

# **Applying Statistical Techniques on Traffic Features for Intrusion Detection**

A Dissertation

Submitted in partial fulfilment of requirements for the award of the Degree of

**Master of Science**

In

**Mathematics**

**Submitted By:**

Somya Sharma (2K19/MSCMAT/29)

Yash Sharma (2K19/MSCMAT/06)

**Under the guidance of:**

Dr. Anshul Arora



**DEPARTMENT OF APPLIED MATHEMATICS  
DELHI TECHNOLOGICAL UNIVERSITY**

Shahbad Daultapur, Main Bawana Road, Delhi-110042, India

May, 2021

## DEPARTMENT OF APPLIED MATHEMATICS

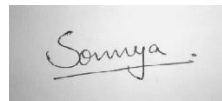
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Bawana Road, New Delhi-110042

### DECLARATION

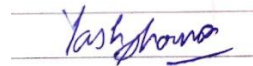
We hereby declare that the work presented in this report entitled “Applying Statistical Techniques on Traffic Features for Intrusion Detection”, in fulfilment of the requirement for the award of the degree Master of Science in Mathematics, submitted in the Applied Mathematics Department of Delhi Technological University, New Delhi, is an authentic record of our own work carried out during our degree under the guidance of Dr. Anshul Arora.

The work reported in this has not been submitted by us for award of any other degree or diploma.

Date: 23<sup>rd</sup> May 2021



Somya Sharma (2K19/MSCMAT/29)



Yash Sharma (2K19/MSCMAT/06)

## **DEPARTMENT OF APPLIED MATHEMATICS**

DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Bawana Road, New Delhi-110042

### **CERTIFICATE**

This is to certify that the Project work entitled “Applying Statistical Techniques on Traffic Features for Intrusion Detection” submitted by Somya Sharma and Yash Sharma in fulfilment for the requirements of the award of Degree Master of Science in Mathematics is an authentic work carried out by them under my supervision and guidance. To the best of my knowledge, the matter embodied in the Dissertation has not been submitted to any other University / Institute for the award of any Degree.



Date: 23<sup>rd</sup> May, 2021

Anshul Arora  
Assistant Professor  
Department of Applied Mathematics

## **DEPARTMENT OF APPLIED MATHEMATICS**

DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering) Bawana  
Road, New Delhi-110042

### **ACKNOWLEDGEMENT**

We express our sincere gratitude to Dr. Anshul Arora (Assistant Professor, Department of Applied Mathematics), for his valuable guidance and timely suggestions during the entire duration of our dissertation work, without which this work would not have been possible. We would also like to convey our deep regards to all other faculty members of the Department of Applied Mathematics, who have bestowed their great effort and guidance at appropriate times without which it would have been very difficult on our part to finish this work. Finally, we would also like to thank our friends for their advice and pointing out our mistakes.

## ABSTRACT

In this age when network attacks and malicious activities seem to be at their peak, cybersecurity plays a key role in detecting network intrusion and prevention of illegitimate access to one's data. In this thesis, we emphasize on the need of Network Security. Needless to say, network attacks can cause immense financial and practical loss to the companies, associations or even individuals too. It is often noticed that antiviruses and firewalls, which are used to provide enough network security, are not at any level adequate to guarantee the insurance of an organization now against these everchanging attacks. These conventional tools have been found to be unsuccessful in defending network systems satisfactorily from increasing refined attacks and malwares. This kind of situation requires smart counter measures to keep up the security of networks and important systems. Hence, in this work, we aim to build effective system to detect intrusions based on network traffic.

Chapter 1 of this thesis gives an insight about the meaning and role of intrusion detection system. Inspecting network traffic and computer cases to recognize malicious or unauthorized activities is a process called " intrusion detection". Intrusion Detection System (IDS) can be defined as any device or any software design application whose motive is to direct an intrusion detection. IDSs can screen exercises inside the secured network and not exactly at its perimeter. In contrast to a firewall, IDSs just have an inspection job. Further, we classify all the different types of IDSs namely signature based, anomaly based and hybrid detection. The main features that an effective Intrusion Detection System ought to have are effectiveness, ability to adapt and extensibility. We conclude chapter 1 with some limitations and problems raised by the existing systems. Chapter 2 revolves around the works proposed in the literature for network traffic- based intrusion detection. We review more than 50 papers in this chapter.

In Chapter 3, we introduce all the 12 features with their names and their meanings that are being used in our research. Then we use statistical tests to rank these features in order of their efficiency to detect network intrusion. The main objective of our work is to obtain a certain set of features that will show higher accuracy than all the other features individually

or combined. For this research, we use two different statistical tests namely ANOVA test and CHI-SQUARE test. ANOVA test stands for analysis of variances. It is a statistical method which verifies the impact of a number of factors by comparing the average or means of various sample data whereas the Chi Square statistic is used to determine if the variables of different categories defined are independent of each other or not. Then we move on to the machine learning classifiers and their purpose. For our research we use three different types of machine learning classifiers namely, Decision tree, SVM and Random Forest. The features are ranked in a way such that the feature fixed at the bottom will be the least efficient and vice versa. In this way, we prepare three columns: One for ANOVA, second for Chi-Square normal and another for Chi-Square Malware. Further, we prepare another table taking all the single features one at a time and separately apply all three of the machine learning classifiers. Now to take into consideration the possibility that the combination of these features might be even more effective, we make different combinations of features and after applying all three machine learning classifiers again we obtain 3 different sets of columns.

Further, the above method is repeated on these 3 columns and we obtain a separate set of features in the end which are best in terms of network intrusion. In Chapter 4, we present all the tables, calculations and proofs to reach at a conclusion that according to our research a particular combination of 5 features namely “Bytes received”, “Bytes sent”, “Time interval between packets sent”, “Packet size sent” and “Packet size received” gives the highest accuracy of 99.49% after applying Decision Tree amongst all possible features and combinations. In Chapter 5, we conclude the thesis with future work directions.

# Contents

Declaration .....	(i)
Certificate .....	(ii)
Acknowledgement .....	(iii)
Abstract .....	(iv)
List Of Tables .....	(v)
List Of Figures .....	(vi)
Chapter -1 .....	1-4
1.1 The Security Problem .....	1
1.2 Intrusion Detection System .....	2
1.3 The Role of Intrusion Detection.....	3-4
CHAPTER-2 Related Work.....	5-11
2.1. Related Work in Intrusion Detection .....	5-11
2.2 Summary .....	11
CHAPTER-3 METHODOLOGY .....	12- 18
3.1 Features Names .....	12
3.2 About Feature Ranking Methods .....	12
3.2.1 ANOVA Test .....	12-13
3.2.2 Chi Square Test .....	13-14
3.3 Machine Learning Classifiers .....	14-18
3.3.1 Decision Tree .....	15-16
3.3.2 Support Vector Machine (SVM) .....	16
3.3.3 Random Forest .....	16
3.4 Proposed Approach .....	17
3.5. SUMMARY .....	17-18

<b>CHAPTER – 4 RESULTS .....</b>	<b>19- 25</b>
<b>4.1. Ranking of Features .....</b>	<b>19-20</b>
<b>4.2. Detection Results with Individual Features.....</b>	<b>20-21</b>
<b>4.3. Detection Results with Propose Approach.....</b>	<b>21-25</b>
<b>4.4. Summary .....</b>	<b>25</b>
<b>5. Conclusions .....</b>	<b>26-27</b>
<b>6. References .....</b>	<b>28-30</b>



