values are described by a "nominal-specification" listing the possible values: "nominal-name1", "nominal-name2", "nominal-name3", ... String attributes help user to create attributes having arbitrary textual values.

```

@relation whatever

@attribute skewness numeric
@attribute kurtosis numeric
@attribute class { 0 , 1 }

@data
0.419261845 , 1.697438133 , 0
0.217194617 , 1.723448941 , 0
-0.516242789 , 1.402417347 , 0
-1.005794424 , 1.184376549 , 0
-1.685229308 , 0.887742941 , 0
-0.597444947 , 1.176612862 , 0
-0.502404799 , 1.308785816 , 0
0.699066403 , 2.345413589 , 1
1.014559834 , 3.017298827 , 1
1.269496003 , 3.67142761 , 1
1.884851132 , 5.010356664 , 1
0.447333334 , 1.947971768 , 1
0.828208611 , 2.535833874 , 1
1.033644233 , 2.969544535 , 1

```

Figure 4.11. CAIDA dataset in attribute relation file format.

Date attributes allow user to create attributes having dates with specific date formats.

(ii) Data Information

- Data Declaration: Single line indicates the start of the data information part. The format is @data.
- Instance Lines: Represent instances and their attribute values. Attribute values are in the same order with the header information part and separated with commas.

4.4.2. WEKA Explorer

There are four different Graphical User Interfaces (GUI) are available in the application; Explorer, Experimenter, KnowledgeFlow and Simple CLI. Explorer GUI is used for the performance calculations. All the built in functions and algorithms are stated. Basically Data preprocessing, Classifying, Clustering, Attribute selection and visualization choices are available. Before using the algorithms brief information is

given about the options. Applications are user friendly and easy to use. Also, sensible default values for the tools are selected to simplify users's task [41].

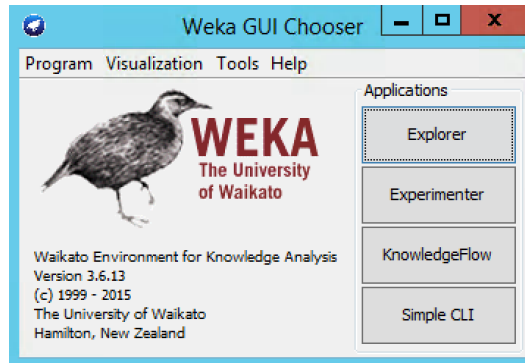


Figure 4.12. WEKA gui [41].

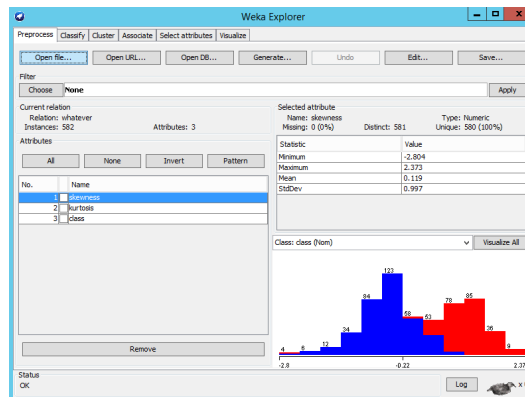


Figure 4.13. WEKA explorer interface [41].

Detailed information about the Experimenter, KnowledgeFlow and Simple CLI options can be found in [41]. In this thesis Explorer option is used.

5. EXPERIMENTS AND RESULTS

5.1. Statistical Results

5.1.1. Overview

There are lots of different datasets in literature for data analysts to test their proposed methods on network security platform. CAIDA (Center for Applied Internet Data Analysis) is a platform on network security for collection and sharing of internet traffic data [42]. CAIDA 2007 is attack traffic whereas CAIDA 2008 is normal traffic. Almost 80% of the traffic on internet uses TCP transport layer 4 protocol [43] and CAIDA traffic is also TCP based.

As it is mentioned before; main purpose of this thesis is to differentiate the specifications of attack and normal traffic depending on the time series analysis. Investigation process is given below;

- (i) Datasets are taken into time series. (Sampling every 1 ms and then 128 windowing applied.)
- (ii) Skewness and Kurtosis parameters are found and histograms of these parameters shown in graphical results.
- (iii) QQ plot (graphical) and Shapiro-Wilk tests (numerical) are applied in order to find the normality.
- (iv) Different distribution fitting tests are used to find the best distribution which represents the data.
- (v) In order to find the mathematical result of the best distribution; Kolmogorov-Smirnoff and Chi-Square tests are applied.
- (vi) Depending on the results best matching distribution type is found.

CAIDA attack dataset and CAIDA normal dataset are evaluated by applying the above mentioned methods. Classification features for network traffic will be Skewness

and Kurtosis parameters of both CAIDA dataset.

Later on BOXCOS transformation is applied. Thresholding, kNN and Naive Bayes classification (using WEKA) methods are used to differentiate attack data from normal data.

5.1.2. Histogram of the Datasets

The duration of CAIDA attack data is approximately 25 minute long whereas CAIDA normal data is 1 minute long. Traffic is sampled in 1 mili seconds and then 128 windowing is applied. After filtering the useful information in the data; 306 samples are found in CAIDA attack data and 468 samples are obtained for CAIDA normal data for skewness and kurtosis parameters.

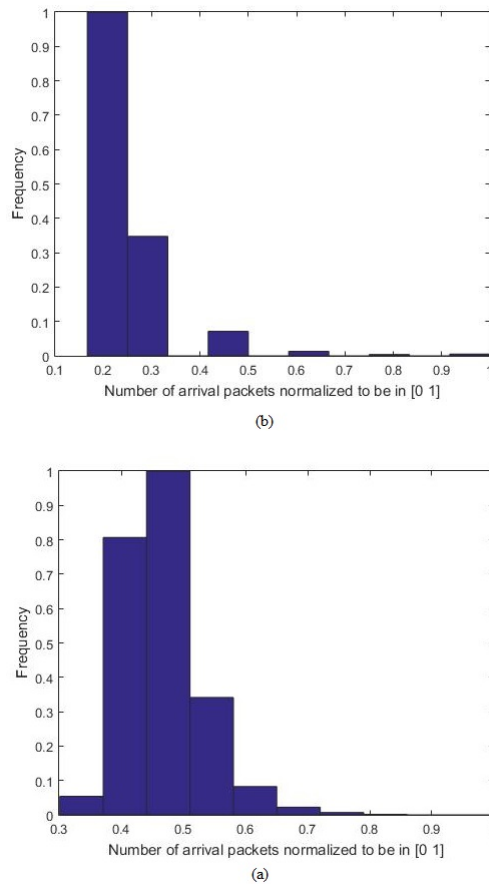


Figure 5.1. Normalized histogram of arrival packet numbers, (a) normal, (b) attack [14].

Table 5.1. CAIDA statistics.

	Mean	Median	Mode	Std	Kurtosis	Skewness
CAIDA Normal	0.47	0.46	0.45	0.06	4.63	0.98
CAIDA Attack	0.23	0.21	0.19	0.12	11.42	2.42

Figure 5.1 displays the histogram of both normal and attack traffic packet number obtained by $x(n)$. Both histograms (packet number) are normalized for making the number of arrival packets exists in $[0 \ 1]$ intervals. Maximum frequency is normalized to be 1. Histogram is normalized according to the area under the curve; total area is calculated and each bin is divided by that value. From the Figure 5.1 we can conclude that packet number by itself can not be a useful feature for network traffic to differentiate attack from normal traffic because of the reason that two histograms overlap each other; thus appropriate features should be extracted from the packet number time series.

Statistical information about the CAIDA dataset is stated in Table 5.1. Both normal and attack traffics are heavy-tailed and right-skewed. It can be concluded from Table 5.1 that in order to differentiate normal and attack traffic Kurtosis and Skewness can be used as features for network traffic.

In Figure 5.2, the vertical axis represents the kurtosis values whereas horizontal axis represents the values of skewness. Figure 5.2 shows that skewness and kurtosis attack values have stronger correlation than skewness and kurtosis normal values. This correlation factor has positive effect on classification performance. Correlation coefficients are 0.8584 and 0.6031 respectively.

