INTRUSION DETECTION AND PREVENTION OF A MALICIOUS ACTIVITY IN A FEDERATED CLOUD COMPUTING PARADIGM

By

Akash Gerard

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfilment of the requirements for the degree of MS in Information Security

November 2018

Supervisor Certificate

This is to certify that **Akash Gerard** Student of **MSIS-15** Course Reg.No **00000171430** has completed his MS Thesis title **"INTRUSION DETECTION AND PREVENTION OF A MALICIOUS ACTIVITY IN A FEDERATED CLOUD COMPUTING PARADIGM"** under my supervision. I have reviewed his final thesis copy and I am satisfied with his work.

Thesis Supervisor

(Assoc Prof Dr. Haider Abbas)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by **Akash Gerard** Registration No. **00000171430**, of **Military College of Signals** has been vetted by undersigned, found complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial, fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have been also incorporated in the said thesis.

Signature: _____

Name of Supervisor:_____

Date: _____

Signature (HOD):_____

Date: _____

Signature (Dean/Principal):_____

Date: _____

Declaration

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

Dedication

I dedicate this thesis to my PARENTS, TEACHERS, COUSINS and FRIENDS for their love, endless support and encouragement.

I also want to offer special thanks to my parents, Ruth Gerard and Gerard Emmanuel, who instilled in me the love of learning from an early age. My parents have been constant cheerleaders through every academic and personal endeavor in my life. Thanks mom and dad for always believing in me and for encouraging me to strive for my dreams.

Acknowledgement

All praise to God, the Almighty, for blessing me and providing me the strength to complete this thesis.

I would like to convey my gratitude to my supervisor Dr. Haider Abbas and co-supervisor Dr. Rabia Latif for their supervision, guidance and continuous support. Their invaluable help of constructive comments and suggestions throughout the experimental and thesis works are major contributions to the success of this research. Also, I would thank my committee members; Asst. Prof. Mian M. Waseem Iqbal and Maj. (R) Dr. Faisal Amjad, for their support and guidance regarding this topic. I also want to express my sincere gratitude to the rest of the professors; Dr. Naima Altaf, Dr. Ayesha Maqbool and Asst. Prof. Waleed Bin Shahid that have advised and helped me during my thesis.

Last, but not the least, I am highly thankful to my mother (Mrs. Ruth Gerard), and my father (Gerard Emmanuel). They have always stood by my dreams and aspirations and have been a great source of inspiration for me. I would like to thank them for all their care, love and support through my times of stress and excitement.

Copyright Notice

Abstract

The concept of federated cloud computing has reduced the computational cost of an individual user. On the other side, it has increased the security and privacy issues of the information of a cloud client. Users sharing computational resources in a federated cloud can be malicious and can gain access to the sensitive data of users of other cloud providers. Therefore, there is a need to monitor the behavior of consumers, exchanging data between different cloud servers.

Using intrusion detection techniques, we can avoid the malicious traffic from gaining an access of the critical data of the clients using cloud services of other cloud providers. Moreover, by adding intrusion prevention technique, we can make our system more robust and efficient. Intrusion Detection and Prevention techniques helped a lot in detecting the malicious activities performed by the intruders. In this domain of securing data, a lot of research is being done. Recently, artificial intelligence and machine learning have greatly attracted the attention of researchers to integrate the concepts of network security with artificial intelligence.

In this research, artificial intelligence techniques have been studied and artificial neural network (ANN) model has been finalized to detect the intrusions. Prevention methods are also discussed in this thesis to deploy a properly secured federation in cloud computing environment.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Overview

In this era, where we all are surrounded by the fastest growing developments being done in the field of technology. Every day, we can see the advancements in the fields of science. Once, it was a time when it was just a thought to travel around the world within seconds but today it is no more just a thought, but it is possible. You can do anything, anytime, anywhere simply by getting connected to the Internet.

Few years back if you wanted to carry your data, you needed hard drives, data storing devices, or simply your computing resources for that but today that is made so easy to access your data anytime, at any place by just the availability of Internet i.e. Cloud Computing (CC).

Cloud Computing is an emerging technology. It has root connections to the distributed computing, cluster computing and grid computing [1]. It has the similar aim as of the previous computing technologies (mentioned above) i.e. resource virtualization. The difference between grid computing and cloud computing is that the grid computing focuses on achieving maximum computing resources whereas cloud computing focuses on optimizing the overall computing resources [1]. Cloud computing introduced a totally new concept of accessing your online resources, data from anywhere throughout the world by just getting connected to the internet. Cloud computing offers multiple services to its users like for storage it offers infrastructure services, for development of

application it offers platform, and software services as well. Customer when chooses cloud services, the data that is present in local repositories moves to remote data centers. Those services can be accessed by user with the help of the services provided by cloud service providers (CSPs). We can conclude from here that to get/post anything from/in the cloud a secured medium is required that won't allow the eavesdroppers to access that channel to avoid data leakage from that medium. If the proper mechanisms aren't implemented over the cloud infrastructure, then there are many chances of data breach. Malicious user (MU) who wants to get access of the data that is transmitted in cloud will get an access of the channel between user and remote location. Likewise, MU can also intrude into the users accounts to perform any action.

The combination of such multiple cloud providers sharing resources, services, infrastructures, or platforms is cloud federation. In cloud-federated environment, different cloud providers based on trust and assumptions share their resources with each other. Mainly, federated organization share attributes of user's identity and access permission using service access requirements [2].

In cloud federation security, the main goal is to develop the mutual trust relationship between multiple federated identities and the components that are involved within the cloud. Such mutual trust can be developed by securing the components that are involved in the cloud-computing framework. Those components [3] are server, client, networking and communication platforms.

Federation of cloud allows users to exchange data within federated identities that can allow malicious users to gain access of the servers of other cloud providers. Gaining an access to data of other CSPs can allow a malicious user to perform certain type of cyber-attacks like DDOS (Distributed Denial of Service), cloud malware injection. These attacks cannot only result in the poor performance but can also affect the economic stability of a cloud provider. The behavior of multiple users of a cloud could be monitored by using intrusion detection techniques. Integration of cloud security with intrusion detection systems (IDSs) is a new challenge in this era of technology. By combining these two, each activity of every user can be monitored. If a malicious user tries to perform any activity against cloud, his actions can be prevented. In this research, the prevention of malicious data has also been covered that will allow the computing

paradigms to detect and prevent the malicious content of a user in worst scenarios. This study presents an approach towards the detection and prevention of an activity performed by a malicious user.

1.2 Aims, Objectives and Research Questions

This section focuses on the aims and objectives of this thesis work and concludes the answer to proposed research questions.

1.2.1 Aims

The aims of this thesis are:

- After studying the network architecture of FCC, mitigating possible threats in it.
- Propose a model for intrusion detection and prevention that could be deployed in federated cloud computing environment.

1.2.2 Objectives

The objectives are:

- To study the architecture of federation in cloud computing.
- To study the possible attacks in cloud federation.
- Identifying inherited cloud computing challenges.
- Studying the artificial intelligence techniques that could be deployed in cloud computing framework to detect and prevent the intrusions.
- To propose a intrusion detection framework to detect attacks in real time.

1.2.3 Research Questions

This thesis focuses to answer the following research questions.

1. What are the challenges in cloud computing security?
2. Is it possible to integrate artificial intelligence techniques with cloud computing to provide a better framework for data security?
3. Among all the artificial intelligence techniques which technique will best suit to detect an intrusion in the cloud paradigm?

1.3 Motivation

To contribute to this domain the strong willingness came after reading the publications and research articles. To get the basic understanding of cloud computing, a thorough study of [3], [4], [5] was done. After that, to get an understanding of cloud computing architecture [6] was studied. To give a proper direction to this research [7], [8], [9] guided me. A detailed study of a survey paper [9] and a confluence report [11] was done to understand the relationship between cloud computing security and artificial intelligence techniques.

1.4 Problem Statement

Trust and privacy of data are two major challenging concerns of cloud computing. Trusted computing can be utilized to fortify solutions of existing security problems by building relationship of trust between federated identities that are participating, segments of the infrastructure of cloud and frameworks inside the cloud. In federated environment, there is no trust platform between the communicating parties. Exchange of data between multiple cloud providers must be ensured w.r.t. integrity. There is a lot of research being carried out in the domain of cloud computing security. Federated

cloud computing as being the emerging technology of this era has its own security challenges that need to be focused.

Cloud Federation allows multiple parties to exchange data based on trust and assumptions. Assuming that, one malicious user being a part of any party in cloud federation can gain access to highly confidential data of any other user belonging to the other cloud service provider. In the environment based on the definition of cloud federation, that malicious user can easily get access to the data of other users that are the part of that environment. Malicious user can act covertly and can exploit many vulnerabilities that are inherited to cloud environment and based on those vulnerabilities, considering the worst scenario, can attack and bring down the whole cloud infrastructure. Focusing on the bigger picture, multiple malicious users of different CSPs can act as a single unit and can badly effect the cloud paradigm. In a federation environment, a lot of trust on other parties is required that can only be possible if all the parties that are inside the umbrella of federation are providing the best services and are only giving the access to their services to the highly trusted consumers.

But the trust on a user is a factor that is variable that can't be quantified at any certain instance. For example, one person at instance T could be a highly trusted person but on the other instance let's say T+N the trust can't be guaranteed the same as it was on the instance T. So, there must be a system that would be beyond all these variable factors like trust and dependence on consumers. From a long time, a lot of research is being carried out in the domain of Intrusion Detection and Prevention Systems. In Cloud Computing, Intrusion Detection and Prevention Systems (IDS/IPS) can be integrated to monitor the behavior of users. With the help of IDS/IPS we can educate our systems to monitor the overall network and in case of any malicious activity these systems can send alerts and can terminate the opponent's action based on the severity of intrusion being done.