## List of Tables

1.	Table 4.1.1: Test Values .....	19-20
2.	Table 4.1.2: Ranking of Features .....	20
3.	TABLE 4.2: Single Feature .....	21
4.	TABLE 4.3.1 ANOVA Test Results .....	21-22
5.	TABLE 4.3.2 CHI SQUARE NORMAL Test Results .....	23
6.	TABLE 4.3.3 CHI SQUARE MALWARE Test Results .....	24
7.	TABLE 4.3.4: Accuracy Of Final Set .....	25

## List Of Figures

1. Figure 3.2.2.1: Chi Square Formula.....13

# CHAPTER 1

## INTRODUCTION

Nowadays, network security is perhaps the most basic thing that the enterprises and associations need to deal with. An ever-increasing number of attacks are witnessed in the expertly corporate organizations network or in private organizations network. New dangers are arising each day and piles of old dangers stay current. The result of an effectively performed network assault can be exceptionally pivotal, either in business or individual organization conditions. Network assault can cause immense financial and practical loss in a large number of cases. With the consistent development of the Internet, computer attacks are expanding in numbers as well as in variety, ransomware are on the ascent more than ever, and zero-day misuses become so important that they are acquiring media inclusion. Antiviruses and firewalls are not, at this point adequate to guarantee the insurance of an organization, which ought to be founded on numerous layers of safety. Quite possibly the main layers, intended to secure its objective against any likely assault through a consistent monitoring of the framework, is given by an Intrusion Detection System (IDS).

### **1.1.The Security Problem**

In the cutting-edge danger scene, the internet and computer networks have been a huge security centre throughout the last twenty years. The cyberspace security is a collection of different tools and methodology expected to safeguard computers, networks, programming software and information records from attack, illegitimate access, modification or harm. Intrusions are known as tries to sabotage the privacy, originality, or accessibility of a system. Recognition of intrusions in distributed, asset confined and constantly advancing conditions is a troublesome task. Within the sight of such complications, people are not normally ready to deliver satisfactory planned decisions. Moreover, intruders profile document, goals and abilities have additionally changed impressively. In the past years, cyber intruders were considered to be youthful socially inaccessible people inspired by few things including interest, the adventure of the illegitimate activities and additionally (for further developed hackers) peer acknowledgment. While most were gifted, these interlopers didn't have adequate financial resources to produce advanced attacks. Though, nowadays erudite attacks and inspirations have created. For instance, we presently see the utilization of Advanced Persistent Threats using numerous highly advanced methods. Risks have lately become more varied and it is hard for security groups to stay aware of the everyday arriving of loop holes, dangers, attacks and comebacks using suitable countermeasures.

Security dangers against network foundation can be sorted as active and inactive. Concerning active dangers, attackers attempt to disrupt network operations.

Though, for inactive threats, intruders stay covered up and focus around acquiring the data exchanged. Conventional safety tools, like firewalls, security strategies, access control and encryption have been unsuccessful to defend network and system satisfactorily from increasing refined attacks and malware. Moreover, security groups have significant challenges taking care of enormous data. The key point is minimising the danger of loss to the extent is viable. The conditions require more smart counter measures to keep up the security of networks and important systems. Machine learning (ML) has had accomplishment in numerous spaces of software engineering, where it was accepted in true applications like product recommendations, optical character recognition. Subsequently, security teams considered embracing ML to decrease the confusion of the conflict with attackers. The programmed arrangement of malicious practices on the web holds huge potential for improving security. Furthermore, the extricated information permits disclosure of new assault techniques, intrusion situations and attacker's targets and procedures, and can upgrade the capacity to recognize assaults and genuine solicitations.

## **1.2. Intrusion Detection System**

Inspecting network traffic and computer cases to recognize malicious or unauthorized activities are a process called " intrusion detection ". Each device or software design application whose objective is to direct an intrusion detection is considered as an Intrusion Detection System (IDS). IDSs can screen exercises inside the secured network and not exactly at its perimeter. In contrast to a firewall, IDSs just have an inspection job: they can't make a move to obstruct doubtful movements and consequently need an executive to deal with their alarms. IDSs can likewise be characterized by the detection strategy they use. They fall into three classifications: signature-based detection, anomaly-based detection, and hybrid detection.

- Signature-based detection (also called as " misuse detection ") accompanies a database of known assault marks. Its distinctions monitored information and the signature database. A misuse detection IDS checks the information stream for the presence of an attack pattern like a classic antivirus. This mark can appear as a sequence of bytes or characters. To be effective, the database of this sort of IDSs should be updated consistently.
- Anomaly detection tries to learn a "normal" or "expected" behaviour of the system. Any deviation from this conduct is considered as a possible attack and will generate an attention. This strategy doesn't need updates or even the presence of a database. It can distinguish doubtful attacks yet in addition makes a lot

of false positives which are hard to measure. It is additionally harder to gather data about the attack since it isn't distinctly recognized by a signature.

- Hybrid detection consolidates the two reactions for alleviate shortcomings of each category: anomaly detection then misuse detection, misuse detection then anomaly detection, or both simultaneously. The objective is to identify known attacks with their signatures, and to utilize Anomaly location to distinguish obscure interruptions normally detection to identify unknown intrusions.

### **1.3.The Role of Intrusion Detection**

A large number of detection and protection approaches have been given in the past but the core element of a good defence system is an Intrusion Detection System (IDS), which provides proper attack detection before any reaction. The main objective of an IDS is to detect intrusions before they seriously damage the network. The term intrusion literally means any unauthorised attempt to access the parts of a network with the objective of making the system corrupt.

The main features that all the IDS ought to have are effectiveness, ability to adapt and extensibility. Effectiveness refers to the flexibility of a system in detecting slight variations of best-known attacks associated with the network. Extensibility implies that an IDS can be easily adjusted according to various environments.

A majority of planned IDSs within the past are of signature detection type which is all due to their effectiveness in detection of best-known attacks. Two key examples of famous and generally used ASCII text file signature-based tools are Snort and ClamAV, each have their own information of signatures (more than 4000 rules in Snort information and over 800,000 in ClamAV repository). Despite the fact they are very efficient these systems raise 2 necessary problems for more research:

- An individual is responsible for the creation of signatures and the need to change the information and as a result the price of developing and maintaining this information is a vital issue. They conjointly suffer from the pause between facing new attacks and manually change the signatures.
- they are doing not work against new and unknown attacks.

To automatise the method of signature generation, the utilization of machine algorithms may be a promising technique with adaptation, fault tolerance, high machine speed and error resilience within the face of vociferous data characteristics and moreover to update the information of signatures for a dynamic environment, associate progressive learning approach are often needed to repeatedly learn new attacks. To summarise, the extreme manual method of signature creation for IDSs, change the non-adaptability of them to dynamically dynamical surroundings and considering extensibility as a vital feature for the planning of IDSs introduce a replacement direction of research and also the motivation for this thesis.

**Contributions of the Thesis:** Keeping these limitations in mind, in this Thesis, we have proposed an anomaly-based intrusion detection model. The main contributions of the Thesis are summarized below:

1. We extracted the network traffic features from the traffic files of normal and intrusion traffic.
2. We applied the statistical techniques of the Anova and Chi-Square Test on the set of traffic features intending to rank the traffic features. We obtained the ranking of features with these statistical tests.
3. We proposed a novel algorithm for intrusion detection by applying machine learning classifiers to the ranked features.

**Organization of the Thesis:** The remainder of the Thesis is structured as follows. We discuss the related works in the field of intrusion detection in Chapter 2. We explain the detailed methodology of our proposed model in Chapter 3. We review the results obtained from the proposed model in 4 and conclude with future work directions in Chapter 5.