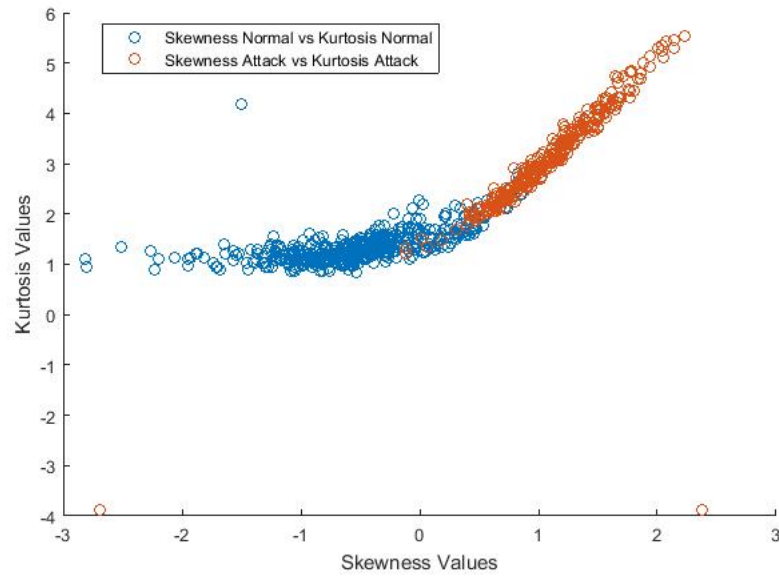


Figure 5.2. Scatterplot

Table 5.2. Scatterplot measures.

	Entropy
Skewness Attack	3.7632
Skewness Normal	1.6947
Kurtosis Attack	0.0795
Kurtosis Normal	0.6382

As it can be seen from Table 5.2, skewness entropy values are higher than kurtosis entropy values which means for classification, skewness performance is expected to be higher than kurtosis. Classification results shows that skewness is a useful feature than kurtosis for differentiating attack from normal traffic.

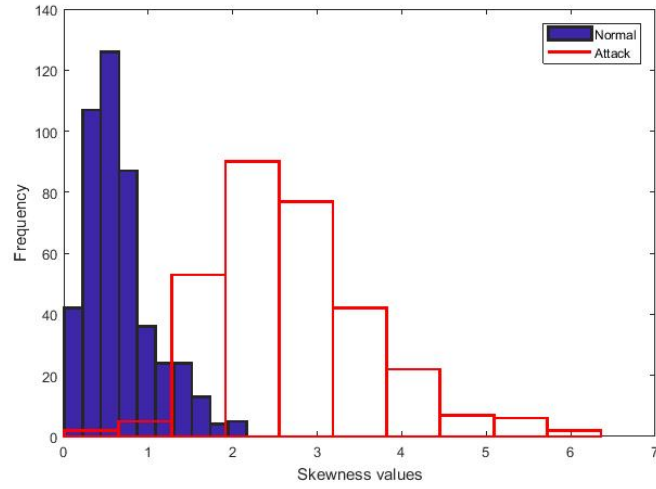


Figure 5.3. CAIDA skewness normal and attack.

Table 5.3. Skewness parameter table.

	Mean	Mode	Median	Var	Std
CAIDA Attack	2.7047	0.0100	2.5595	0.8799	0.9380
CAIDA Normal	0.6627	0.0064	0.5807	0.1665	0.4081

From the Figure 5.3 and Table 5.3, it can be concluded that both CAIDA Normal and Attack are skewed to the right. For both dataset, mean value is higher than the median. Right tail is longer than the left tail which means they are positively skewed. CAIDA attack is more skewed to the right then the CAIDA normal data.

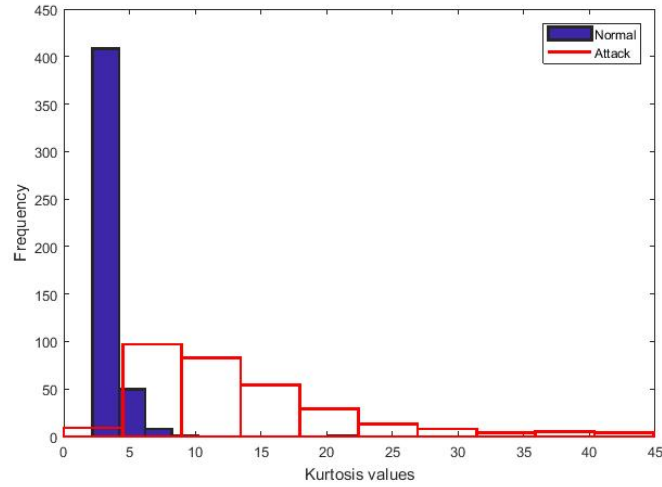


Figure 5.4. CAIDA kurtosis normal and attack.

Table 5.4. Kurtosis parameter table.

	Mean	Mode	Median	Var	Std
CAIDA Attack	13.2775	$1.0e - 03$	10.9903	64.1031	8.0064
CAIDA Normal	3.4463	2.2080	3.1762	1.6042	1.2666

From the Figure 5.4 and Table 5.4, mean value for CAIDA attack kurtosis is 13.2775 whereas CAIDA normal kurtosis is 3.4463. In order to find excess kurtosis 3 must be subtracted. After this calculation CAIDA attack excess kurtosis becomes 10.2775 and CAIDA normal excess kurtosis becomes 0.4463. Both of the excess kurtosis values are higher than zero, which means that they are leptokurtic.

For Kurtosis, if the values are greater than +1 distribution is too peaked and if the values are smaller than -1 , it means that distribution is too flat. When compared, CAIDA Attack has more leptokurtic distribution than CAIDA Normal. Tail is fatter and longer also its central peak is sharper.

5.1.3. Normality Tests

After having the skewness and kurtosis results; QQ Plot and Shapiro-Wilk tests are used for checking the normality.

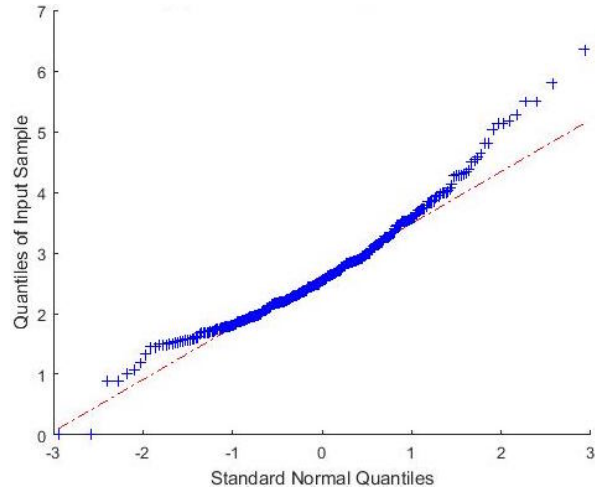


Figure 5.5. QQPlot CAIDA skewness attack.

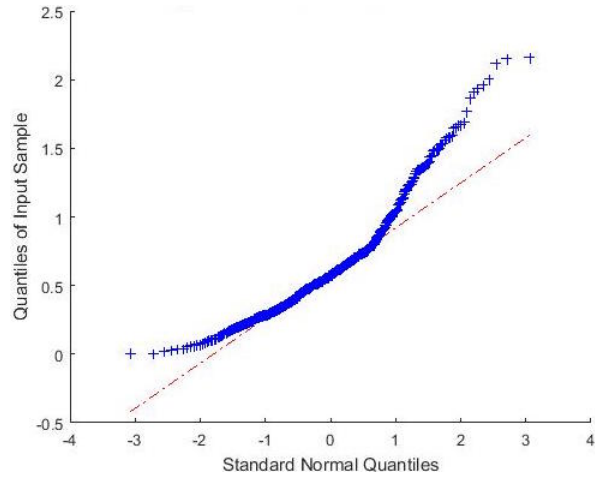


Figure 5.6. QQPlot CAIDA skewness normal.