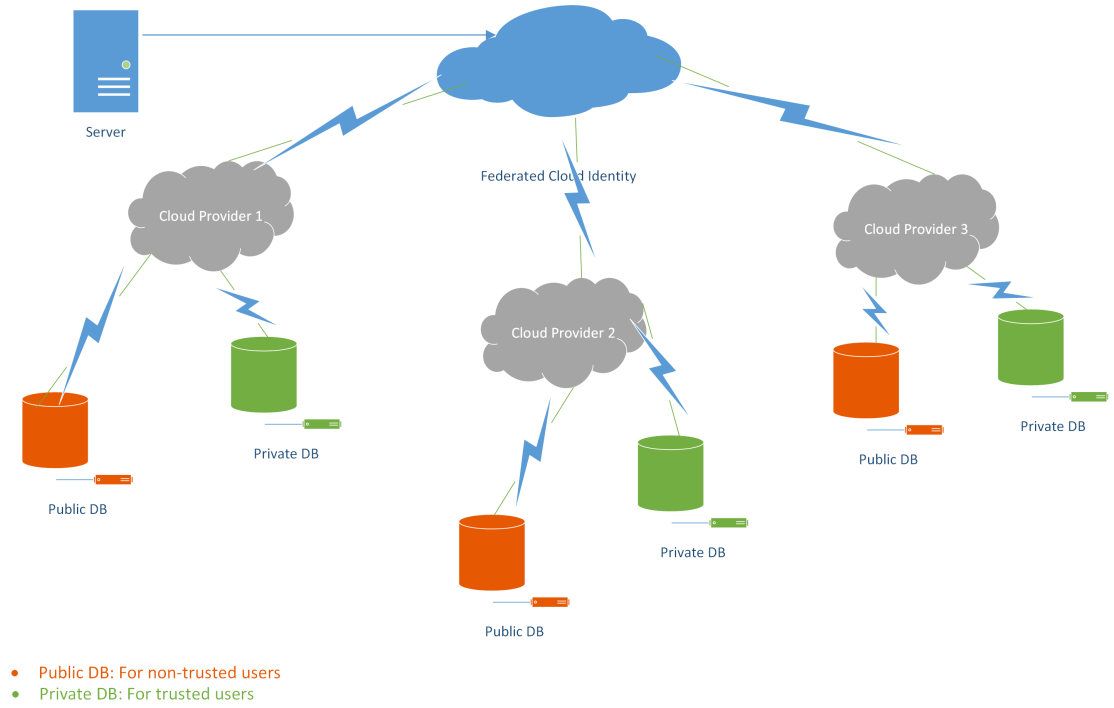

Figure 1.1: Federated Cloud Computing Environment

1.5 Research Contribution and Evaluation Process

The major contribution towards this domain is the proposition of an intrusion detection system based on machine learning technique that could be deployed in federated cloud computing environment. Secondly, algorithm has been suggested knowing that intrusion prevention systems are the extension to intrusion detection systems. Taking an output from intrusion detection system we can give it to an intrusion prevention system as an input that will decide either to deny or allow that network transmission.

1.6 Thesis Organization

This research document is divided into six chapters. Following is the summary of each chapter.

- **Chapter 1:** This chapter includes overview of thesis, aims, objectives and research questions. It also highlights motivation in choosing this domain, contribu-

tions of this research and the evaluation process.

- **Chapter 2:** This chapter covers a comprehensive literature review related to the topic of thesis. A brief introduction and analysis of IDS and IPS is discussed. Furthermore, the role of artificial intelligence is elaborated in the field of computing frameworks. Integration of artificial intelligence wih the principles of network security in the form of intrusion detection ad prevention frameworks is covered as well.

- **Chapter 3:** In this chapter, a brief introduction of cloud computing framework is discussed along with the characteristics of cloud computing framework that made it unique from the other computing technologies. A comparative study of control over security in service and deployment models is highlighted.

- **Chapter 4:** This chapter contains the security concerns that revolve around securing the data in cloud computing framework. Concepts related to data in transit and data at rest were covered are discussed. Further, data encryption is covered as a part of this chapter followed by mistakes of data encryption in cloud computing.

- **Chapter 5:** In this chapter, the proposed solution and proposed framework of the problem statement is discussed. After that UNSW-NB15, the dataset used in the experiments of this thesis is discussed briefly. Lastly, the terminologies that will be used in later chapters to understand the results and the work of this thesis are covered. Moreover, the experiments and results are discussed in detail followed by the introduction of new terms like cross entropy and percentage error that gives the description of the designed algorithm that how well it can perform. The results are shown using different plots of confusion matrix, ROC, performance, training state and error histogram.

- **Chapter 6:** This chapter covers the discussion of topic of this thesis. Later, the conclusion is discussed that answers the research questions of this thesis. At the end, future work of this work is covered. Future work includes few suggestions and propositions for future researchers who are willing to contribute in this domain.

Chapter 2

Literature Review

2.1 Introduction

The concept of federated cloud computing has reduced the computational cost of an individual user. On the other side, it has increased the security and privacy issues of the information of a user. Users sharing computational resources in a federated cloud can be malicious and can gain access to the sensitive data of users of other cloud providers. Therefore, there is a need to monitor the behavior of users, exchanging data between different cloud servers. Using intrusion detection techniques, we can avoid the malicious traffic from gaining an access of the critical data of the users of others CSPs. Moreover, by adding intrusion prevention technique, the system become more robust and efficient.

2.2 Intrusion Detection System

Intrusion Detection Systems (IDSs) are widely being used in the domain of computer and network security not only to protect the systems from external users but from insiders as well.

In network security, IDS [10] is basically considered as the second layer of defense against the attacks that usually harm the network. Intrusion is defined as, "Any malicious activity that is done to compromise any system and its security triad."

The security triad of information security (IS) are:

- Confidentiality
- Integrity
- Availability

The system that detects an intrusion is known as intrusion detection system. Based on their physical presence, intrusion detection systems are mainly divided into two categories i.e. Network based IDS and Host based IDS. The summarized analysis of these two categories are shown in table 2.1.

Network based Intrusion Detection Systems (NIDS): It is placed at some places inside the network that examines the network traffic and based on some rules it allows or denies the traffic to pass through it.

Host based Intrusion Detection Systems (HIDS): It is basically a kind of agent that is employed only to protect an individual host. HIDS are installed locally on host machines. It does not examine the network traffic of any other host except itself.

Table 2.1: Summarized Analysis of HIDS and NIDS

Function	HIDS	NIDS
Rejection of packets	No	Yes
Experts Knowledge	Less	More
Type of attacks	Local attacks	Network attacks
Packet Inspection	No	Yes
Dependency on Host	Yes	No
Bandwidth	No	Yes
False Positive Rate	High	Low
Examples	OSSEC	Snort

IDS is consisted of three components:

- Source of data
- Analysis of data
- Response

Source of data could be gathered by any mean like for instance for a single host, HIDS, or for a traffic over a network, NIDS. There are two common methods to analyze the data:

Analysis based on Anomalies: Figure 2.1 depicts the categories of anomaly based IDS. In this analysis, the normal behavior of a user is considered. It can either be programmed in which a system programmer provides a pattern of normal usage of user or it could either be self-learned in which the system is trained to build the normal behavior of a user automatically. It can identify zero day/ unknown attacks while on the other hand it is inherited with high false alarm rate. In case of different type of attacks, it [9] cannot differentiate between them.

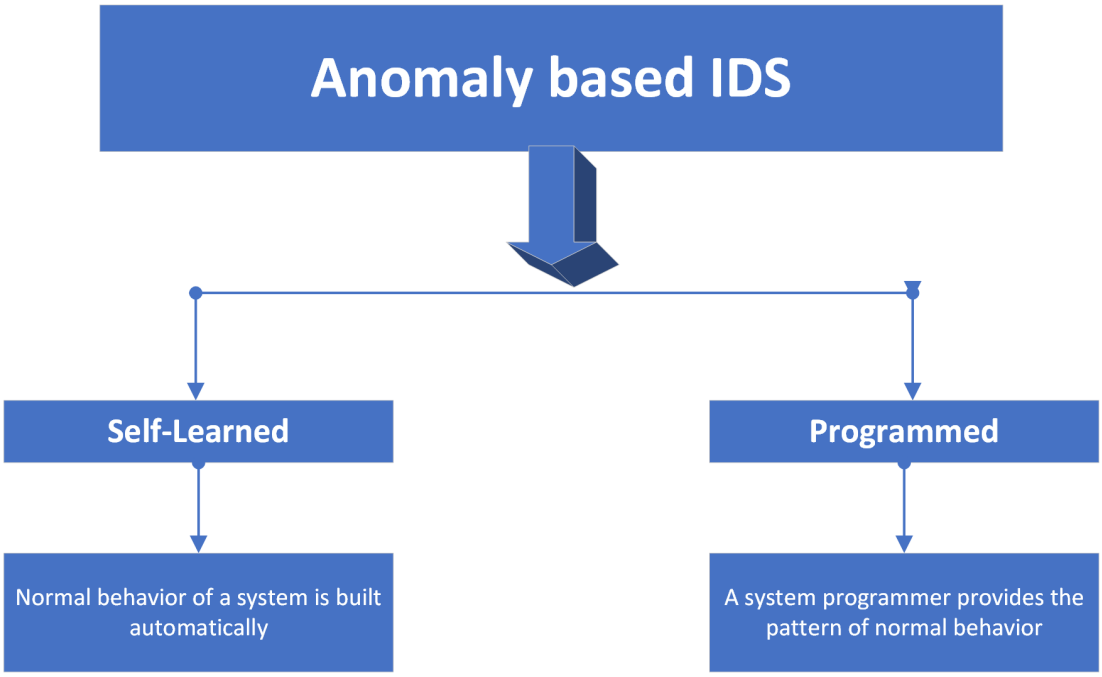

Figure 2.1: Anomaly based IDS

Analysis based on Signatures: Fig. 2.2 shows the signature based IDS. In this type of analysis, the IDS is programmed which compares the pre-existing signatures with the normal traffic and in case of similar pattern or behavior it notifies the system about an intrusion. A signature is basically a byte sequence of data. It is so accurate in detecting the known attacks that it checks every byte sequence of the payload and after comparing the results it allows the traffic to pass through it, but it is not inherited [9] to identify zero-day attacks.