## CHAPTER-2 RELATED WORK

In this chapter, we discuss the works proposed in the literature for network traffic-based intrusion detection, in Section 2.1. Further, we summarize the chapter in Section 2.2.

### 2.1. Related Work in Intrusion Detection

The authors in [1] presented an approach named Clustered Generalization oriented Ensemble Learning Model (CGELM) for predicting the network attacks and malware. Due to the increase in the rate of network anomalies, a hybridized multi-model system was proposed with Stack Ensemble methods, and the authors achieved 98.93% detection accuracy. The authors in [2] focused on creating a more realistic data set and building a Multi-layer Stack Ensemble (MLS) model for the intrusion detection system. They exploited the strengths of various base-level model predictions to create a more robust meta-classifier with multiple techniques 1.

To enhance the security and improve the detection ability of malicious intrusion behaviour in a cluster of network traffic, Yang et al. [3] proposed a wireless network intrusion detection method based on improved convolutional neural network (ICNN). In their work, the network traffic data was first characterized and reprocessed, then the network intrusion traffic data was modelled by ICNN. The test results indicated that the method proposed in this paper had good detection accuracy. Srivastava et al. [4] used machine learning algorithms for discovering weird patterns in the recently provided dataset. Zeng et al. [5] presented a lightweight mechanism with the help of deep learning for hidden traffic classification, known as Deep-Full-Range (DFR) as a replacement for the existing systems that required a burdensome analysis of various features. DFR was able to monitor and learn the patterns from raw traffic without needing the aid of manual intervention. This kind of a method required a significantly lower amount of storage. The authors in [6] introduced a real-time network intrusion detection system that solved the problem of low accuracy, low detection efficiency, high false-positive rate, etc. Their model was based on natural language processing technology and deep learning technology for intrusion detection with traffic features. The authors in [7] proposed CBIDP (Cluster-Based Intrusion Detection Planning), an effective clustering algorithm for intrusion detection. This technique proved to be really useful for detecting intrusions . It also helped in overcoming the other drawbacks like traffics, connections, and node mobility on the network.

Yang et al. [8] proposed a novel malicious SSL traffic detection method that resembled SSL records and inspected the characteristics of records for deep learning-based classification. The experimental results have shown that this approach has good capability of separating between benign and malicious traffic flows on an encrypted SSL channel. The authors in [9] proposed a flow-based approach to detect abnormal Neighbour Discovery Protocol, i.e., NDP traffic behaviour which is considered as an indicator of the presence of NDP-based attacks. This approach relied on flow-based network traffic representation and the adoption of the entropy algorithm to detect the randomness in the network traffic. The authors in [10] talked about a novel intrusion detection model which is kind of double action mechanism with meta classifiers. The model included an ensemble learning meta classifier for stacking through cross-validation to prevent overfitting.

Nie et al. [11] have applied deep learning based on GPA in the presence of a feature selection module for intrusion detection which aims at a single type of attack. The authors in [12] applied an anomaly-based machine learning algorithm to detect DDOS attacks in the networks. The experimental results proved that their model is better than other signature-based intrusion detection techniques. On similar lines, Kapil et al. [13] applied several machine learning classifiers at different network layers for intrusion detection. Naseer et al. [14] attempted to provide a shorter list of features regarding their role in detecting intrusions. They used a 5-step method in which the primary steps were all about dropping basic features while the couple of end steps were performed to detect the significant ones. The authors in [15] designed a framework consisting of different deep reinforcement algorithms applied on different data sets. The authors interpreted the trend of traffic flows by extracting statistical features on the prior network traffic for network prediction.

The authors in [16] emphasized the need to detect anomalous network traffic to protect network security with the help of multiple learning automata. They applied such multiple learning automata to remove redundant traffic features for detection. The authors in [17] described the need and importance of security and cyberattacks in the Internet of Things (IoT) framework and proposed an ensemble intrusion detection method to detect malicious activities in IoT design. Shenfield et al. [18] presented the idea to use artificial neural networks suited for use in deep packet inspection-based intrusion detection systems to detect harmful network traffic. The neural network system proposed in this paper displayed good accuracy with the potential to enhance the utility of intrusion detection systems. The authors in [19] described a Gated Recurrent Unit (GRU) and Long Short-Term Memory unit (LSTM), both being the variants of the



Recurrent Neural Network (RNN) for intrusion detection on network traffic. Zavrak et al. [20] emphasized the importance of a flow-based intrusion detection system based on the small amount of traffic data due to the large growth in network traffic. Particularly, Autoencoder and variational autoencoder methods were put to use to detect unknown attacks using the help of flow features. The authors in [21] modelled network traffic as time series with the use of supervised learning methods for intrusion detection. To measure the effectiveness of such methods, the synthetic ID data set was formed using the KDDCup 99. After various time series tests, it was concluded that Convolutional Neural Network (CNN) and its variant architectures performed significantly well when compared with the classical machine learning classifiers.

The authors in [22] described the need for extra research efforts to put a leash on to the rapid increase in cyber-attacks. Similar to some other works, they also applied deep neural networks and other machine learning classifiers for real-time data fusion problems in intrusion detection. Atli et al. [23] described an intrusion detection system based on modelling distributions of network statistics and Extreme Learning Machines (ELM). The authors in [24] did a study to find the most efficient model by comparing several machine learning classifiers. For their evaluation, they used the NSL-KDD dataset. They found that ANN-based machine learning with wrapper feature selection is more efficient than SVM when it comes to classifying network traffic. The authors in [25] proposed a PCCN (Parallel Cross Convolutional Neural Network) based method on several multi-class flow features for detecting abnormalities in the traffic. The results demonstrated that the proposed PCCN algorithm was better than other machine learning algorithms in terms of detection accuracy. Vu et al. [26] proposed a deep learning-based model named Multi-distributed Variational Auto Encoder (MAVE) for intrusion detection. The authors labelled the data samples into the Kullback-Leibler (KL) loss function of VAE and further applied various machine learning classifiers including SVM, Decision Trees, and Random Forest for the detection.

Santos et al. [27] proposed an IP flow-based (IDS) to recognize and secure Internet of things data from all types of threats. The authors collected IP flows from an IoT network to analyse them and for further detecting threats, malware activities and other types of irregularities at various IoT structures. This type of system was created to consider the IoT network architecture. The authors in [28] talked about the growing need for an intrusion detection system and cybersecurity as a primary defence mechanism that can adapt and secure the computing infrastructures from the ever-changing sophisticated threat landscape. This work presented an unsupervised deep learning approach for intrusion detection and

trained their classifier in two separate stages. Firstly, with a One-Dimensional Convolutional Auto Encoder (1D CAE) and then with a One-Class Support Vector Machine (OCSVM).

The authors in [29] talked about NID being a crucial method of network traffic security. It can monitor and recognize the flow of external intrusion in the network, and carry out early precautionary system. The NID based on FCMA transforms the system into the classification of network traffic, establishes a network traffic classification model, and classifies the network traffic by using machine learning classification algorithms, and detects the external intrusion traffic. Dasari et al. [30] used meta-heuristic association scale to derive a minimum threshold value for the transaction and thus the ensemble classifier is used to analyse the transaction is normal or attack. The Classifier used by the author is based on drift detection which has the potential to analyse the request at stream level.