According to the Figure 5.5 and Figure 5.6, Quantile-Quantile plots, both CAIDA Skewness Attack and Normal data is not exactly normally distributed. In order to check the normality numerically, Shapiro-Wilk test is used. Results are given in Table 5.5.

Table 5.5. Shapiro Wilk test results.

	W	p value
CAIDA Normal Skewness	0.91998	$4.634e - 15$
CAIDA Attack Skewness	0.96078	$2.461e - 07$
CAIDA Normal Kurtosis	0.56018	$< 2.2e - 16$
CAIDA Attack Kurtosis	0.86848	$1.742e - 15$

On significance level, p-values are smaller than the alpha 0.05 value, it rejects the null hypothesis that, these distributions do not have Gaussian distribution. W statistics for all data is smaller than 1 where skewness is closer to 1 then the kurtosis. For Shapiro-Wilk test calculation R programming is used which is an environment in statistics like Matlab.

Both QQ Plot and Shapiro-Wilk test results show that these datasets are not normally distributed. It is proven by both graphically and numerically. After this point, main concern is to find which distribution best represents the given datasets. When the best matching distribution is found, it will be easier to classify the attack traffic from normal traffic.

5.1.4. Distribution Fitting Test

In this section, depending on the dataset, best matching distribution is trying to be found depending on the Kolmogorov-Smirnoff and Chi-Square measuring tests. Randomly chosen 9 different distribution types are applied to the dataset. CAIDA normal and attack data's skewness and kurtosis features' distributions have log logistic distribution properties. Graphical representation and results are given in Figure 5.7 to Figure 5.10.

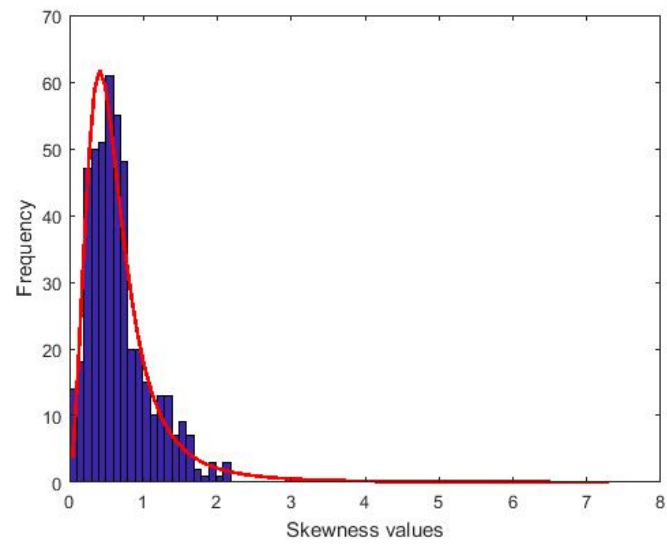


Figure 5.7. CAIDA normal skewness log logistic distribution.

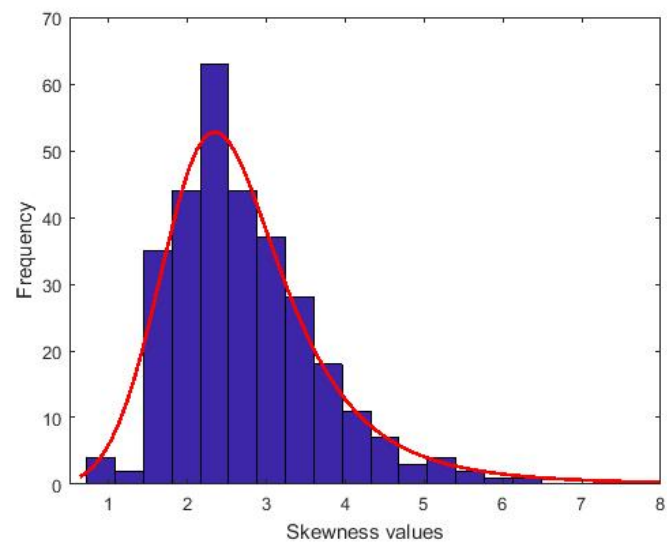


Figure 5.8. CAIDA attack skewness log logistic distribution.

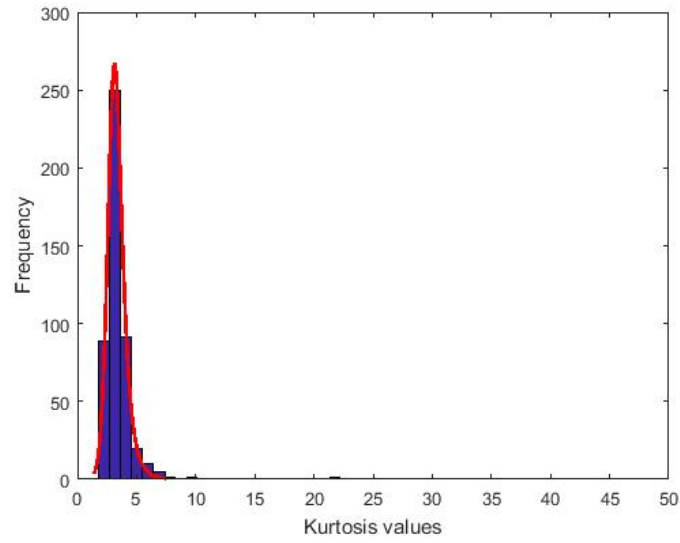


Figure 5.9. CAIDA normal kurtosis log logistic distribution.

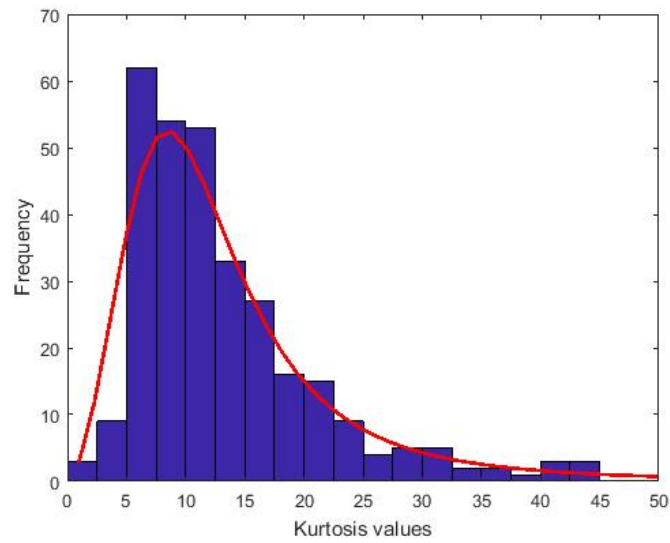


Figure 5.10. CAIDA attack kurtosis log logistic distribution.

In Figures 5.7 to Figure 5.10, y-axis represents sample size and x-axis represents bin size. Bin size is adjusted automatically according to the distribution. KS test and Chi-Square test results are given in Table 5.6 and Table 5.7; according to the p value

results, datasets have log logistic distribution.

Table 5.6. KS test log logistic distribution (p value).

	Skewness	Kurtosis
CAIDA Attack	0.6521	0.5836
CAIDA Normal	0.2822	0.6692

Table 5.7. Chi-Square test log logistic distribution (p value).

	Skewness	Kurtosis
CAIDA Attack	0.5836	0.0650
CAIDA Normal	0.1712	0.0614

As it can be concluded from the Table 5.6 and Table 5.7; the dataset is best represented with Log logistic distribution. Also KS test results are higher than Chi-Square test results. Both KS test and Chi-Square tests accepts the null hypothesis that datasets have the Log logistic distribution.

5.2. Classification Results

5.2.1. Naive Bayes Classification

Before Box Cox transformation, CAIDA normal and attack dataset have log logistic distribution. Classification of the attack and normal traffic is done by using Naive Bayes classification method in WEKA tool. Obtained results is going to be compared with the Box Cox transformed data results. Dataset information is given in Table 5.8.