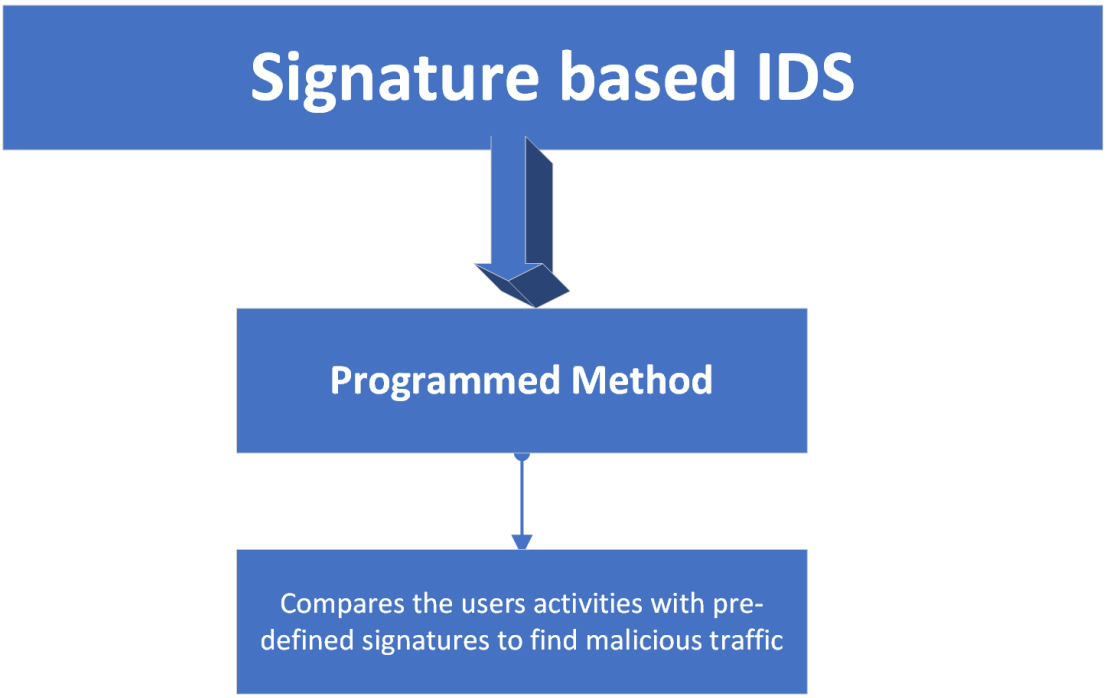

Figure 2.2: Signature based IDS

In case of any intrusion the third component of an IDS gives a certain response based on defined rules. There are two type of responses:

Active Response: In case of any intrusion, if the IDS is programmed to take decision, after the analysis of traffic, and to block the incoming packets of traffic by its own then such response is said to be an active response [14] of an IDS.

Passive Response: In case of an intrusion, if the IDS notify the administrator to act, then such a response is called passive response of an IDS.

Most of the IDSs are passive [14].

2.3 Intrusion Prevention System

In this era of computing world, where everything is going to be shifted to artificial intelligence. Machines are taking their own decisions to perform different responses to events. Everything is evolving with such a great speed that it won't be considered a regret if we claim that in a coming future, machines will analyze [11] what is good for them and after analyzing they will start responding to those events.
A system having an ability to correspond to network events, in such a way that if it monitors any intrusion in a network it prevents that intrusion in a real time, is known as Intrusion Prevention System (IPS). In a real-time network, IPS can detect the traffic passing through it, in case of any threat it is able to take actions like dropping a packet or terminating the whole session of communication. It further makes a detailed analysis [14] report of a network and forwards it to the network administrator.
Intrusion prevention systems may vary with respect to how they sense the stream of data for the detection of an intrusion. Few of the methods used in an IPSs are:

Profile Method: In this method, IPS collects data that is sent to and from the computer network in trusted states. From the collected data, a profile is created that is considered a normal profile. After that, the normal profile is matched to the traffic passing through a network in real time. If pattern of a real time traffic deviates from the normal profile, it is treated as an intrusion and the preventive action is taken against it. Normal profile [15] is based on the trusted behavior of a user, network connection, and computer applications. IPS can also be trained to detect the deviation against normal traffic pattern by using artificial intelligence techniques (discussed in section 2.5). It has few disadvantages:

- High false alarm rates.
- Any change (even a trusted one) might be considered as an attack.

- Due to the frequent changes in network topologies, it is hard to maintain a normal profile of a network.

Signature Method: In this method, IPS compares the real-time network traffic with the signatures of attack traffic stored in databases. It simply examines every byte of data for a known pattern that is associated with any attack. The pattern could be anything like a special string or may be any command. For example, the usernames and passwords used in repetitive login attempts. If an intrusion prevention system is configured to compare usernames and passwords with the list of usernames and passwords generated from a known attack, then this match will also generate an alarm. In signature method, it is very important to update a database of signatures in form of software patches in the same way as antivirus gets updated after the addition of a known pattern of detected attack. In case of any match, an alert is generated for a network administrator to make sure that an intrusion [15] doesn't cause any harm to the computer system. The disadvantages of signature method are:

- Zero day or unknown attacks will be undetected by an IPS.
- If a known attack, works in an unusual manner even then there is a possibility that it may go undetected.
- A simple modification in the known attack will be considered as a normal traffic pattern.

Stateful Protocol Method: A payload of a network packet is always covered with a lot of headers of various protocols. As the packet is reached to each of the layer of OSI model, an additional header is attached to it. Protocols follow a standard document format known as RFCs (Request for Comments). It completely explains the use of each protocol forming the basis of stateful protocol method. In stateful protocol method, each packet is peeled off and is scanned for its consistency as compared to what RFC specifies. A minor deviation observed, will result in an alert. Other than monitoring the ideal behavior of protocol, IPS is that much intelligent to know about the implementation of a protocol in real world to make sure that a normal practiced RFC violation

isn't treated as an intrusion attempt to compromise security of a computer system. It is almost same a profile method but varies in such a way that profile method host or network-based rules while stateful protocol method uses protocol specific rules as per RFCs. It properly scans the protocol states to make sure that the protocol [15] is being used in proper manner.

There are different types of Intrusion Prevention Systems depending upon their physical locations. Popular types of IPSs are:

Network based IPS (NIPS): Network based IPS (NIPS), as the name suggests, keep a check on incoming and outgoing traffic of a network. In a network, NIPS is installed at key positions like in a range of firewall, router, or switch. Most of the network based IPSs are designed to scan the network traffic at an application layer but due to the advancements in technology, NIPS now can scan traffic at internet or transport layer [16]. Network based IPSs detect and prevent network attacks. In case any intrusion is detected in a network, NIPS are configured to respond to the central management servers that specifies the preventive action to be taken against an attack.

Host based IPS (HIPS): Host based IPS are installed at host and monitors the incoming and outgoing traffic. On a host, HIPS is installed as a software. To prevent attack attempts against any system, it scans the various characteristics of a host where it is installed. Agent is installed on a host system as a component of an IPS. Agent is responsible to scan the characteristics of a host and to take preventive measures. All the agents that are installed on hosts report to centralized management servers that manage those hosts [16]. All the hosts in a network have their own IPS, whereas the complete information of a network can be collected from the centralized management systems.

Wireless based IPS (WIPS): Wireless based IPS scans the wireless traffic of a network to prevent an unauthorized access on local area networks (LANs) or any other network connection. WIPS monitors the spectrum of microwave and radio wave for the presence of rogue access point. It is also responsible for detecting attack tools that can attack a network. In case of detection, WIPS reports an event to security administrator

of a network [14]. In a network, it is deployed in the form of sensors that keep on sensing network traffic.

2.4 IDS/IPS in Cloud Framework

Lo et al. [8] proposed a cooperative intrusion detection framework for cloud computing networks. They mainly focused on to reduce the impact of DOS or DDOS attacks in cloud environment. In cloud region, they deployed IDSs that keep on communicating with each other. A cooperative agent was used that determines either to accept the alert from other entities or not that helped in identifying the same type of attacks. Though the approach was good, but the work resulted in an increase of the computational effort as compared to the snort based IDS.

In the work of Calheiros et al. [13], they proposed a solution based on InterCloud. They used Cloud Coordinators to increase the reliability, performance, and scalability of applications that are elastic. Their work can be used in intrusion detection domain by placing Cloud Coordinators in every CSP to interact with the multiple cloud entities. Clients who need to acquire services and do not want to claim the facilities, provided by CSP, in the market utilize Cloud Coordinators.

Montes et al. [18] suggested a solution based on cloud monitoring. In their paper, they proposed a unified cloud monitoring taxonomy that is based on layered cloud monitoring architecture. They have implemented GMonE that covers each aspect of cloud monitoring, addressing all the requirements of modern cloud infrastructures. The work of Chen et al. [17] aims a practical collaborative Network Security Management System with an effective collaborative UTM (Unified Threat Management). In their work they merged the distributed security network with centralized security data centre to leverage the P2P communication protocol used in the module of UTM that allows to update the events and the rules of security. They implemented a fully cloud based security centre for the forensic analysis of network security. They used the cloud storage to keep the collected traffic data and then processed it through cloud computing infrastructures to analyze the malicious patterns. The rules they set in their proposed model are enforced

by collaborative UTM and the feedback is returned to the security data centre to identify the patterns of new attacks more effectively.