The problem of intrusion detection in MANET's was explored in [31] and then a novel system network information-based moderation model has been proposed. The system proposed used time variant snapshots to detect routing attacks. The authors in [32] proposed novel intrusion detection method with a fully automatic threat sign generation capability. This view exploited a Hanif for traffic data analysis to break an attack scenarios database, used to detect potential intrusions. DARPA'99 and UNSW-NB15 traffic was used to evaluate the approach proposed. The authors in [33] introduced a lightweight machine learning based intrusion detection method with great results for resource limited IOT wireless networks.

Multi structure modular deep neural network model has been proposed to bring down the false positive rate of anomaly money type detection methods by Atefni et al. [34]. The experimental results showed a crazy response as high as 100% for network level attacks compare to related works. Gupta et al. The authors in [35] proposed a multi demeanour fusion-based IDS where stream data mining is done by using ST-SR (stochastic relaxation). Methods used in this paper can easily predict the quality of service of each network during the network traffic and it also enabled the user to get connected with a secured network which held a high packet delivery ratio, less packet loss and high output.

A deep multi-scale convolution neural network is being proposed by Wang at al. [36] for network intrusion detection. The author used two different testing sets of which second one is more difficult to be classified. Experimental results have shown that its accuracy rate has reached up to 98%, it had high intrusion detection accuracy low false alarm rate which can help to overcome the limitation of using traditional detection methods. Hsu et al. [37] used a deep learning scheme to overcome difficulties faced by applying machine learning to IDS. The proposed model in this paper is based on convolutional neural network CNN layers using long short term memory layers called CNN-LSTM classify every single-track

network. This data set included two testing sets. The work proposed in [38] is based on univariate ensemble feature selection techniques which is used for the selection of valuable reduced feature set from the given intrusion data set. An ensemble method proposed here effectively classifies whether the network traffic behaviour is normal or attack. The authors in [39] developed a combining classified model which is based on tree-based algorithms for network intrusion detection. The experimental values showed 89.24 % detection accuracy which was achieved by using the combination of random tree and NB Tree algorithm based on the sum rule scheme. It proved to be better than individual random tree algorithm.

Caligari et al. [40] used an approach using different families of Markovian models called high order and nonhomogeneous Markov chains for modelling network traffic running over TCP. This statistical approach provided an edge over other statistical technique used before. The authors in [41] presented a new online and real time unsupervised network anomaly detection algorithm: ORUNADA. The solution in this paper is based on a discrete time sliding window to update continuously the future space and an incremental grid clustering to detect rapidly the anomalies. The evaluation showed that ORUNADA can process online large network traffic while ensuring a low action delay and good detection performance. The results have proved to be better in terms of TPR and FPR existing techniques reported earlier. The authors in [42] used the traffic model and developed it to describe the dynamic characteristic property of network traffic in visual sensor networks. On the basis of this model the optimal feature set for traffic pattern learning can be extracted. Then self-organizing map is employed to learn traffic patterns and detection intrusion further it is also devised to accelerate the training process of HSOM to learn the pattern of attacks better.

The authors in [43] paper analysed intrusion detection using flow-based IDS and cross layered approach. The proposed detection method had a two-phase approach. In the first phase, flow-based anomaly detection method was used to find potential anomalies in the network. During second phase cross-layer features are correlated to narrow down the possible attacks. The authors in [44] paper discussed about the problem of approximate reduction of non-deterministic automata that can appear in NIDSs and a method to resolve it. The authors designed an approximate reduction procedure that achieves a great size reduction. They implemented their approach and evaluated its uses on NIDSs. Results indicated that the method can be highly effective in practice. The authors in [45] emphasized the need to efficiently identify any intrusion in the network due to the current increase in the network attacks exponentially. This research work introduced a deep learning-based system for hybrid intrusion detection and signature

generation of unknown web attacks known as D-Sign which is fully capable of detecting and creating attack signatures with great accuracy and specificity. D sign is known to have significantly low False positives and false negatives and moreover it generates signatures which helps in reducing the damage due to network attacks.

The authors in [46] paper proposed an intrusion detection method based on multiple-kernel clustering (MKC) algorithms which can find specific correspondence between traffic attributes and attack behaviours and increase the detection accuracy. The proposed method completes the absent traffic property through similarity calculation. In the conclusion of the paper, the results showed that this method can improve the clustering accuracy of incomplete sampled data. Moreover it can decrease the sensitivity of the anomaly detection model to the choosing of traffic feature. The authors in [47] proposed Improved Siam-IDS (I-SiamIDS) an algorithm based view, which is a two-layer ensemble for handling class imbalance problem. It recognizes both majority and minority classes at the algorithm-level without using any data-level balancing methods. The first layer of I-Siam uses an ensemble of binary b-XG Boost, Siamese Neural Network (Siamese-NN) and Deep Neural Network (DNN) for hierarchical filtration of input samples to identify attacks. These attacks are then sent to the second layer of I-Siam IDS for classification into different attack classes using multi-class extreme Gradient Boosting classifier (m-XG Boost). I-SiamIDS showed significant improvement in terms of Accuracy, Recall, Precision, F1-score and values of Area Under the Curve (AUC) for both NSL-KDD and CIDDS-001 datasets. The authors in [48] utilized Deep learning strategies to actualize an anomaly based Novel Intrusion detection system. It's a known fact that anomaly-based methodologies are productive and efficient yet signature-based detection is favoured for general use of intrusion detection systems. The main reason why ventures don't support the idea of using anomaly-based intrusion detection systems can be understood by approving the efficiencies of every one of the strategies.

The authors in [49] emphasized on the exponential growth of services in the internet with rapid development of technologies which directly results in increase in network traffic. Increase in traffic means increase in attacks. One of the many approaches found to stop network attack is the machine learning approach which relay on features to extract the knowledge from the traffic. In such a scenario the performance depends on the features extracted at the packet level. This paper introduced a set of unique flow features that are defined to extract the traffic from the network at flow level and train the system with diversity of the flow characteristics identified using Kolmogorov–Smirnov Test (K–S Test). The author in [50] talked about the cloud computing. It is an innovative paradigm technology that is quite

known for its versatility and qualities such as being cost efficient and reliable. Although these services are easily accessible but they are exposed to intrusion attempts. This paper focused on the security service offered to the cloud tenants to restrain intruders. The authors intended to provide a flexible, on-demand, efficient and pay-as-you-go multi-tenant intrusion detection system as a service that targets the security of the public cloud.

## **2.2. Summary**

In this chapter we have reviewed the related work proposed in the past in the field of intrusion detection based upon network traffic behaviour. We have reviewed more than 50 papers in this chapter.

## **CHAPTER-3 METHODOLOGY**

This chapter includes all the necessary information that one should be aware of to understand this paper. We describe features for intrusion detection, statistical tests, machine learning classifiers followed by the approach we use.

### **3.1. Features Names**

Table 3.1: Summarizes The Features Names And Notations That Are Being Used Throughout The Thesis.

<b>Sr.no.</b>	<b>Features</b>	<b>Feature Notations</b>
1.	Average Packet Size	F1
2.	Time interval between packets sent	F2
3.	Time interval between packets received	F3
4.	Flow Duration	F4
5.	Ratio between Incoming to outgoing bytes	F5
6.	Ratio of Incoming to outgoing packets	F6
7.	Packet size sent	F7
8.	Packet size received	F8
9.	Bytes sent	F9
10.	Bytes received	F10
11.	Number of packets sent	F11
12.	Number of packets received	F12

### **3.2 About Feature Ranking Methods**

Here, in this work, we have used statistical test to rank the features obtained. We used two different statistical tests to get the ranking: ANOVA and Chi-Square test. We discuss them now.