Table 5.8. Dataset.

	Total	Train (%75)	Test (%25)
CAIDA Normal	468	351	117
CAIDA Attack	306	231	75

For classification; system should be trained first with the %75 of the total data and then the remaining %25 data is used for testing. While evaluating the results; true positive rate, false positive rate and accuracy values are used. Definition of this statistics are given below.

True positive rate (TPR), False positive rate (FPR), accuracy, Receiver operating characteristic (ROC) and Area under curve (AUC) are mainly used as performance metrics in statistics. Performance metrics measure the classifiers' prediction performance and for better understanding, confusion matrix should be known.

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP-True Positive	FN-False Negative
	Negative	FP-False Positive	TN-True Negative

Figure 5.11. Confusion matrix.

When the incoming traffic is an attack traffic and the system identifies it as an attack; identification is true and because the incoming traffic is identified as an attack, it means positivity. In the same manner if the incoming traffic is normal but identified as an attack, identification process is wrong that is why it is false. It is classified as an attack and it becomes false positive. Namely positivity and negativity represents; classification as an attack or normal traffic respectively, whereas true and

false represents; classification is correct or not.

In detection estimation, accuracy is the most commonly used performance metric. In order to evaluate the results objectively true positive rate, false positive rate and accuracy should be taken into account all together. The confusion matrix will be helpful to understand TPR, FPR and accuracy.

$$TPR = \frac{TP}{TP + FN} \quad (5.1)$$

True positive rate is a ratio of correctly classified attack traffic divided by the sum of correctly classified attack and incorrectly classified attack traffic.

$$FPR = \frac{FP}{FP + TN} \quad (5.2)$$

False positive rate is a ratio of incorrectly classified normal traffic divided by the sum of incorrectly classified normal traffic and correctly classified normal traffic.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.3)$$

Accuracy is the ratio of summation of correctly classified attack and correctly classified normal traffic divided by the total traffic. As it is mentioned before, skewness and kurtosis are taken as features for Naive Bayes Classification and obtained results are given in Table 5.9.

Table 5.9. Naive bayes classification results.

	TP	FP	TN	FN	TPR(%)	FPR (%)	Accuracy (%)
Skewness	71	9	108	4	94.6667	7.6923	93.2292
Kurtosis	56	2	115	19	74.6667	1.7094	89.0625
Combined	67	4	113	8	89.3333	3.4188	93.75

For better understanding, bar chart representation is given in Figure 5.12.

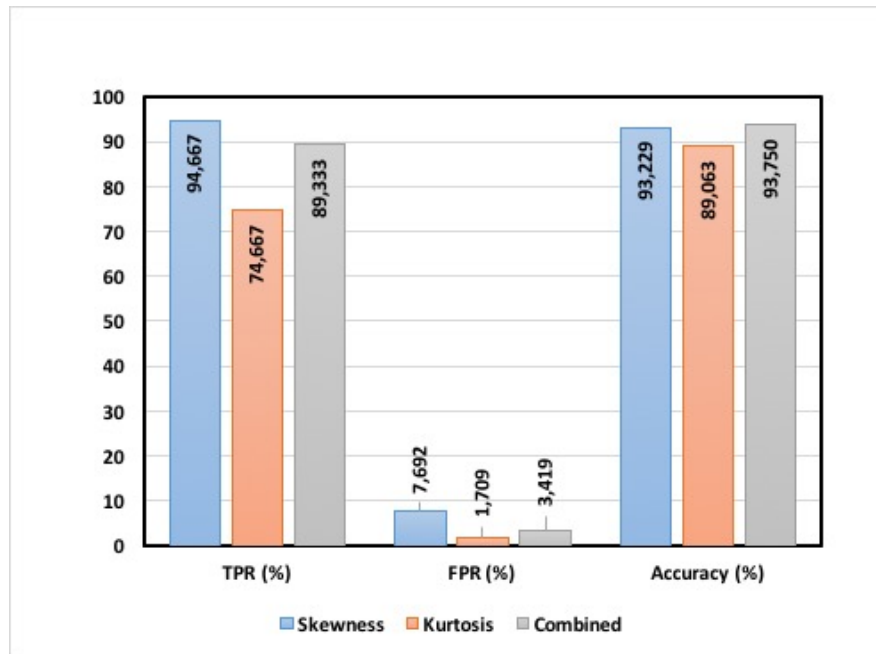


Figure 5.12. NB classification results.

Skewness resulted in the best true positive rate whereas kurtosis's false positive rate is the lowest. When both skewness and kurtosis features are combined as two features of network traffic in order to increase classification performance, it resulted in the best accuracy.

5.2.2. Analysis After Box Cox Transformation

After seeing that the distributions are log logistic, in order to improve statistical efficiency Box Cox transformation is applied. Figure 5.13 to Figure 5.16 are the graphical representation of before and after Box Cox transformation.

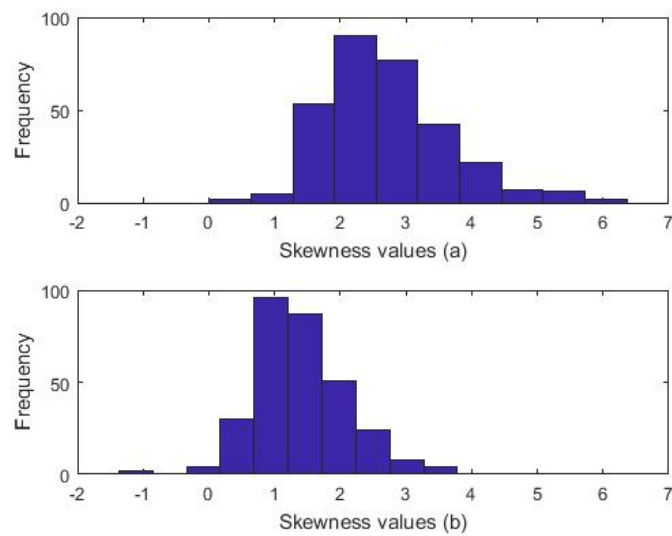


Figure 5.13. Attack skewness (a) Before Box Cox (b) After Box Cox.

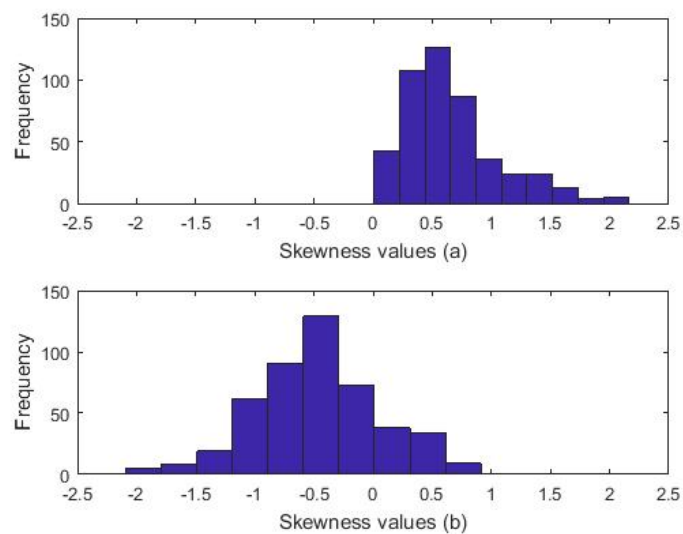


Figure 5.14. Normal skewness (a) Before Box Cox (b) After Box Cox.