Currently available Intrusion Detection Systems are not able to handle effectively the real time network analyzing environments. They detect intrusions based on signatures and predefined patterns. It is difficult for them to analyze and detect the zero day attacks. They can also ignore the changes done to the operating environment that may eliminate or result in vulnerabilities. Because of this incapability, they may result in the detection of intrusions that are irrelevant only because of doing changes to the environment. The ability to be aware of the threat will help IDS to improve its detection rate. In that capacity, there are no unified detection and anticipation ways to deal with recognize intrusions in the Cloud environment, nor is there any inclusive acknowledged metric or standard to assess against [14].

Table 2.3: Analysis of Intrusion Detection and Prevention Systems

IDS	**IPS**
It is a device that monitors traffic for an intrusion and sends alert to network administrator.	It is a device that inspects network traffic, detects and intrusion and classifies it and then prevents malicious traffic.
Analyse traffic patterns.	Analyse traffic patterns.
Alarms on the detection of an anomaly.	Prevents traffic on detection of an anomaly.
Must be placed non-inline through port scan.	Must be placed inline generally after firewall.
Detects traffic in real time and captures signatures of attacks and then generate alarms.	Inspects traffic in real time and captures signatures of attacks and then prevents an attack on detection.
Must be configured in inline mode.	Must be configured in inline mode, generally at layer two.
Detection; anomaly based, and signature based.	Detection: profile based, signature based, stateful protocol based.
Types: NIDS and HIDS.	Types: NIPS, HIPS and WIPS.

Table 2.3 shows the analysis of intrusion detection and prevention systems (IDS/IPS).

2.5 Artificial Intelligence

Artificial Intelligence (AI) is a term used to describe an ability of a machine to observe, think, and react [20]. In AI, scientists and engineers maps the human intelligence over computing systems that respond to certain events in a particular manner. Researchers [20] are finding ways to discover those aspects that can solve multiple challenges in the field of computing.

2.5.1 A Brief History of Artificial Intelligence

Back to the year 1943, Pitts and McCulloch gave an idea of "Boolean Brain" in their paper [38]. Later to that, J.V. Neumann [22] reflected on the ideas of Pitts and McCulloch to design a digital computer. In 1950, Turing Test suggested that a computer can work intelligently if a person communicating by teletype wasn't able to distinguish the machine from a normal human being, based upon their response to certain questioning. The term "Artificial Intelligence" was first coined in the conference held at Dartmouth College in 1956. AI theorists focused on the value of symbolic logic in the development of computer programs that are intelligent. From that time to today, AI has attracted a lot of researchers and many studies have been carried out in this domain. Today almost in every field of science, AI [20],[21], [22] is playing its vital role that can never be denied. With every passing day, machines are becoming more and more intelligent that is all because of the scientists and engineers contributing in this domain.

2.6 Artificial Intelligence in IDS

Artificial Intelligence has contributed a lot in every field of science. In the field of intrusion detection systems many researchers have used different techniques of AI to detect the intrusions in network. Due to the flexibility, adaptability, efficiency and high accuracy to the computing system, AI techniques are highly likable as compared to the other techniques in intrusion detection systems.

There are few artificial intelligence (AI) based techniques, that are discussed below:

Pattern Matching: This technique [20] allows to generate a profile of a network by matching the pattern of traffic with the given sequence. After matching, those profiles are used for detecting the anomalies. But this technique, consumes much time in case of new networks.

Machine Learning (ML) Techniques: It is defined [18] as the training of computer applications to enhance their performance by discovering different tasks over the span of time. It is an ability of a machine to learn something without being explicitly learned. Tom Mitchell has defined Machine Learning as, "An application of a computer is said to gain experience E as for some task T and some execution measure P, if its execution on T, as estimated by P, enhances with experience E." For instance, in checkers, the experience E would be the experience of having the program play a huge number of diversions itself. The task T would be the assignment of playing checkers, and the execution measure P will be the likelihood that wins the following round of checkers against some new adversary. Generally, all the ML related problems are classified into two main categories as depicted in figure 2.3.

1. **Supervised Learning:** When we are given a data set and already have some idea about the correct output or in simple words an idea of the relationship between an input and an output, this type of learning is considered as a supervised learning. Supervised Learning is further categorized into two categories: regression and classification. Regression is the prediction of results in continuous output environment or a continuous function. In classification, we predict results in a discrete output. In simple words, classification is a mapping of input variables into discrete variables [20].

2. **Unsupervised Learning:** In unsupervised learning, we don't have any idea how the results or outputs would look like. In unsupervised learning, it is possible to derive structure from data where we don't necessarily know how the variable will affect. In unsupervised learning, we use clustering based on relationships among the data [20].

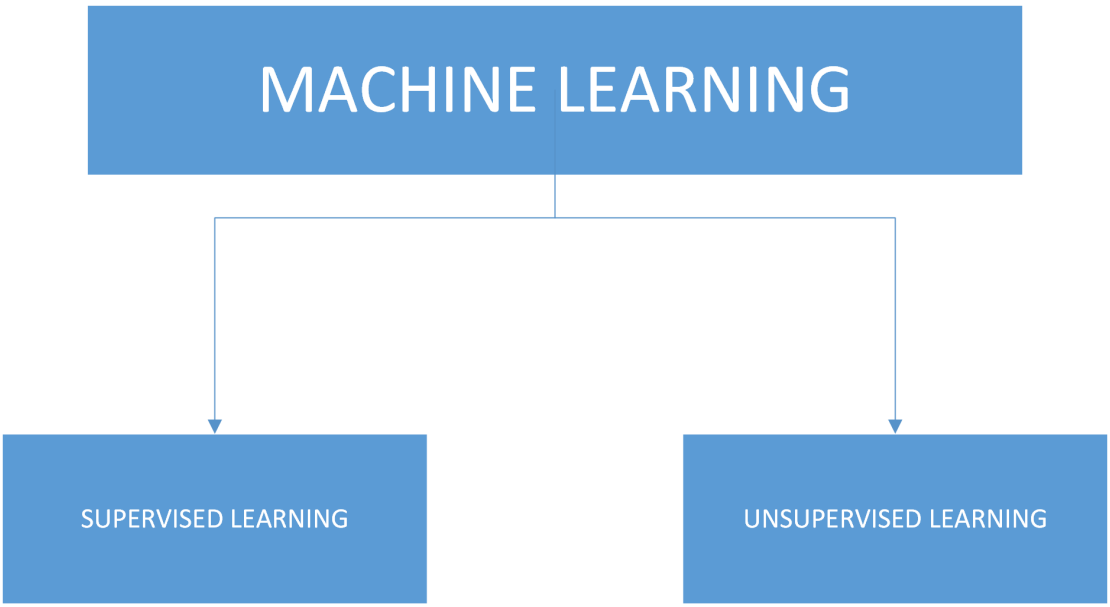

Figure 2.3: Types of Machine Learning

Machine Learning techniques develop an intelligent system based on the previous scenarios. ML techniques are widely being used in IDSs. Few of the ML techniques are:

- **Artificial Neural Networks:** ANNs are basically derived after being inspired from the biological neural networks of the brain. Couple of those neurons are connected to each other. To detect the intrusions and the future behavior of an individual, ANNs need must to be deployed properly. These networks are capable of learning after few steps of training so that can identify complex threats. There are two types of architectures of ANNs; feedback and feedforward as shown in table 2.5.

Table 2.5: Architectures of ANN

Feedback ANN	**Feedforward ANN**
Bi-directional signals	Uni-directional signals

  - *Feedback Artificial Neural Network*

    In feedback ANN, signals move in both directions i.e. from input to output and vice versa.

  - *Feedforward Artificial Neural Network*

In feedforward ANN, signals only move in one direction i.e. from input to output only.

The advantages of ANNs are:

1. Tolerance to inaccurate data.
2. Tolerance to uncertain information.
3. Ability to deduce solution from data without having previous information of it.

IDS based on ANNs are more efficient in working when properly trained with data sets. They are mostly used in recognizing the pattern of data, classification of data and even to predict the behavior of data. It has three layers; input, hidden and output layer as shown in figure 2.4.