#### **3.2.1 ANOVA Test**

It stands for Analysis of Variances between the two population or data we have. When the comparison of two or more than two groups/data is required one-way analysis of variance that is One-way ANOVA is suitable method for it. This method evaluates the relative size of variants between groups called group variance compared to the average variance within the group called within group variance.

It is a statistical method which verifies the impact of a number of factors by comparing the average or means of various sample data. ANOVA finds out if the groups formed by the level of independent variable are statistically different by calculating if the means of the treatment levels are distinct from the overall mean of the dependent variable.

Assumptions made while performing ANOVA test

- 1) We considered that there is no dependence among the data we collected.
- 2) We considered the dependent variable is following a normal distribution.
- 3) The variances of the population or the data are equal.

### **Working of ANOVA**

It is a statistical method and hence like all of them it also works on the principle of hypothesis which is null and alternate hypothesis.

- 1) Null hypothesis in ANOVA is valid when the averages of all sample data are equal or we can say that they don't have any remarkable difference between them.
- 2) alternate hypothesis in ANOVA is valid when at least one of the sample data means varies from the other values in the sample data.

### **3.2.2 Chi Square Test**

The Chi Square statistic is used to determine if the variables of different categories defined are independent of each other or not. It can also be defined as if there is any significant difference between variables and their expected values.

**Chi square formula:**

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Figure 3.2.2.1: Chi Square Formula

where:

c=Degrees of freedom

O=Observed value(s)

E=Expected value(s)

### **Assumptions of chi-square test**

- 1) Data are considered to be independent of each other.
- 2) The data follows chi square distribution.

### **The hypothesis of this test is**

- 1) Null hypothesis says that there is no link between the original and expected data. That is independent of each other.
- 2) Alternate hypothesis says that the original and expected data are dependent of each other

### **3.3 Machine Learning Classifiers**

Classification can be visualized as two-step process the first step includes learning step and the second one is the prediction in machine learning. In the first step of learning a model is developed which is based on the given training data. Decision tree is considered as one of the easiest and most common algorithms used for classification which helps in understanding and interpretation.

Let us try to understand this with a simple example.

Spam detection in electronic mail service providers can be recognized as a classification problem, an email classifier that sorts emails to filter them by a class label : Spam or Non-Spam.

There is a huge number of applications in classification in many areas such as medical diagnosis, mail filtering etc.

### **Types of learners in classification**

- 1) Lazy learners: which type of learners waits and stores the data until testing data appears. When the testing data appears then the classification is done and the most appropriate data in the training data is then used. The prediction time is more compared to that of eager learners.
- 2) Eager learners: This type of learners first constructs a model of classification which is based on the training data provided before receiving data which is to be classified. They use a single hypothesis which can be applied to the entire data for classification. Due to which the time taken in training is more and the time taken for prediction is very less as compared to lazy learners.



In machine learning there are many classification algorithms which are now available but it is not possible to conclude which one is better than the other. The comparison is dependent on the application and nature of data available.

### **3.3.1 Decision Tree**

Decision tree builds classification models in a tree structured way. Produces if-then rules set which makes it mutually exclusive in classification. Rules can be learned sequentially with the use of training data one at a time. Whenever a rule comes into knowledge it eliminates the tuple covered by the rules. This process will be repeated on the training set and till it made sir termination condition. The final structure we get looks like a tree with nodes in leaves. All the characteristics must be categorical. Otherwise, they should be discretized earlier. Characteristics at the topmost position of the tree have more influence towards in the classification and they are identified using the information gain concept. Decision tree can contain too many branches which can reflect irregularities due to noise or outliers. Decision tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervisors learning algorithms the DT algorithm is able to solve regression and organization problems. The goal of using decision trees for constructing a training model that can be further used to predict the value of the target variable by using simple rules of decision which are concluded from the training data.

In decision tree, for predicting a class label for a record we start from root of the tree. The value of the roots directs attribute and the record attribute are compared. On the basis of this evaluation, we will follow the next branch adjacent to the value of root and move on to the next node. Decision tree classifier examples by categorizing them from root to some leaf for the terminal node with the terminal node or leaf providing the classification off the example. Every node in the decision tree will act as test case for some attribute and every e edge descending from the nodes will correspond to the probable answer to the test case. This process will be repeated recursively for every subtree rooted at new node.

Some of the assumptions we make while using DT algorithm are mentioned below:

Initially in the decision tree all of the training set is taken as the root. Feature values are preferred to be categorical. If the values given are continuous then their discretized before building the model. Distribution of records is recursive based on characteristic values. Order to insertion characteristics as root or internal node of tree is done using some statistical approach. The trees accuracy depends most on

the decision of making planned splits. The decision criteria for classification in regression trees are different.

To decide whether to split a node into two or more sub nodes decision tree uses multiple algorithms. Formation of sub nodes increase the similarity of subsequent sub nodes. We can also say that the vitality of node increases with respect to the target variable. Decision tree first splits the node on all accessible variable then it fixes the split having most alike sub nodes.

### **3.3.2 Support Vector Machine (SVM)**

SVM is a type of supervised machine learning model which uses the classification algorithm for two-group classification problems. After we provide SVM model with the training data for each category they are able to easily categorise new text. SVM can also be used for regression. Although it is used mostly in classification problems.

In SVM algorithm, each data item is mapped in an n-dimensional space with number of features equals to 'n' and value of each feature represents the value of some particular co-ordinate. The classification is performed by finding the hyperplane that distance two classes significantly. Support vectors can be observed as the coordinator of individual observation.

### **3.3.3 Random Forest**

Random Forest is an aggregate learning method for classification and regression and other applications that operates by creating a decision tree of multitude at training time. Random Forest takes the base of decision tree and eliminates its disadvantage of overfitting their training set. Random Forest output from decision tree in most cases but they are less reactive than gradient boosted trees. Although characteristics of data is also one of the factors affecting their performances. As the name implies the random Forest consists of several individual decision trees which operates as one. Each tree in the random Forest provides a class prediction and then the class with the most votes become our model's prediction. In random Forest many independently relative models of the trees operate as one which outperforms any individual constituent model. The correlation between model is the key. While some can be wrong, various other trees will be right so as a group the trees are able to move in specific direction so the prerequisite of for the random forest to execute well are:

- Include actual signal the features by which the model will use those features can perform better rather than a random guess.
- The prediction which are made by individual trees must have low correlation with each other.

### **3.4 Proposed Approach**

With respect to above section 3.2 where the statistical test has been performed, we have organized it in a manner such that the feature which is best ranked according to statistical test is capped at the top successively being followed by the one with less ranked features and so on. The least ranked feature will be at the bottom of the column. In this manner we will get three different columns first one will be of ANOVA test second will be Chi square normal and third will be of Chi square malware. Later we have created another table which has accuracy according to the machine learning algorithm. These accuracies are calculated on single features at a time. The three machine learning algorithms which we have used in the paper of Decision Tree SVM and Random Forest as mentioned above in section 3.3.

After getting the accuracy of single features for all three algorithms. The process applied for the involves the combination of the highest ranked feature with the next highest ranked feature for each statistical test (column). Every time we combine two features, we have calculated the new accuracy of the combined features using Decision Tree, SVM, Random Forest.

Here we have two possible scenarios of combining the features:

- 1) If the new accuracy so obtained is less than or equal to that of the highest ranked feature's accuracy then we will not combine that feature.
- 2) If the new accuracy so obtained is greater than that of the highest ranked feature's accuracy then we will combine that feature.