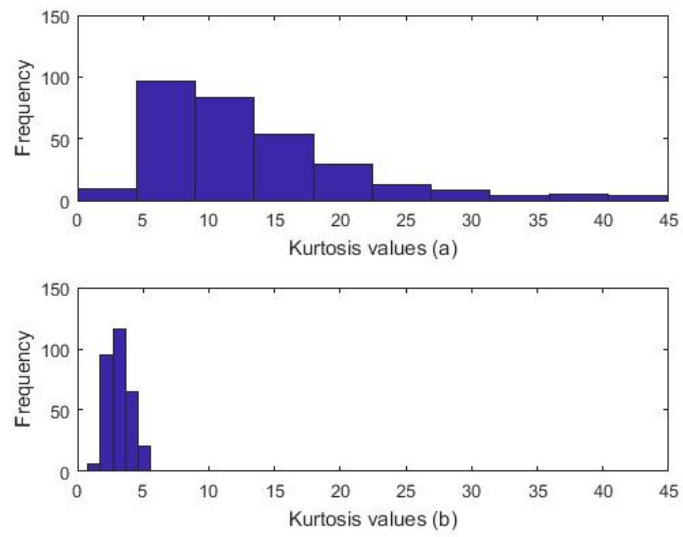


Figure 5.15. Attack kurtosis (a) Before Box Cox (b) After Box Cox.

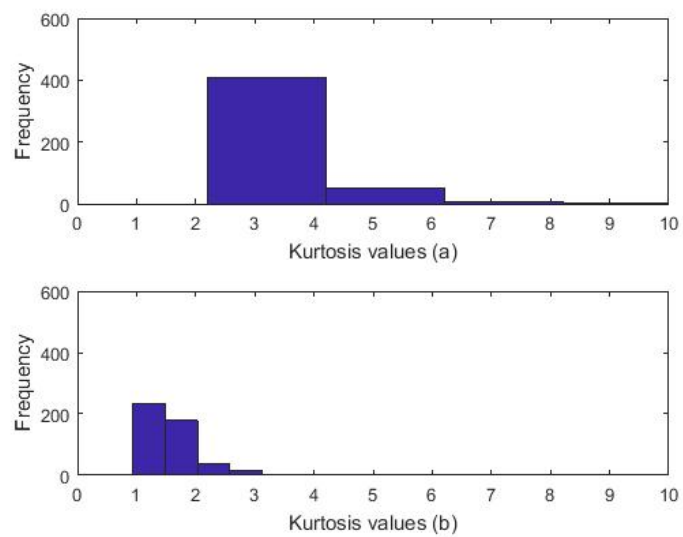


Figure 5.16. Normal kurtosis (a) Before Box Cox (b) After Box Cox.

After the Box Cox transformation, variance and the standard deviation of the data decreased.

5.2.3. Thresholding Method

Box Cox transformation has a positive statistical effect on the data. Naive Bayes classification is applied to see the transformation's effect on the dataset. Thresholding method is also used as an additional classification method to detect attack from normal traffics. In the training phase, thresholds depending on different $(\mu + 1\sigma)$, $(\mu + 2\sigma)$, $(\mu + 3\sigma)$ etc. standard deviations around the mean are applied to the dataset. The threshold which have the highest TPR and accuracy in the training data is chosen for the test data. Data is known to have log logistic distribution and positive skewness. From the training results, for skewness feature 1.5 *std* around the mean is calculated whereas for kurtosis feature 2 *std* around the mean gave highest TPR and the accuracy.

Table 5.10. Skewness Thresholding results.

	Mean(M)	Std	M+1Std	M+1.5Std	M+2Std	M+3Std	M+4Std
Skew	-0.542	0.586	0.337	0.630	0.337	1.217	1.803
Acc(%)			88.54	92.708	90.10	66.66	61.97

As in the classification, 75% of the data is used for training and 25% is used for testing. Skewness graph for thresholding is given in Figure 5.17. The calculated threshold value is 0.337.

Related ROC Curve is given in Figure 5.18.

In Receiver Operating Characteristic (ROC) curve; TPR is Sensitivity is plotted in function of FPR (100-Specificity) for different cut-off points. Each point on the curve represents a Sensitivity/Specificity pair that belongs to a particular decision threshold.

If two distributions does not overlap and perfectly discriminated, ROC curve has a straight line to the top left corner. Thus, when the curve get closer to the upper left corner, higher accuracy is obtained. It is desirable for the curve to be as close

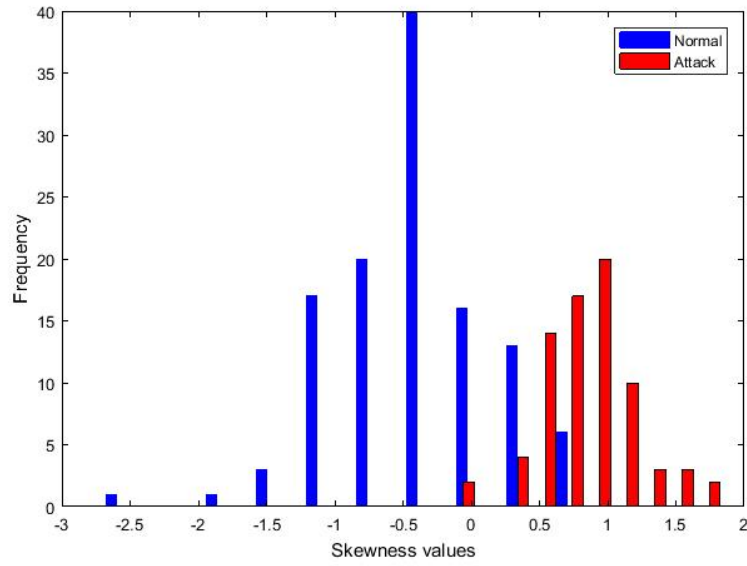


Figure 5.17. Skewness thresholding test.

as possible to the upper left corner in statistical aspect. In Figure 5.18, the red line represents 45-degree diagonal. When the curve gets closer to the red line, test means less accuracy.

Accuracy is calculated according to the area under the ROC curve. If the area equals to 1; it means test is perfect whereas if the area equals or less than 0.5, test is worthless. AUC is also calculated as; 0.9812 and 0.9764 for skewness and kurtosis respectively.

For the Kurtosis; threshold value is chosen as 1.997 in Figure 5.19.

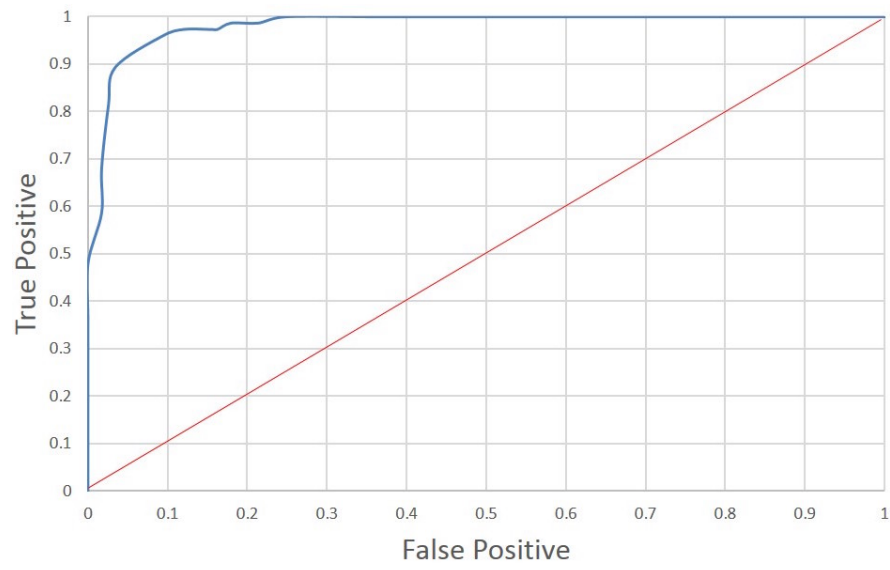


Figure 5.18. Skewness thresholding test ROC.