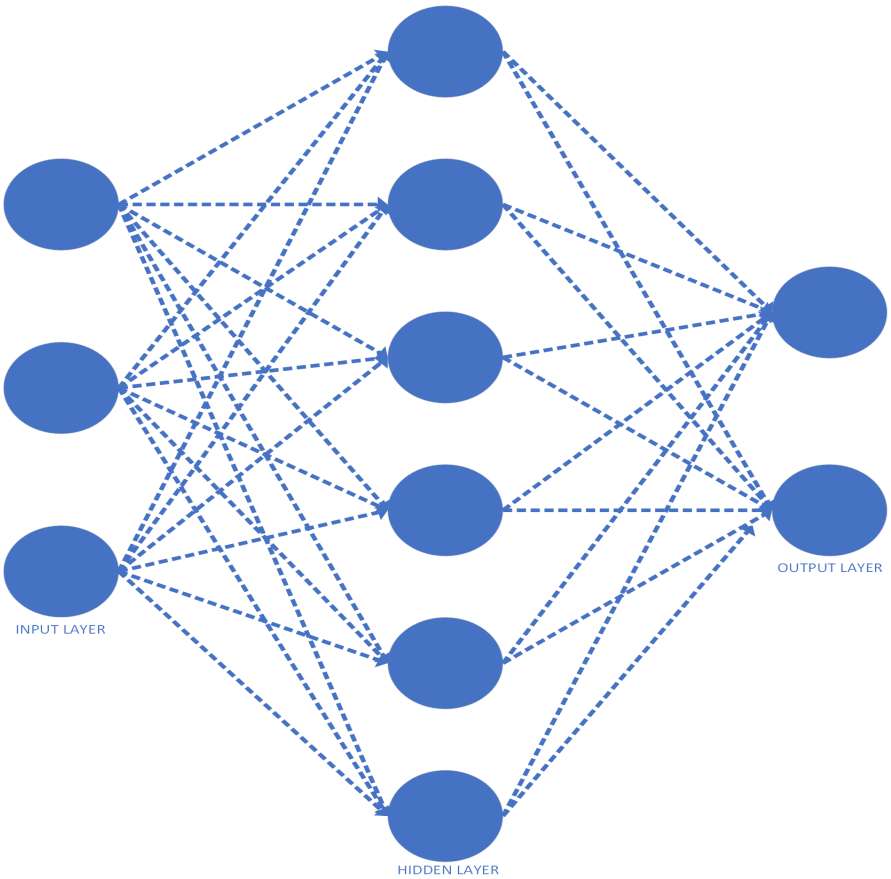

Figure 2.4: Basic Framework of ANN

Bahram and Nima [22] have introduced an IDS that combines the concepts of Fuzzy Clustering, Artificial Bee Colony (ABC), and Multilayer Perceptron (MLP). The proposed system works on three steps; training of components, validation of data and testing the system. Training is done using the fuzzy clustering technique while the performance of the system is examined in the validation phase. Testing phase include the passing of test data through model that was previously trained for intrusion detection. Clustering is performed to gather the homogenous data into clusters. Multilayer perceptron (MLP) differentiates between normal and abnormal data of the traffic. To train the MLP, ABC algorithm is used. Artificial Bee Colony (ABC) is a very fast, responsive and nature inspired algorithm that is based on the basic framework of population. ABC includes three groups of bees; employed, scouts, and onlookers. All of them have different functions. Employed bees share the information regarding food while staying in the hive and scouts are always looking for the new food sources while the onlookers search for the source of the food depending on the values of probability of the fitness that is provided by the employed bees. The formula to calculate this probability is given as:

$$P(r) = \frac{fitt(z)}{\sum_{r=1}^{M} fitt(z)} \; where \; \{r = 1,2,3,...M : Size\,of\,Population\} \qquad (2.6.1)$$

Here food source (z) is chosen by onlooker to calculate the probability (P) and fitt is the fitness of the food source i.e. nectar amount.

- **Bayesian Network:** It is a method of encoding probabilistic associations [20] among the variables of interest. This method is used in the domain of IDS by using statistical schemes. This combination of two domains allows to detect the intrusions more effectively.

- **Information Theoretic Security:** It is the security that is properly derived after studying how the information is being stored, quantified and communicated [20]. It assumes that any intrusion in the content of an information can cause irregularity in the data set.

Data Mining Methods: It is defined as discovering patterns, changes, intrusions, and the major constituents of data. In other words, data mining is a technique which takes data as an input and gives output of what it has extracted from the data [20]. The famous data mining techniques include:

- **Fuzzy Logic Techniques:** Fuzzy logic techniques are helpful in the domain of network and computer security because of few reasons like there are certain parameters that are used in intrusion detection e.g. the interval of connection, usage time of the CPU etc. that can be represented in the form of fuzzy variables. Also, the concept of the fuzzy behavior allows to smooth out the rapid transition of normal behavior to abnormality. The only limitation that it inherits is that the fuzzy rules are developed with the assistance of an experts of this domain. Nitin [21] in his work have integrated SNORT and Fuzzy Inference System. Snort is an open source [23] IDS and IPS module that performs analysis on real time network traffic and create the log files of IP networks. It is a tool that allows the user to design security frameworks very easily. It uses a library, libpcap (library packet capture), that most sniffers [24] and packet analysers use in TCP/IP model. Snort is hybrid IDS that used the concepts of both signature and anomaly-based intrusion detection systems. Fuzzy inference system [3] is based on the conclusions drawn from fuzzy rules. Fuzzy rule is just like an English statement; if-then. Using this system, we can easily map fuzzy inputs on fuzzy outputs. This mapping can easily be done by using simple fuzzy rules. These rules are the components of fuzzy inference system.

- **Genetic Algorithm:** It is a search technique that is used to find maximum solutions to the problems that are related to optimization. These are based on the basic concepts of natural selection, mutation theory, theory of evolution and inheritance. These techniques allow to differentiate between the anomalous and the normal traffic. Different types of attacks [26] could be classified using GA and applying rules to those attacks is possible. Dhak and Lade [27] proposed an IDS approach to detect the malicious activities. The model is connected with a firewall, it first gets the entries of firewall and then after the examination of

those entries it transfers them to GA based system. Output obtained from GA is forwarded to IDS. The proposed model is very effective and have detected maximum number of malicious connections. The main advantage is the robustness and flexibility of this technique, but it consumes high resources that is the biggest disadvantage.

Clustering Technique: This technique [20] work by clustering the trusted data into group based on their similarities. There are many ways to measure the similarity such as Cosine formula etc. The data that isn't the part of any cluster is considered an anomaly.

Decision Tree: Decision tree based techniques are used to classify and predict the data points. It has three [20] components:

- *Node*: Every node of a tree is labeled with unique feature or attribute that gives most of the information about its path from the root of the tree.
- *Arc*: Arc is labelled with the unique feature of the node.
- *Leaf*: Leaf is characterized by its class.

So, by the help of all these components decision tree locates the data point by initiating from the root moving through nodes and then reaching towards the leaves of tree.

Artificial Immune System: Artificial Immune System is a theoretical approach of human like immune system. It is defined as, "Systems that are adaptive after being inspired by theoretical immunology system and its principles and functions which are practically applied in catering different problems." This [29] approach is simply copied from the natural immune system, the way it reacts to external pathogens and fight against the harmful cells from bodies of individuals. It is designed to detect anomalous behavior of an intruder. The immune system of human is very complex and contain B-cells, T-cells and D-cells. B-cells (B Lymphocytes) are basically white blood cells (WBCs) present in bone marrow and they are liable for preparing antibodies. T-cells (named after Thymus) are formed in bone marrow but later they move to thymus where

they mature. They are responsible for recognizing antigens and destroying infected cells that are present in blood. D-cells or Dendritic cells [29] are present in blood that simply are responsible for collecting information about dead cells. It has few unique features like:

- Self-Organization
- Self-Adaption
- Self-Learning
- Distributed
- Diversity
- Dynamic

These features have encouraged a lot of computer security scientists to adapt this technique in different applications. A lot of researchers have worked on AIS based IDS. It also has feature of self-discrimination and non-self-discrimination that could be used to detect the anomalous behavior in IDS. Exactly like human immune system works against the pathogens to protect the human bodies, AIS provides a multiple layer structure that helps to protect the computer systems against cyber-attacks. Farhoud [28] et al. in their work have proposed a distributed framework for IDS based on AIS. They also have used the concepts of Genetic Algorithm (GA) in their work to improve the response of secondary immune. To apply AIS there are two approaches:

- **Self or Non-Self Discrimination:** AIS is designed in such a way that it can differentiate between self and non-self-space which is obtained by T-cells. T-cells contain set of non-self-detectors. Negative Selection Algorithm helps to differentiate between these two spaces. At first stage, detectors are trained with normal set of data to check either any detector is/isn't sensitive to self-data [28]. In case, if any detector matches with self-entity the system will remove such detectors and will keep the remaining. This method is helpful in efficient detection of IDS.

- **Danger Theory:** Danger theory works on a principle that human body is mainly dependent upon the tissues of the body rather than immune cells. Danger signals are released by the tissues that are distressed to activate the immune response while calm signals released by tissues that are healthy and they provide tolerance to the immune system [27]. In this approach, agents are used that communicate intelligently to detect the intrusions in system. Properly designing an IDS is important for the efficient detection of intrusions while the poor design will simply reduce its capability of detection.

2.7 Summary

This chapter gives a brief introduction of an intrusion detection and intrusion prevention systems. Section 1 covered the detection methods, components and types of an intrusion detection systems (IDSs). In a next section, discussion related to the introduction, methods of prevention and types of intrusion prevention systems (IPSs) was covered. Later, the analysis part of intrusion detection and prevention systems is presented in a tabular form. In next section, the role of artificial intelligence in the field of computing frameworks. The introduction of artificial intelligence is discussed followed by the brief history of artificial intelligence. The way artificial intelligence evolved in an era of information technology. In the last section, it has been discussed that how artificial intelligence integrates with the principles of network security in the form of intrusion detection and prevention frameworks.

Chapter 3

Cloud Computing

3.1 Introduction

Cloud computing is generally defined as “Computing Everywhere”. Cloud computing is IT model capable of computing composed of components (networking, software, services, and hardware) of information technology to deliver the services of cloud using internet of any private network. It stores, manages, designs, and provides the services as per the requirements. National Institute of Standards and Technology (NIST) has defined cloud computing as, “Cloud computing [30] is a model for empowering pervasive, helpful, on-request organize access to a mutual pool of configurable processing resources (e.g. systems, servers, applications, and administrations) that can be quickly provisioned with negligible administration effort or with an interaction of organizations. The model of cloud computing is consisted of five characteristics, three administration models, and four models of deployment.” It is clearly stated in the definition that cloud computing helps any organization to reduce its computational resources and mainly the burden of its users. In short, we can say that it can help an organization in many ways.

3.2 Characteristics of Cloud Computing

According to NIST [30] cloud computing has some unique characteristics that make its different from tradition computing technologies. Those characteristics are given below:

On demand self-service: Users can select the services (storage, platform or infrastructure) as per their needs without needing any interaction of humans with the service provider for that.

Resource pooling: On the providers' end the services are pooled together to cater numerous consumers altogether by utilizing multi-tenant model. A sense of location independence is observed in cloud computing framework where the user has no access to control the physical location of its data but general information at a higher level of abstraction can be provided by CSP.

Broad network access: All the services for which the user has signed up are available over the network though standard mechanisms that can be accessed any time and at any place in this world.