Following this step till we reach the lowest ranked feature. In this way we have obtained 3 different sets of each column. Now we have combined these 3 sets following the above method. And removing the repeated features. We get the final set of features which are the best ones to detect network intrusions.

### **3.5. Summary**

In this chapter we divided our methodology into four different sections.

- In the first section of this chapter section 3.1 includes feature description of the features used in our detection model.

- In section 3.2 we gave description of statistical tests that we used for ranking features discussed in section 3.1.
- Later we also discuss about machine learning classifiers in section 3.3 and how we use them to get accuracy of features.
- In section 3.4 we describe the approach we use in this paper to achieve best outcomes.

## CHAPTER – 4 RESULTS

In this chapter, we discuss the results obtained from the proposed work. Firstly, in section 4.1 we highlight the ranking of the features obtained from the ANOVA test and Chi Square. In section 4.2 we review the detection results with individual features. Further, in section 4.3, we discuss our approach with results in detail.

### 4.1. Ranking of Features

When applied ANOVA tests gives you the p-value (Table 4.1.1).

On applying chi-square test on normal data and the malware data separately we get their respective chi square value of each feature as follows(Table 4.1.1):

**TABLE 4.1.1: TEST VALUES**

Features	P- Value	Chi-Square value Normal	Chi-Square value Malware
<b>F1</b>	4.30E -72	2122715.611	5926304.712
<b>F2</b>	0	37495.33308	81805.3058
<b>F3</b>	0	40352.17939	70452.7675
<b>F4</b>	3E-128	19016927.82	581449.6579
<b>F5</b>	0	4445.849973	24584.32825
<b>F6</b>	8E-147	339624.5391	209980.1579
<b>F7</b>	0	541806.182	1052382.224
<b>F8</b>	1.5E-18	3249576.305	18539138.24
<b>F9</b>	0.12025	5.12082E+11	3401902043
<b>F10</b>	7.3E-12	1.94795E+11	56885444238
<b>F11</b>	1E-11	3081397.304	23331382.89
<b>F12</b>	5.7E-21	111805954.4	31027067.28

The P-value is directly proportional to rank of feature. Therefore, the features are placed in a descending order from top to bottom, where top most feature is the best ranked according to the ANOVA test and the last feature is the least in ranking. Higher the chi-square value for the feature higher it is ranked . so, we arranged them accordingly as given below Table 4.1.2 :

**TABLE 4.1.2: RANKING OF FEATURES**

ANOVA	Chi square Normal	Chi square Malware
F9	F9	F10
F2	F10	F9
F3	F12	F12
F5	F8	F11
F7	F11	F8
F11	F1	F1
F10	F4	F7
F8	F7	F4
F12	F6	F6
F1	F3	F2
F4	F2	F3
F6	F5	F5

#### 4.2. Detection Results with Individual Features

Here we have calculated accuracy for different features individually using three different algorithms (machine learning classifiers) Decision Tree, SVM and Random Forest. All the of these classifiers give different results for the highest accuracy feature. For Decision Tree and Random Forest, the result is same F10 is the highest accuracy giving feature with 99.199% while for SVM F2 remains highest

accuracy feature with 95.753%. Now if we compare between all the Algorithms, we find that F10 is the most accurate individual feature of all.

**TABLE 4.2: SINGLE FEATURE**

Features	DT	SVM	RF
F1	98.082	90.609	98.082
F2	98.786	<b>95.753</b>	98.786
F3	98.859	95.098	98.859
F4	98.980	90.876	98.980
F5	88.789	73.477	88.789
F6	97.282	89.686	97.282
F7	95.850	80.805	95.826
F8	98.859	90.075	98.859
F9	94.370	73.477	94.418
F10	<b>99.199</b>	73.477	<b>99.199</b>
F11	89.080	73.453	89.080
F12	88.740	73.477	88.740

### 4.3. Detection Results with Proposed Approach

Our focus is to get a set of features using the ranking in table 4.1 that gives more accurate results than individual feature. We followed the approach discussed in section 3.4 for each test column.

**TABLE 4.3.1 ANOVA TEST RESULTS**

ANOVA	DT		SVM		RF
F9	<b>94.370</b>	F9	<b>73.477</b>	F9	94.370
F9F2	94.370	F9F2	73.477	F9F2	94.418
F9F3	94.370	F9F3	73.477	F9F3	94.248

F9F5	94.370	F9F5	73.477	F9F5	94.346
F9F7	94.370	F9F7	73.477	F9F7	94.394
F9F11	94.079	F9F1 1	74.326	F9F2F11	94.273
F9F10	94.370	F9F1 0	73.477	F9F2F10	94.200
F9F8	94.370	F9F8	73.477	F9F2F8	94.224
F9F12	94.370	F9F1 2	73.477	F9F2F12	94.346
F9F1	94.370	F9F1	73.477	F9F2F1	<b>94.942</b>
F9F4	94.370	F9F4	73.477	F9F2F4	94.346
F9F6	94.370	F9F6	73.477	F9F2F6	94.273

Starting with ANOVA test results highest ranked feature according to ANOVA test is F9 so we start with Decision Tree taking feature F9, for ANOVA we have 4 features (F2, F3, F5, F7) on rank 2. We proceed by combining F9 with (F2, F3, F5, F7) one by one but there is no change in accuracy therefore we proceed to next features which is F11 but we get lesser accuracy than F9 alone. So, we skip this feature and proceed further in a same manner till last feature F6. But none was greater than F9 therefore our most accurate feature in this category F9. Similarly preceding for or SVM starting with F9 followed by (F2, F3, F5, F7) on combining these with F9 the results are not distinct so we proceed further for F11 which is less than F9 therefore we skip this and go on in a same manner. Here decision tree and SVM we have found that individual feature F9 is the most accurate in their categories with 94.370% and 73.477% respectively but when we applied Random Forest, we found out that F9F2F1 is the most accurate set in its category with 94.942% accuracy following the same steps as earlier. On comparing the results of all the algorithms, we get F9F2F1 is the best set of features with 94.942%.



**TABLE 4.3.2 CHI SQUARE NORMAL TEST RESULTS**

<b>Chi Square Normal</b>	<b>DT</b>		<b>SVM</b>		<b>RF</b>
F9	<b>94.370</b>	F9	<b>73.477</b>	F9	94.370
F9F10	94.370	F9F10	73.477	F9F10	94.370
F9F12	94.370	F9F12	73.477	F9F12	94.273
F9F8	94.370	F9F8	73.477	F9F8	94.370
F9F11	94.079	F9F11	74.326	F9F11	94.273
F9F1	94.370	F9F1	73.477	F9F1	94.248
F9F4	94.370	F9F4	73.477	F9F4	94.273
F9F7	94.370	F9F7	73.477	F9F7	<b>94.394</b>
F9F6	94.370	F9F6	73.477	F9F7F6	94.346
F9F3	94.370	F9F3	73.477	F9F7F3	94.346
F9F2	94.370	F9F2	73.477	F9F7F2	94.370
F9F5	94.370	F9F5	73.477	F9F7F5	94.297

For chi square (normal) we get F9 as the best feature while ranking so we proceed with all the Algorithms we are using. On combining F9 with F10 and so on. We get best feature for Decision Tree, SVM and Random Forest with 94.370% and 73.477% respectively. After comparing we get F9 is the best chi-square (normal) with 94.370%.