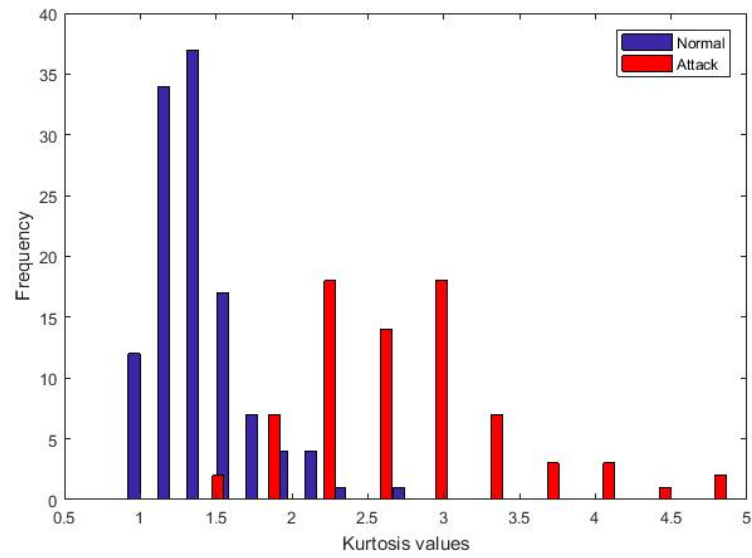


Figure 5.19. Kurtosis thresholding test.

Table 5.11. Kurtosis Thresholding results.

	Mean(M)	Std	M+1Std	M+2Std	M+3Std	M+4Std	M+5Std
Kurtosis	1.343	0.327	1.670	1.997	2.324	2.651	2.978
Accuracy(%)			90.104	92.708	87.5	80.729	73.958

Related ROC Curve is given in Figure 5.20.

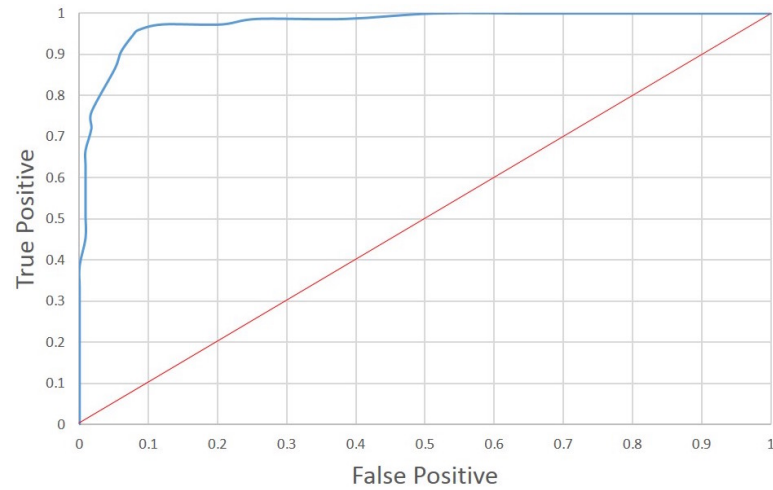


Figure 5.20. Kurtosis thresholding test ROC.

Both ROC curves are visually put together in Figure 5.21 for better comparison.

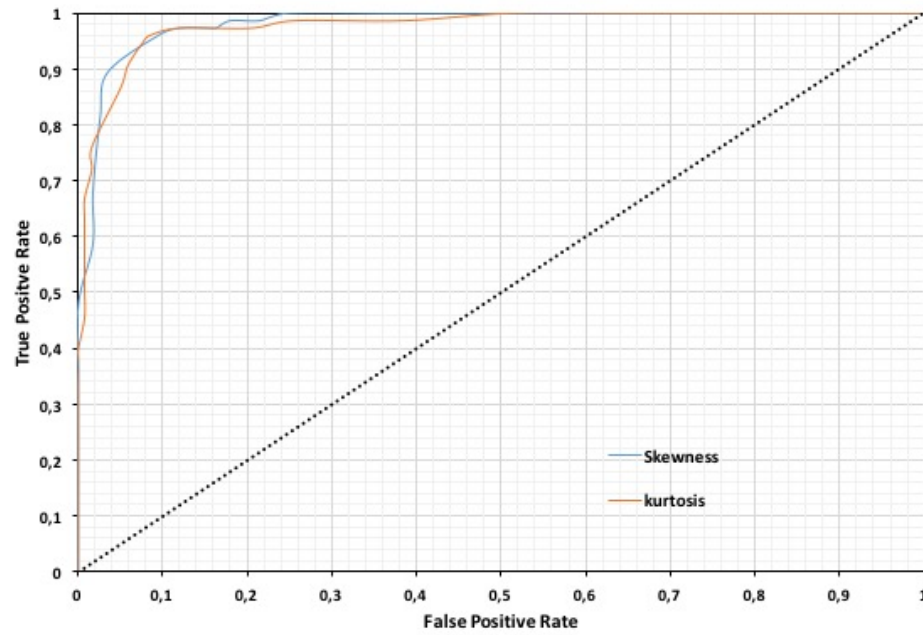


Figure 5.21. Combined ROC

After the thresholding method, kNN and Naive Bayes classification is applied. Overall thresholding, kNN and Naive Bayes classification results are stated in Table 5.12 and Figure 5.22.

Table 5.12. Thresholding and NB classification results after Box Cox.

	TP	FP	TN	FN	TPR(%)	FPR (%)	Accuracy (%)
Thresholding Skew.	72	11	106	3	96	9.402	92.708
Thresholding Kurto.	68	7	110	7	90.667	5.983	92.708
NB Class. Skew.	73	11	106	2	97.3333	9.4017	93.2292
NB Class. Kurto.	68	6	111	7	90.6667	5.1282	93.2292
NB Class. Comb.	71	8	109	4	94.6667	6.8376	93.75
kNN (3) Comb.	72	6	111	3	96	5.1282051	95.31
Entropy	74	5	112	1	98.7	4.3	96.9

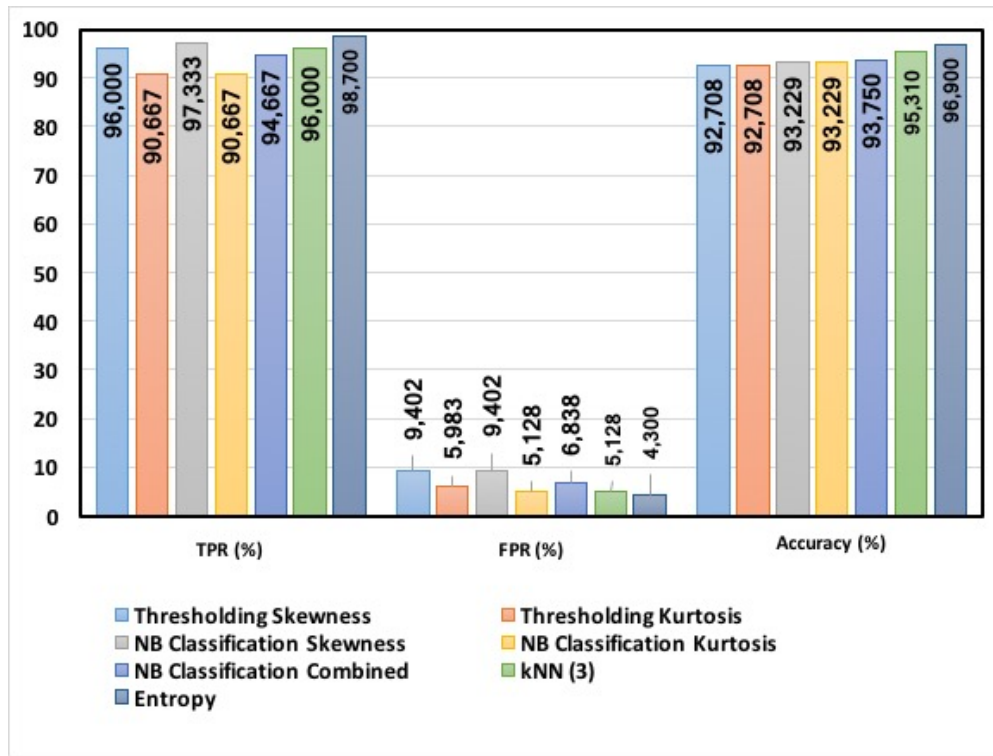


Figure 5.22. Thresholding and NB classification results.