Measured Service: To keep the transparency in mind for both the consumer and provider, it is possible to control, report and monitor the resources.

Rapid elasticity: The services provided by the CSP can be elastically modified as per the needs and the requirements of the user. If at anytime the user wants to increase the capacity of its services, user can immediately notify it to the provider and from the provider end the capability of the resources can be increased and decreased as well.

Programmatic Access: The services provided by CSPs can be accessed by using different APIs (Application Programming Interfaces). APIs are also highly likeable by the users because they can access their resources anytime around the world.

Greater Automation: Scaling is a factor that is impossible to control manually in a computing world, so it must need to be automated. Automation also benefits the cloud users for more competitive services.

Reliability: In cloud computing, reliability is considered as the core feature in security.

Today, many organizations are shifting from the tradition computing technologies to cloud computing because of its good features and characteristics. Traditional computing frameworks were lacking few of these features that makes cloud computing highly recommended among the large-scale organizations.

3.3 Service Models

Service models are basically the frameworks that are consumed and delivered over the private network or an internet. Depending upon the multiple requirements of the user service models may vary. NIST [30] has defined three type of service models:

Software as a Service (SaaS): In this type of service model, an access of a software or an application is provider to the user that is running on the cloud computing framework. An application can be accessed by the user on any device. User isn't given any access to the underlying infrastructure like servers, storage capacity, networks etc. but only the interface of an application.

Platform as a Service (PaaS): In this model, user is provided with a capability to deploy the applications created by cloud users. These service models also support the programming languages tools that are required to create different type of applications on these service models. In PaaS, user isn't provided with an option to manage the cloud infrastructure like OS, storage, servers, but is only given an access of deployed programs or applications.

Infrastructure as a Service (IaaS): In this service model, user is provided with a capability to control and manage storage, processing and other computing resources like operating systems and different applications. In IaaS, user is given an access to

deployed apps, OS, storage and a limited control of network tools.

All these service models are referred as SPI model and can be layered as Infrastructure as a Service (IaaS) as a building block of Platform as a Service (PaaS), and Platform as a Service (PaaS) as a building block of Software as a Service (SaaS) otherwise stand alone as shown in figure 3.1. It can be said that, "Platform [31] as a Service (PaaS) and Infrastructure as a Service (IaaS) are the special versions of Software as a Service (SaaS) that allow to deploy new cloud services."

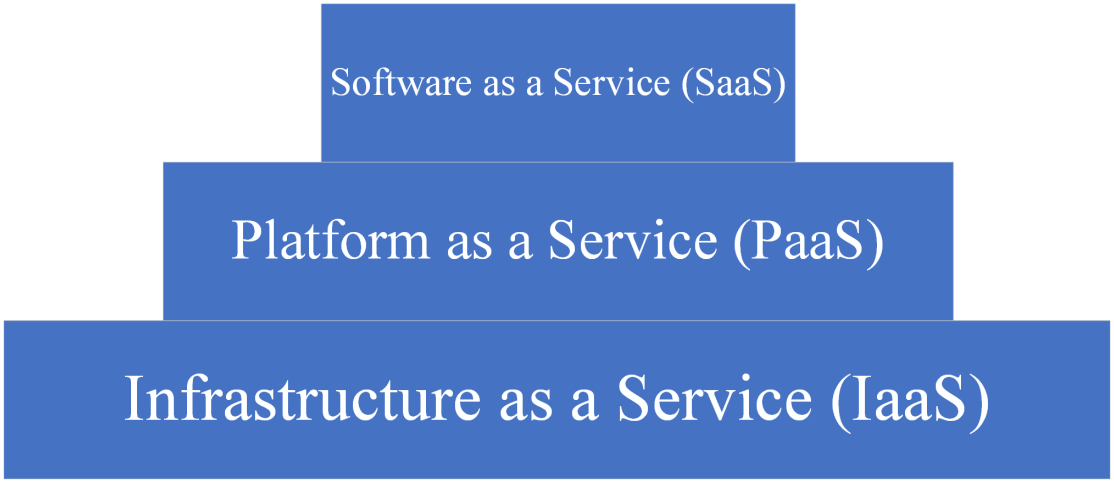

Figure 3.1: SPI Model: Tier Down Approach

In a view taken by Cloud Security Alliance [31], "Infrastructure as a Service (IaaS) is the building square of all cloud administrations, with Platform as a Service (PaaS) expanding upon Infrastructure as a Service (IaaS), and Software as a Service (SaaS) thus expanding upon Platform as a Service (PaaS) as shown in figure 3.2. In this way, just as the capabilities are inherited, so are the information security risks."

Figure 3.2: SPI Infrastructure: Underlying Services to Frameworks

Because of the level of abstraction of the services offered by CSPs, the control over security by a user may vary. Moving from IaaS to PaaS and then from PaaS to SaaS, computing functions get abstracted more and more. So, in other words the highest level of control over security is offered by IaaS, after IaaS comes PaaS and at the end SaaS as shown in figure 3.3.

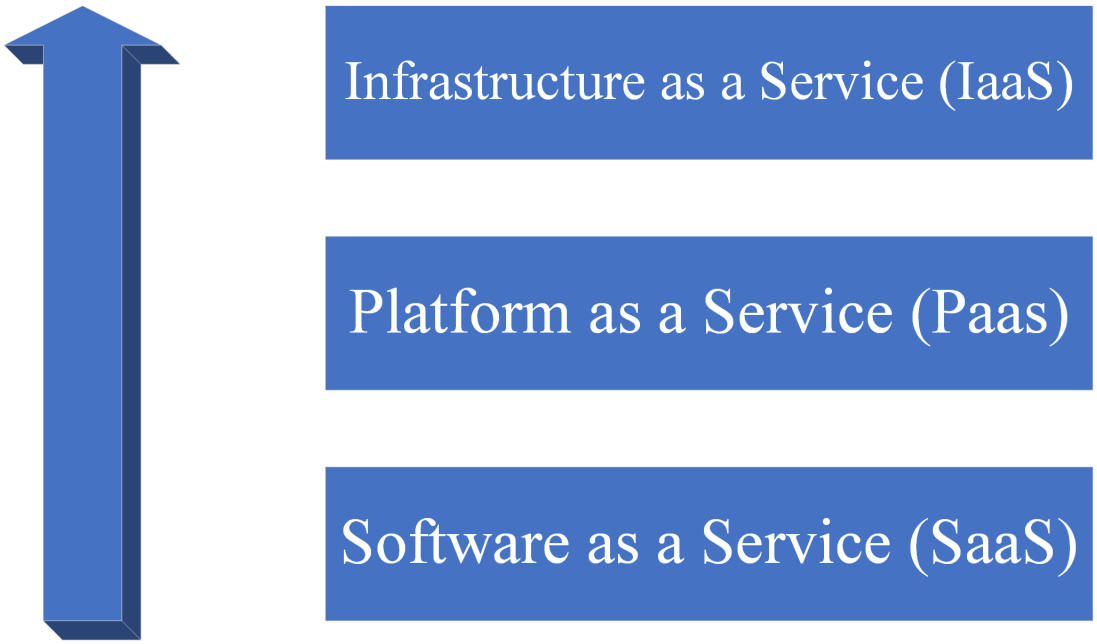

Figure 3.3: Users Control over Security

Beyond these basic service models, now-a-days because of the innovations in technology other service models have been proposed that are:

- IoT as a Service (IoTaaS)

- Identity as a Service (IDaaS)
- Security as a Service (SECaaS)
- Data Center as a Service (DCaaS)
- Monitoring as a Service (MaaS)
- Hardware as a Service (HaaS)

Generally, deployment of a service can be represented as XaaS, where X can be any service delivered to the consumer. Due to this expansion in the service models it is evident that the SPI model isn't necessary to be universal.

3.4 Deployment Models

In the previous section, we have covered the service models of cloud computing that are SaaS, PaaS and IaaS and few other extended service models. But now the point is that where will these services model will be deployed. According to Mell and Grance [30], there are four cloud deployment models: public, private, community and hybrid.

Public Cloud: A cloud model in which the services are delivered to the public and large organizations providing cloud services [30]. Public cloud exists in external premises to its users and is normally accessible with very little restrictions. Among all the deployment models, it is most commonly used by the users. Examples are EC2, IBM's BlueCloud, Cloud Front etc.

Private Cloud: It delivers [30] the services only to an organization. It is a responsibility of an organization to manage it or an organization may hire a third party to operate it. Unlike public cloud, private clouds are hosted internally.

Community Cloud: The infrastructure of cloud that is shared by multiple organizations that share the same concerns and targets a common group of people [30]. Like a

private cloud, it also can be managed by the people of organization or an organization may hire a third party to manage their cloud services.

Hybrid Cloud A combination of two or more cloud deployment models that are bound together by policies based on different standards as depicted in figure 3.4. But it doesn't mean that the entities also share their characteristics [30].