**TABLE 4.3.3 CHI SQUARE MALWARE TEST RESULTS**

<b>Chi Square Malware</b>	<b>DT</b>		<b>SVM</b>		<b>RF</b>
F10	<b>99.199</b>	F10	<b>73.477</b>	F10	99.199
F10F9	94.370	F10F9	73.477	F10F9	94.346
F10F12	99.199	F10F12	73.477	F10F12	99.199
F10F11	99.199	F10F11	73.477	F10F11	99.199
F10F8	99.199	F10F8	73.477	F10F8	<b>99.223</b>
F10F1	99.199	F10F1	73.477	F10F8F1	99.199
F10F7	99.199	F10F7	73.477	F10F8F7	99.199
F10F4	99.199	F10F4	73.477	F10F8F4	99.223
F10F6	99.199	F10F6	73.477	F10F8F6	99.199
F10F2	99.199	F10F2	73.477	F10F8F2	99.223
F10F3	99.199	F10F3	73.477	F10F8F3	99.174
F10F5	99.199	F10F5	73.477	F10F8F5	99.199

Finally, we did the same with chi-square (malware) as well. For chi-square (malware) F10 is the highest ranked feature. Hence, starting with F10 we follow our approach and reach to a conclusion that with Decision Tree and SVM gives F10 as best feature but with different accuracy 99.199% and 73.477% respectively. But for Random Forest we get F10F8 as the best set with 99.223% accuracy.

Now we have 3 different set F9F2F1 from ANOVA test F9F7 from Chi-square (normal) and F10F8 Chi-square (malware). When we combine these 3 sets, we get a single set F10F9F2F7F8. After calculating the accuracy of the new set so formed we get the following results

**TABLE 4.3.4: ACCURACY OF FINAL SET (F10F9F2F7F8)**

DT	<b>99.490</b>
SVM	82.164
RF	99.417

In the end from the proposed model, we have got the highest accuracy of **99.490** % combining 5 features together that is F10F9F2F7F8 which is more than any single feature or different sets of features we obtained from the three tests.

#### **4.4. Summary**

We used different tests and machine learning algorithms to find out the best set of features which can detect intrusion. Using the approach in section 3.4 we found out that the results in chapter 4 has provided us with the set of 5 features that proved to be the best set which gave 99.490% accuracy.

## CHAPTER – 5

### CONCLUSION

In this thesis, we emphasized on the need of Network Security in this age of growing network attacks. Needless to say, network attacks can cause immense financial and practical loss to the companies, associations or even individuals too. It is often noticed that antiviruses and firewalls, which used to provide enough network security, are not at any level adequate to guarantee the insurance of an organization now against these everchanging attacks. This kind of situation requires smart counter measures to keep up the security of networks and important systems. Chapter 1 gave a deeper understanding to the meaning and role of Intrusion detection system. Intrusion detection system (IDS) can be defined as any device or any software design application whose motive is to direct an intrusion detection. Further, we explained all the different types of IDS's namely signature based, anomaly based and hybrid detection. In simple words it can be easily stated that the main objective of an IDS is to detect intrusions before they seriously damage the network.

Chapter 2 revolved around the works proposed in the literature for network traffic- based intrusion detection. We reviewed more than 50 papers in this chapter. In Chapter 3 firstly we introduced all the 12 features with their names and their meanings. Then we used statistical tests to rank these features in order of their efficiency to detect network intrusion which was the whole motive of our paper. We used two different statistical tests for this research namely ANOVA test and CHI SQUARE test. After giving a brief explanation about the meaning, the formulas and the working of the two tests we moved on to the machine learning classifiers and their purpose. For our research we used three different types of machine classifiers namely, Decision tree, SVM and Random Forest. Chapter 3 revolved mainly around our approach that we used to rank our features. The features are ranked in a way such that the feature fixed at the bottom will be the least efficient and vice versa. In this way, we prepared three columns. One for ANOVA, second for chi square normal and another for chi square malware. Further, we prepared another table taking all the single features one at a time and separately applied all three of the machine classifiers. Now to take into consideration the possibility that the combination of these features might be even more effective, we made different combinations of features and after applying all three machine classifiers again we obtained 3 different sets of columns. 2 possibilities came up while we combined the features, either the new combination showed higher accuracy or it didn't.

In cases where the new feature obtained had greater accuracy than the original one, it was ranked higher and in cases where the accuracy was left unchanged or was found to be less than the original feature, it was discarded or rejected.

Further, the above method was repeated on these 3 columns and we obtained a separate set of features at the end which are best in terms of network intrusion. In chapter 4 we presented all the tables, calculations and proofs to reach at a conclusion that according to our research a particular combination of 5 features namely “Bytes received”, “Bytes sent”, “Time interval between packets sent”, “Packet size sent” and “Packet size received” gave the highest accuracy of 99.490 % after applying Decision Tree amongst all possible features and combinations.

## REFERENCES

- [1] C. Radhakrishnan, K. Karthick and R. Asokan, "Ensemble Learning based Network Anomaly Detection using Clustered Generalization of the Features," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 157-162.
- [2] F. L. Aryeh and B. K. Alese, "A Multi-layer Stack Ensemble Approach to Improve Intrusion Detection System's Prediction Accuracy," 2020 15th International Conference for Internet Technology and Secured Transactions (ICITST), 2020, pp. 1-6.
- [3] H. Yang and F. Wang, "Wireless Network Intrusion Detection Based on Improved Convolutional Neural Network," in *IEEE Access*, vol. 7, pp. 64366-64374, 2019.
- [4] A. Srivastava, A. Agarwal and G. Kaur, "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 524-528.
- [5] Y. Zeng, H. Gu, W. Wei and Y. Guo, "Deep-Full-Range: A Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework," in *IEEE Access*, vol. 7, pp. 45182-45190, 2019.
- [6] Y. Dong, R. Wang and J. He, "Real-Time Network Intrusion Detection System Based on Deep Learning," 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019, pp. 1-4.
- [7] S. Gopalakrishnan and A. Rajesh, "Cluster based Intrusion Detection System for Mobile Ad-hoc Network," 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2019, pp. 11-15.
- [8] Yang and H. Lim, "Deep Learning Approach for Detecting Malicious Activities Over Encrypted Secure Channels," in *IEEE Access*, vol. 9, pp. 39229-39244, 2021.
- [9] A. A. Bahashwan, M. Anbar, I. H. Hasbullah, Z. R. Alashhab and A. Bin-Salem, "Flow-Based Approach to Detect Abnormal Behaviour in Neighbour Discovery Protocol (NDP)," in *IEEE Access*, vol. 9, pp. 45512-45526, 2021.
- [10] A. Saber, M. Abbas and B. Fergani, "Two-dimensional Intrusion Detection System: A New Feature Selection Technique," 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), 2021, pp. 69-74.
- [11] L. Nie et al., "Intrusion Detection for Secure Social Internet of Things Based on Collaborative Edge Computing: A Generative Adversarial Network-Based Approach," in *IEEE Transactions on Computational Social Systems*.
- [12] D. Gupta and R. Singh, "Empirical Analysis of NIDPS using Machine Learning Models," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 679-685.
- [13] D. Kapil, N. Mehra, A. Gupta, S. Maurya and A. Sharma, "Network Security: Threat Model, Attacks, and IDS Using Machine Learning," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 203-208.
- [14] Z. Naseer, M. Rizwan, S. Sarwar and M. B. Khan, "Optimality of Feature set for Intrusion Detection System," 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), 2021, pp. 693-698.
- [15] L. Nie et al., "Intrusion Detection in Green Internet of Things: A Deep Deterministic Policy Gradient-based Algorithm," in *IEEE Transactions on Green Communications and Networking*.
- [16] Y. Su, K. Qi, C. Di, Y. Ma and S. Li, "Learning Automata based Feature Selection for Network Traffic Intrusion Detection", 2018 IEEE Third International Conference on Data Science in Cyberspace, 978-1-5386-4210-8/18/\$31.00 ©2018 IEEE.
- [17] N. Moustafa, B. Turnbull, and K.k.R. Choo. "An Ensemble Intrusion Detection Technique based on proposed Statistical Flow Features for Protecting Network Traffic of Internet of Things", *IEEE Internet of Things Journal*.
- [18] A. Shenfield, D. Day and A. Ayesh, "Intelligent Intrusion Detection Systems using Artificial Neural Networks", in *ICT Express*, S2405-9595(18)30049-3.
- [19] A.F.M. Agarap, "A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data", *ICMLC 2018*, February 26-28, 2018, Macau, China © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6353-2/18/02...\$15.00.
- [20] SULTAN, ZAVRAK, AND M. İSKEFIYELI, "Anomaly-Based Intrusion Detection from Network Flow Features Using Variational Autoencoder", *IEEE ACCESS open access Journal*.
- [21] V.R, S.KP and P.Poornachandran, "Applying Convolutional Neural Network for Network Intrusion Detection", 978-1-5090-6367-3/17/\$31.00 ©2017 IEEE.
- [22] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, "Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic", 1949-307X 2018 IEEE.
- [23] B.G. Atli, Y. Miche, A. Kalliola, I. Oliver, S. Holtmanns, and A. Lendasse, "Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space", Springer Science Business Media, LLC, part of Springer Nature 2018.
- [24] K.A.Taher, B. Mohd.Y. Jisan and Md. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection", in 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST).
- [25] Y. ZHANG, X.CHEN, D. GUO, M.SONG, Y.TENG, AND X. WANG, "PCCN: Parallel Cross Convolutional Neural