Naive Bayes classification method results are better than the thresholding method for both skewness and kurtosis features. Except Entropy feature, the highest accuracy belongs to kNN classifier with combined features. Naive Bayes classification has the independency assumption between the features whereas according to the scatter plot in Figure 5.2, data relation can be seen. That is why kNN resulted better than the NB classification. NB produce a model according to training data whereas kNN does not. When test data comes to the system kNN uses the training and test data for classification that is why it takes much more time than NB. kNN needs more storage than NB.

Entropy is another statistical measure which is broadly used in statistics to differentiate attack from normal traffics. As stated in [22, 23], the entropy of the flow packet number in time series is used to compare our selected skewness and kurtosis features' classification results. From Figure 5.22, TPR of the entropy feature has the highest

value and it also outperforms other classification results in accuracy and FPR. Overall our proposed method's results are close enough to the well-used entropy method in DDoS literature.

Box Cox transformation's effect can be clearly seen from the Table 5.13. Figure 5.23 through Figure 5.25 are the bar plot representation of the Table 5.13.

Table 5.13. Thresholding and NB classification results before and after Box Cox.

	TP	FP	TN	FN	TPR(%)	FPR (%)	Accuracy (%)
NB Skew. BFBXCX	71	9	108	4	94.6667	7.6923	93.2292
NB Skew. AFBXCX	73	11	106	2	97.3333	9.4017	93.2292
NB Kurto. BFBXCX	56	2	115	19	74.6667	1.7094	89.0625
NB Kurto. AFBXCX	68	6	111	7	90.6667	5.1282	93.2292
NB Comb. BFBXCX	67	4	113	8	89.3333	3.4188	93.75
NB Comb. AFBXCX	71	8	109	4	94.6667	6.8376	93.75

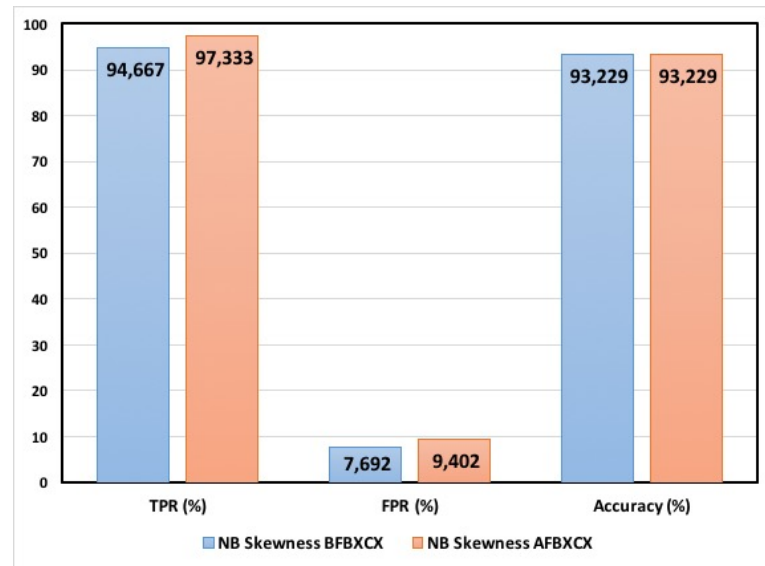


Figure 5.23. Bar plot NB classification results skewness BF and AF BXCX.

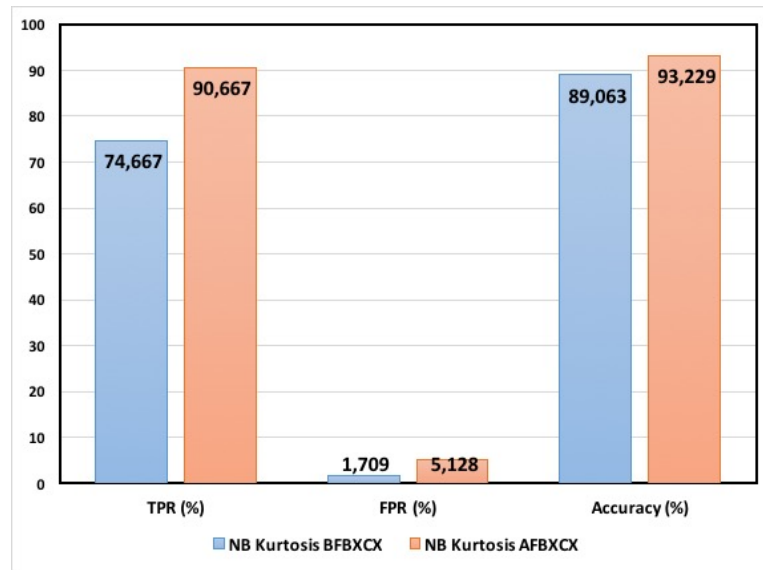


Figure 5.24. Bar plot NB classification results kurtosis BF and AF BXCX.

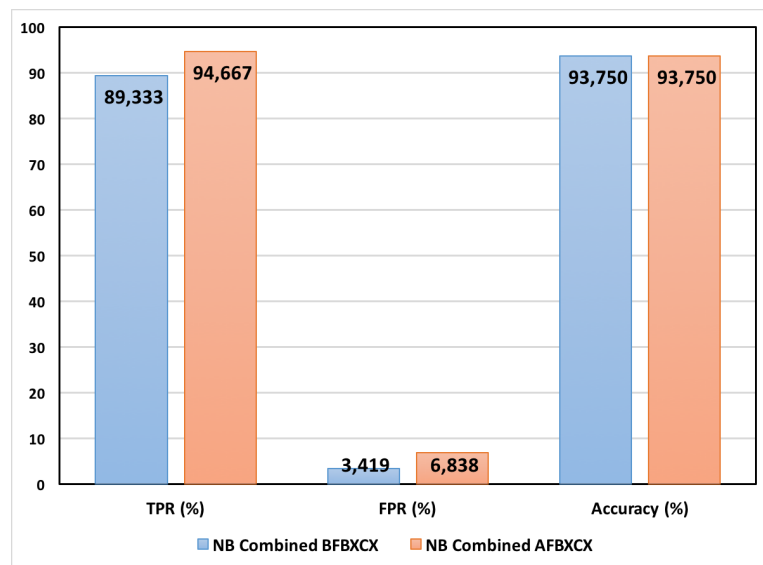


Figure 5.25. Bar plot NB classification results accuracy BF and AF BXCX.

As it is seen from the results above; Box Cox transform increased the overall true positive rate. Attack detection rate increases whereas false classification of the normal data is also negatively effected. Overall accuracy did not change except kurtosis feature. Kurtosis's accuracy increased with Box Cox transformation.

6. CONCLUSION

In our study, a time series analysis of DDoS attacks was investigated by sampling the incoming traffic in each 1 ms. The obtained time series were further divided into 128-length window to create the dataset. For each row in the dataset the skewness and kurtosis parameters are found. These two parameters were used as features to differ attack from normal traffic. To the best of our knowledge this is the first work that used Skewness and Kurtosis in DDoS attack analysis. Skewness and Kurtosis parameters contain useful information for statistical aspect. CAIDA 2007 and 2008 were used to train and evaluate our approach. By testing the normality, we found that both skewness and kurtosis comes from log logistic distribution and for classification of attack and normal traffic Naive Bayes was employed which assumes feature independency.

Later on Box Cox transformation was applied to the obtained dataset for statistical improvement of the attack traffic detection. Box Cox transformation improved the True positive rate in all NB classification and for the Kurtosis feature, it also improved the Accuracy in addition to TPR. KNN was suggested as an alternative classification method to NB. Apart from Naive Bayes and kNN classification method, simple thresholding method was also implemented and depending the results it can be used as an initial step of attack detection in IDS. Overall kNN resulted better than NB because of the reason that CAIDA has correlation in the data which was ignored by NB for independency assumption. A very well known Entropy feature was added to compare proposed method classification results.

By using time series parameters a simple, low cost and powerful detection mechanism for Intrusion Detection Systems is suggested. Third and fourth moment of the data are taken into account as skewness and kurtosis parameters and they give useful statistical information, moments of higher degrees could give better classification performance for future study.