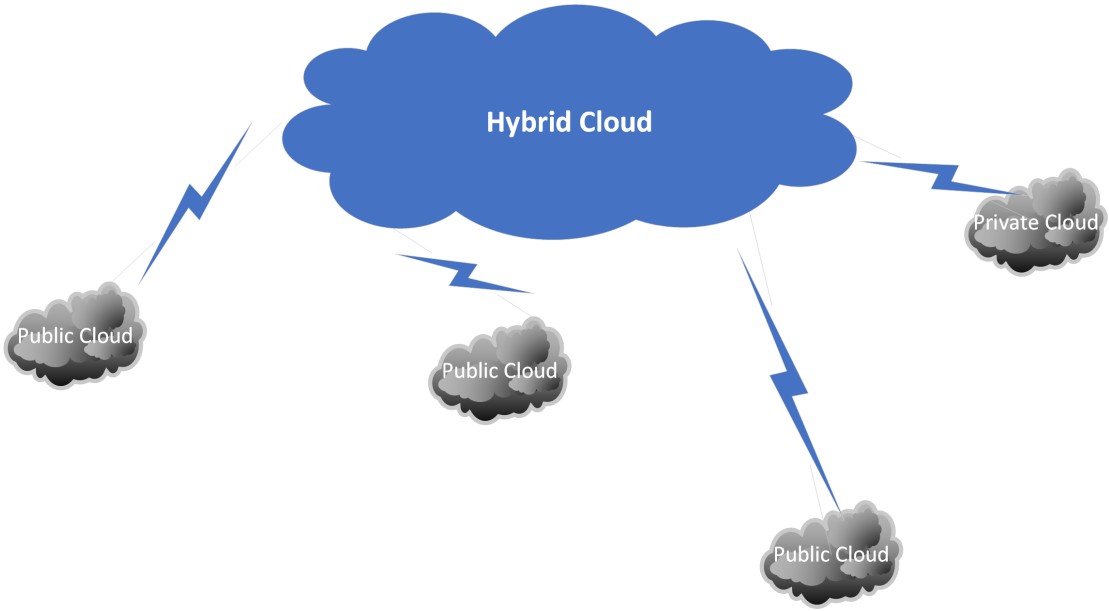

Figure 3.4: Hybrid Cloud

Keeping the point of control over security in mind, we can say that a user gets more control when it moves from public to community cloud and from community to private cloud as shown in figure 3.5. Hence, the private cloud is the most secure amongst deployment models of cloud computing.

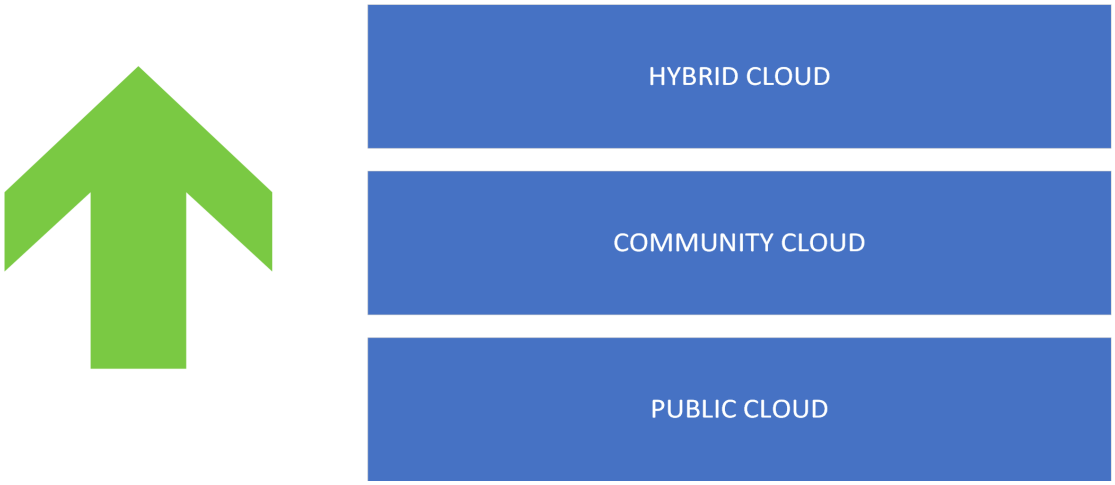

Figure 3.5: Control over Security in Cloud Deployment Models

3.5 Federation in Cloud Computing

The combination of multiple cloud providers sharing resources, services, infrastructures, or platforms is cloud federation as depicted in figure 3.6 [2]. In cloud federated environment, different cloud providers based on trust and assumptions share their resources with each other. Mainly, federated organization share attributes of users' identity and access permission using access requirements. In simple words, let's assume that X cloud service provider is only providing its customers software as a service and the customer A is completely satisfied with the deliverance of services by X cloud service provider. But later, customer A thought that he along with software services he needs a large capacity of storage as well. Now there are two options in front of him.

1. He will go to X cloud service provider and asks for the storage service.
2. He will go to Y cloud service provider who is offering storage services.

In the first scenario, the trust of a user on X cloud service provider will be very high because of his satisfaction with the services he already has used. But as cloud service provider X is only providing software services, the customer A's request for storage service will not be entertained. On the other hand, the trust of the user A on the Y cloud provider would be zero at the start. That simply means, there would be a certain probability that customer A might not be willing to get Y's storage services. Now that's

a problem that what customer A will do to get the storage service? To cater this problem, cloud federation has a solution. Cloud Federation allows multiple cloud entities to share the services based on some policies and standards to meet the trust factor. For the above-mentioned problem another cloud provider Z will develop a mechanism, based on trust, that will get the services from X and Y cloud provider and will manage the both services at a single platform. Now customer A being a part of CSP X, can go to the CSP Z (the federated identity) and will ask for the trust and assumption policy. When customer A agrees, he can sign up for the both services (software and storage) provided by CSP Z.

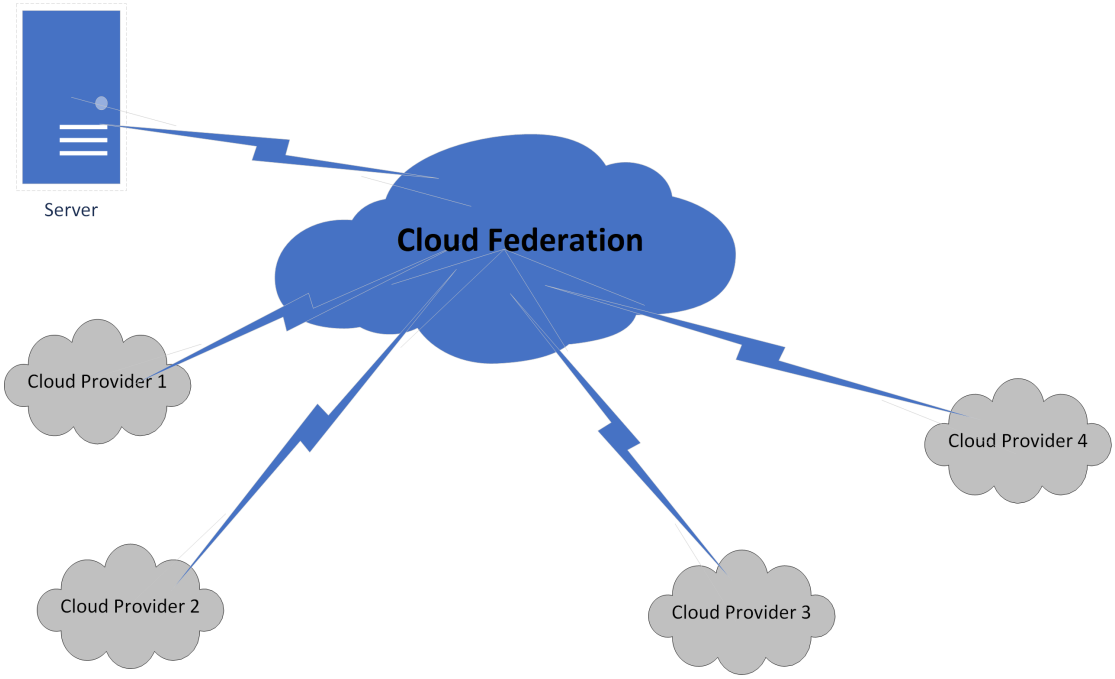

Figure 3.6: Federated Cloud Computing Environment

Federation in cloud was basically an idea that was proposed to reduce the computational resources of users. With the other advancements in cloud computing this concept has also gained an interest of researchers.

3.6 Security Concerns in Cloud Computing

Access to computer systems and mainly to the data, residing in them is the serious topic of concern these days. With every passing day, we can see number of cases reporting data stealing; leakage of sensitive and personal information. Hence, there is a need to secure everything that is available online. Cloud Computing is the growing area in field of IT where users can access there data anytime at any place using internet. Cloud Computing offers many good features, as on demand self-service and wide-ranging network access but still privacy and security of the data are the major issues [2].

In cloud federation security, the main goal is to develop the mutual trust relationship between multiple federated identities and the components that are involved within the cloud. Such mutual [9] trust can be developed by securing the components that are involved in the cloud computing framework. Mainly, those components are servers, clients, networking and communication platforms.

Federation of cloud allows users to exchange data within federated identities that can allow malicious users to gain access of the servers of other cloud providers. Gaining an access to data of other CPSs can allow a malicious user to perform certain type of cyber attacks like DDOS and cloud malware injection. These attacks cannot only result in the poor performance but can also affect the economic stability of a cloud service provider. Few of the security concerns [39] in cloud computing are discussed below:

Transparency: Transparency of the services delivered to users in a cloud computing framework is a major security concern. When a CSP doesn't exposes the details of its security policy or the implementation of their technological framework, the users must trust the CSPs' claims of security. But still a little information of the privacy of data, security mechanisms and how the incidents are being managed must be provided to the user. In an interview [31] it was concluded that, "By the security experts, transparency of security is considered as the main security concern in cloud computing framework."

Disaster Recovery: In case of any disaster, a confidence of users must be maintained for the continuity of their services in cloud.

Availability of network: It is one of the requirements by the cloud users that whenever a cloud service consumer (CSC) needs service the cloud must be available. In case of unavailability, it won't be a different scenario other than a DOS (Denial of Service) case.

Security Incidents: Occurrence of any incidence in cloud computing must be notified to the user on a very immediate notice. Also, in that case users must support and cooperate with their respective CSP.

False Sense of Security: Most of the companies now-a-days are claiming the full deliverance of data protection and security to their users. But on the other hand, there are number of data breaches occurring every year and at the end when the forensics team analyses the reason of breach then customers realize about the standards of the policies the company was practicing. It targets to the false sense of security. To make you a customer, they will completely satisfy you by lying about security protocols they have implemented. But, you'll not find any such protocol being implemented in their systems.