- Network for Abnormal Network Traffic Flows Detection in Multi-Class Imbalanced Network Traffic Flows”, in IEEE ACCESS open access Journal .
- [26] L.Vu,V. L. Cao, Q.U.Nguyen, D. N. Nguyen, D.T.Hoang and Eryk Dutkiewicz , “Learning Latent Distribution for Distinguishing Network Traffic in Intrusion Detection System” , 978-1-5386-8088-9/19/\$31.00 ©2019 IEEE.
- [27] L. Santos,R. Gonçalves C.Rabadao and J.Martins, “ A flow-based intrusion detection framework for internet of things networks”, in SPRINGER Cluster Computing.
- [28] A. Binbusayyis and T. Vaiyapuri , “Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM”, in Springer Applied Intelligence.
- [29] H.Li , “ Research on network intrusion detection technology based on improved FCMA algorithm”, Journal of Ambient Intelligence and Humanized Computing.
- [30] D. B.Dasari , G. Edamadaka , Ch. S. Chowdary and · M. Sobhana, “ Anomaly based network intrusion detection with ensemble classifiers and meta heuristic scale (ECMHS) in traffic flow streams”, Journal of Ambient Intelligence and Humanized Computing.
- [31] A. Binbusayyis and T. Vaiyapuri , “Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM”, in Springer Applied Intelligence.
- [32] A Boulaiche, K Adi , “ An auto-learning approach for network intrusion detection”, Telecommunication System (2018) 68:277–294.
- [33] A. Davahli, M. Shamsi, G. Abaei, “ Hybridizing genetic algorithm and grey wolf optimizer to advance an intelligent and lightweight intrusion detection system for IoT wireless networks”, journal of Ambient Intelligence and Humanized Computing (2020) 11:5581–5609.
- [34] R. Atefni , M. Ahmadi , “ Network intrusion detection using multi architectural modular deep neural network”, The Journal of Supercomputing (2021) 77:3571–3593.
- [35] A. R.Gupta ,J. Agrawal , “ The multi demeanour fusion based robust intrusion detection system for anomaly and misuse detection in computer networks” , journal of Ambient Intelligence and Humanized Computing (2021) 12:303–319.
- [36] X. Wang, S. Yin1, H. Li, J. Wang, L. Ten, “ A Network Intrusion Detection Method Based on Deep Multi scale Convolutional Neural Network”, International Journal of Wireless Information Networks (2020) 27:503–517.
- [37] C.M. Hsu, M. Z Azhari , H.Y.Hsieh, S. W.Prakosa, J.S Leu, “ Robust Network Intrusion Detection Scheme Using Long-Short Term Memory Based Convolutional Neural Networks”, Mobile Networks and Applications.
- [38] S. Krishnaveni,S. Sivamohan, S. S. Sridhar, S. Prabakaran, “ Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing, 123 Cluster Computing.
- [39] J. Kevric, S. Jukic and A. Subas, “ An effective combining classifier approach using tree algorithms or network intrusion detection” , Neural Computer & Application (2017) 28 (Suppl 1):S1051–S1058.
- [40] C. Caligari , S. Vaton and M. Pagano, “ Communication Networks A new statistical method for detecting network anomalies in TCP traffic”, EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS Eur. Trans. Telecoms. 2010; 21:575–588 Published online 15 June 2010 in Wiley Online Library.
- [41] D. Juliette, R. Gilles and O. Philippe , “ Online and Scalable Unsupervised Network Anomaly Detection Method”, NSM.2016.2627340, IEEE.
- [42] K. Huang, Q. Zhang, C. Zhou, N. Xiong, “ An Efficient Intrusion Detection Approach for Visual Sensor Networks Based on Traffic Pattern Learning” , Ieee Transactions On Systems, Man, And Cybernetics: Systems.
- [43] L. Gandhimathia and G. Murugaboopathia, “A Novel Hybrid Intrusion Detection Using Flow-Based Anomaly Detection and Cross-Layer Features in Wireless Sensor Network”, ISSN 0146-4116, Automatic Control and Computer Sciences, 2020, Vol. 54, No. 1, pp. 62–69. © Allerton Press, Inc., 2020.
- [44] M.Ceška ,V. Havlena, L. Holík, O. Lengál and T. Vojnar, “ Approximate reduction of finite automata for high-speed network intrusion detection”, International Journal on Software Tools for Technology Transfer (2020) 22:523–539
- [45] S.Kaur and M. Singh , “ Hybrid intrusion detection and signature generation using Deep Recurrent Neural Networks”, Neural Computing and Applications (2020) 32:7859–7877
- [46] N.Hu,Z.Tian, H. Lu , X. Du, and M.Guizani, “ A multiple-kernel clustering based intrusion detection scheme for 5G and IoT networks”, International Journal of Machine Learning and Cybernetics.
- [47] P. Bedi , N. Gupta & Vinita Jindal, “ I-SiamIDS: an improved Siam-IDS for handling class imbalance in network-based intrusion detection systems”, Applied Intelligence (2021) 51:1133–1151.
- [48] K.Vengatesan, A. Kumar, R. Naik and D.K. Verma , “Anomaly Based Novel Intrusion Detection System For Network Traffic Reduction”, Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018) IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1-5386-1442-6.
- [49] V. Jyothsna, K. M. Prasad,K. Rajiv, G. R.Chandra, “ Flow based anomaly intrusion detection system using ensemble classifier with Feature Impact Scale.”, Cluster Computing
- [50] M. Hawedi ,C. Talhi and H. Boucheneb , “ Multi-tenant intrusion detection system for public cloud (MTIDS)”, ] The Journal of Supercomputing (2018) 74:5199–5230 .