REFERENCES

1. “Symantec Internet Security Threat Report”, Vol. 1, <https://www.symantec.com>, accessed at December 2017.
2. Ramin Fadaei Fouladi, C. E. K. and E. Anarim, “Frequency Based DDoS Attack Detection Approach Using Naïve Bayes Classification”, *Telecommunications and Signal Processing (TSP)*, pp. 104–107, Institute of Electrical and Electronics Engineers (IEEE), June 2016.
3. Zhijun WU, H. Z., Chen WANG, “Research on the Comparison of Flood DDoS and Low-rate DDoS”, *International Conference on Multimedia Technology*, pp. 5503 – 5506, 2011.
4. Charalampos Patrikakis, O. Z., Michalis Masikos, “Distributed Denial of Service Attacks - The Internet Protocol Journal”, <http://www.cisco.com>, accessed at December 2017.
5. Powell D., S. R., *Malicious and accidental fault tolerance for internet applications conceptual model and architecture*, Tech. Rep. D2, Malicious and accidental fault tolerance for internet applications (MAFTIA), November 2001.
6. Axelsson, S., *Intrusion Detection Systems: A Survey and Taxonomy*, 2000.
7. Herve Debar, A. W., Marc Dacier, “Revised Taxonomy for Intrusion Detection Systems”, *Annals of telecommunications*, pp. 361–378, July 2000.
8. Herve Debar, A. W., Marc Dacier, “Towards a Taxonomy of Intrusion Detection Systems”, *Computer Networks*, pp. 805–822, 1999.
9. AlEroud A, K. G., “Toward zero-day attack identification using linear data transformation techniques”, *Institute of Electrical and Electronics Engineers (IEEE)*

- 7th international conference on software security and reliability (SERE'13)*, pp. 159–168, Washington, D.C., June 2013.
10. Marty, R., “Snort Intrusion Detection System”, <https://www.snort.org>, accessed at December 2017.
 11. Chirag Modi, B. B. H. P. A. P., Dhiren Patel and M. Rajarajan, “A Survey of Intrusion Detection Techniques in Cloud”, *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp. 42–57, 2013.
 12. Hung-Jen Liao, Y.-C. L., Chun-Hung Richard Lin and K.-Y. Tung, “Intrusion detection system”, *A comprehensive review. Journal of Network and Computer Applications*, Vol. 36, No. 1, pp. 16–24, 2013.
 13. O. Depren, E. A. M. K. C., M. Topallar, “An Intellegent Intrusion Detection System (IDS) for Anomaly and Misuse Detection in Computer Networks”, Elsevier (Editor), *Expert Systems with Applications*, Vol. 29, pp. 713–722, 2005.
 14. Ramin Fadaei Fouladi, C. E. K. and E. Anarim, “Statistical Measures: Promising Features for Time Series based DDoS Attack Detection”, European Signal Processing Conference (EUSIPCO), September 2017.
 15. Sungkornsarun LONGCHUPOLE, R. V., Noppadol MANEERAT, “Anomaly Detection Through Packet Header Data”, *International Conference on Information and Communication Systems (ICICS)*, 2009.
 16. Luo Hua, Y. X.-m., Hu Guang-min, “Global Detection of DDoS Attack Based on Time and Frequency Analysis”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2007.
 17. Haiqin Liu, M. S. K., “Real Time Detection of Stealthy DDoS Attacks Using Time Series Decomposition”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2010.

18. Xiaozhe Wang, R. H., Kate Smith Miles, “Rule Indication for forecasting method selection: meta-learning the characteristics of univariate time series”, *Neurocomputing*, Vol. 72, No. 10-12, pp. 2581–2594, June 2009.
19. Chen-Mou Cheng, K.-S. T., H.T. Kung, “Use of Spectral Analysis in Defense Against DoS Attacks”, Institute of Electrical and Electronics Engineers (IEEE) Global Communications (GLOBECOM), 2002.
20. Alefiya Hussain, C. P., John Heidemann, “A Framework for Classifying Denial of Service Attacks”, *Association for Computing Machinery’s Special Interest Group on Data Communications (ACM SIGCOMM)*, 2003.
21. Ramin Fadaei Fouladi, E. A., Tina Seifpoor, “Frequency characteristics of DoS and DDoS attacks”, *Sinyal İşleme ve İletişim Uygulamaları (SIU)*, April 2013.
22. Jaswinder Singh, K. K., Monika Sachdeva, “Detection of DDoS Attacks Using Source IP Based Entropy”, *International Journal of Computer Science Engineering and Information Tehnology Research (IJCSEITR)*, Vol. 3, No. 1, pp. 201–210, March 2013.
23. Sidharth Sharma, S. K. J., Santosh Kumar Sahu, “On Selection of Attributes for Entropy Based Detection of DDoS”, *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1096–1100, 2005.
24. Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004.
25. Henry C. THODE, J., *Testing for Normality*, Marcel DEKKER Inc, 2002.
26. S. S. SHAPIRO, M. B. W., “An Analysis of Variance Test for Normality”, *Biometrika*, Vol. 52, pp. 591–611, 1965.

27. B.W. Yap, C. H. S., “Comparisons of Various Types of Normality Tests”, *Journal of Statistical Computation and Simulation*, Vol. 81, No. 12, pp. 2141–2155, December 2011.
28. Vilijandas Bagdonavicius, M. S. N., Julius Kruopis, *Non-parametric Tests for Complete Data*, John Wiley and Sons, 2011.
29. Dodge, Y., *The Concise Encyclopedia of Statistics*, Springer, 2008.
30. *Matlab Statistics and Machine Learning Toolbox User’s Guide*, Mathworks, 2016.
31. Sorana D. Bolboaca, A. F. S. R. E. S., Lorentz Jantschi and D. C. Pamfil, “Pearson-Fisher Chi-Square Statistic Revisited”, *Information*, Vol. 2, No. 2, pp. 528–545, 2011.
32. AGRESTI, A., *An Introduction to Categorical Data Analysis*, Wiley Series in Probability and Statistics, John Wiley and Sons, second edn., 2007.
33. Nir FRIEDMAN, M. G., Dan GEIGER, “Bayesian Network Classifiers”, K. A. Publishers (Editor), *Machine Learning*, 29, pp. 131–163, 1997.
34. T. M. Cover, P. E. H., “Nearest Neighbor Pattern Classification”, *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21–27, 1967.
35. Jiawei Han, J. P., Micheline Kamber, *Data Mining Concepts and Techniques*, Elsevier, 2011.
36. Kotsiantis, S. B., “Supervised Machine Learning: A Review of Classification Techniques”, *Informatica*, Vol. 31, pp. 249–268, 2007.
37. H. Bhavsar, A. G., “A Comparative Study of Training Algorithms for Supervised Machine Learning”, *International Journal of Soft Computing and Engineering (IJSCE)*, Vol. 2, No. 4, pp. 74–81, September 2012.

38. M. Soundarya, R. B., “Survey on Classification Techniques in Data Mining”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, pp. 7550–7552, July 2014.
39. Sayali D. Jadhav, H. P. C., “Comparative Study of KNN, Naive Bayes and Decision Tree Classification Techniques”, *International Journal of Science and Research (IJSR)*, Vol. 5, No. 1, pp. 1842–1845, January 2016.
40. G. E. P. Box, D. R. C., “An Analysis of Transformations”, *Journal of Royal Statistical Society. Series B (Methodological)*, Vol. 26, No. 2, pp. 211–252, 1964.
41. Ian H. Witten, M. A. H., Eibe Frank, *DATA MINING Practical Machine Learning Tools and Techniques*, Morgan KAUFMANN, 3 edn., 2011.
42. Andersen, D., “Center for Applied Internet Data Analysis”, <http://www.caida.org/home>, accessed at December 2017.
43. Kunchan Lan, D. D., Alefiya Hussain, “Effect of Malicious Traffic on the Network”, *Proc. PAM*, April 2003.