Geo-location: It has also become a key concern to monitor the data of the customer where it is physically residing in a cloud. Data migration is quite a familiar term in cloud computing, because to balance the load many times the data of the customer might be migrated from one point to another. But migrating data may violate the Service Level Agreement (SLA). It is possible for a malicious CSP to relocate the data which may jeopardize the privacy and security of customers' data.
Listing the security concerns benefits to validate them with the controls. Keeping these security concerns in mind, we can design an efficient model for cloud computing.

3.7 Summary

In this chapter, a brief introduction of cloud computing framework is discussed along with the characteristics of cloud computing framework that made it unique from the other computing technologies. NIST paper that presented the definition of cloud computing, the three service models and four deployment models were covered. A comparative study of control over security in service and deployment models was highlighted. After that, how cloud federation is different from normal cloud computing model was discussed in detail. In the last section of this chapter, the major security concerns in cloud computing were dealt.

Chapter 4

Data Security in Cloud Computing

4.1 Introduction

This chapter covers the concepts of data security in cloud computing model. It involves multiple complex approaches to secure data in cloud than simple encryption mechanisms in traditional computing models. The requirements for data security in cloud may vary depending upon:

- Service models (SaaS, PaaS, or IaaS)
- Deployment models (Public, Private, Community, or Hybrid)

In a cloud model, data is available in two forms; data at rest, data in motion. To completely secure data in cloud, data in both scenarios must be protected. Encryption is a key component in cloud computing, but in case if the key isn't kept secured then even a most robust cryptographic algorithm is of no use.

4.2 Data in Motion

When the data moves as a file from one location to another location, the data is said to be in motion. The data we upload to the cloud, during uploading the data is basically in motion at that time. To authenticate with any web server or even with a cloud provider,

we enter username and password that is not usually stored in encrypted form. It is possible for an intruder to tamper the data that is in motion and even the third party introduces [39] risk factors. The best possible option to secure the data in motion is data encryption.

4.3 Data at Rest

The stored data in computer, data servers, virtual servers, or anywhere where the backup of data is created is basically considered as a data at rest. Protecting data inside cloud is not different in any way rather than securing data that isn't a part of cloud. The secure the data in both scenarios, similar security mechanisms could be applied [39]. In case if you're going to store your data on external cloud provider a risk of an exposure of the data can't be denied in that scenario. Another requirement is that CSP will guarantee the complete security of data. So, a verifiable and strong attestation of the practices and the implementation of the security modules the CSP is following must be done before selecting any CSP.

4.4 Data Encryption

Data encryption is a first thought when a concept of data security specially when data at rest is discussed. Encryption [39] is an art of making a plain text so confused that it becomes unreadable to those who don't have the key to decrypt that text. There are two methods of data encryption:

Symmetric Encryption: In a symmetric encryption, the key that encrypts the data is used to decrypt the same data.

Asymmetric Encryption: In an asymmetric encryption of data, the key to encrypt data differs from the key that decrypts it.

In both cases, symmetric and asymmetric, the challenge lies in keeping the key secret

from an adversary. Keeping the key secret is the major security concern in cryptography. That is because it is difficult for any person to memorize those lengthy keys. So, the key must be stored on the system or on the network where the data resides. But on the other side, it is too much risky to store the key there because of the network attacks.

4.5 Mistakes with Data Encryption in Cloud Computing

For communication, users are provided with the cryptography mechanisms so that they can secure their private digital files. As commonly as it is available, it is not effectively in use by the users to secure their data. Researchers in this cryptography domain have worked on this to find out the reason behind this failure of cryptography. Few of the major reasons behind this failure are:

- Transfer of sensitive files such as passwords by attaching them with an unencrypted mail.
- You are working on designing your own algorithm that will secure your data in future. But in present you aren't using any of the cryptography protocol to secure your files.
- Even after knowing that cryptographically secured versions of web-based protocols like HTTPS (Hyper Text Transfer Protocol Secure) are available, still using the previously designed protocols like HTTP (Hyper Text Transfer Protocol).

There are so many other such problems that are listed by many researchers. Keeping the scope of our topic, the discussion is kept limited, but the reader is encouraged to cover the widely available research material available on this topic.

4.6 Categorization of Sensitive Data in Cloud Security

In cloud security, after a very little modification in the techniques of securing information that were available already to secure the data in traditional computing models is enough. Few of the techniques that are applicable to achieve the data security in cloud computing are listed below:

Data Encryption: The best way to secure data in motion is to implement a model that integrates the concept of cryptography with data authentication. Two major goals of securing the data in motion:

- When the data is in motion, its confidentiality should remain undisturbed.
- No one, other than sender and receiver, can tamper with the data in transit.

Encryption assures the user that in case of any data breach, the data may remain confidential. Authentication assures that the data can only be accessed by the trusted user or in simple words the person who has rights to read the file.

Access Control: Access control techniques are helpful in a manner where an organization wants to maintain an environment of separation of duties. In any organization, there are different level of employees who are categorized based on their capabilities or roles to perform multiple tasks. Access control mechanisms are highly efficient, but their dependency relies on capability of identity management. Fig. 4.1 shows the fundamental model of access control.

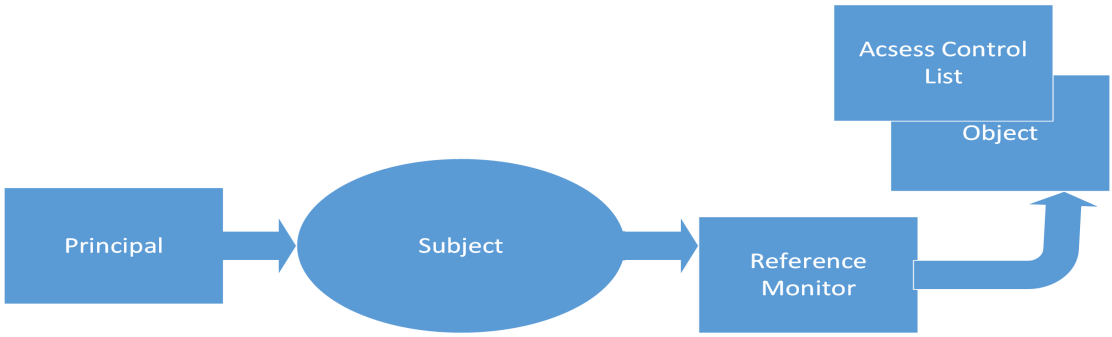

Figure 4.1: Fundamental Model of Access Control [40]

Identity management is based on two things; subjects and objects. Objects refer to files or the documents that could be accessed by people (subjects). There are three types of access control [40] models as depicted in figure 4.2. These includes:

- **Mandatory Access Control (MAC):** The access policy that the system determines and implement by sensitivity labels of data located in the system. Level of trust [40] can be determined by the label of a subject while the objects' label determines the level of trust to access it. In simple words, to access any object, subjects' label must need to be higher than the objects' label.
- **Role based Access Control (RBAC):** RBAC is determined by the system. It is based on the role of a subject. In RBAC, a subject can access an object iff their role permits to access it otherwise not [40].
- **Discretionary Access Control (DAC):** In DAC, owner of an object has the rights to give access to an object. An owner [40] can decide who can access the object and who cannot. DAC is useful where the user group is limited so that the permissions can be maintained easily.

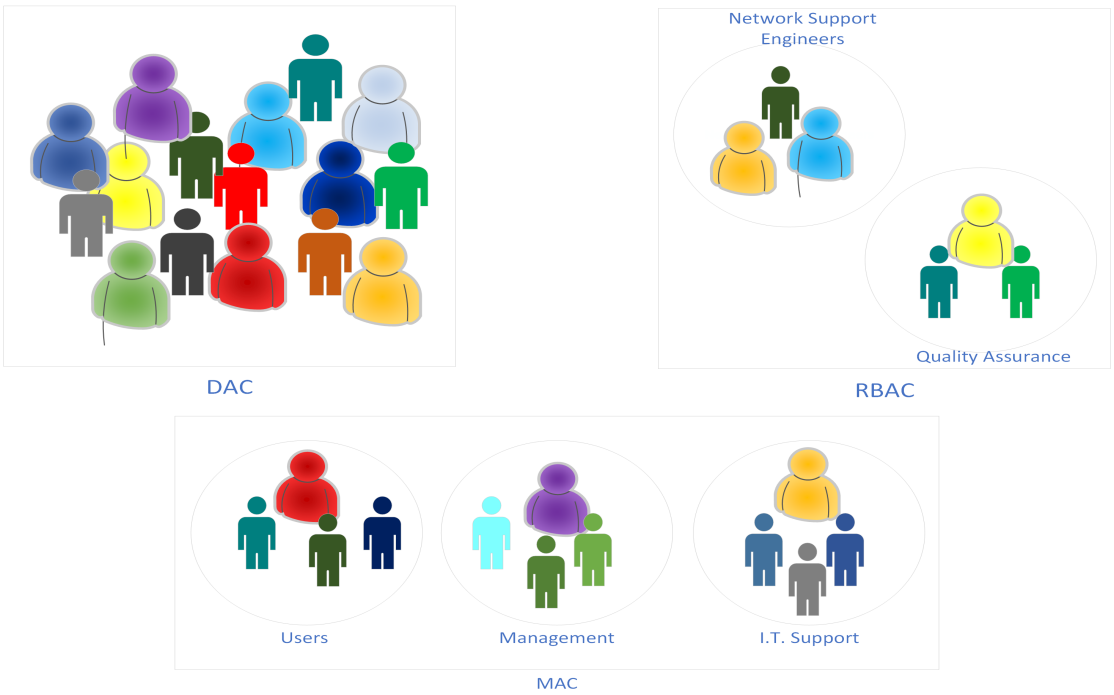

Figure 4.2: Access Control Models

By definitions, we can note that these access controls may vary a lot in different ways but still they can be implemented together to get a desired output from any system [39].

Data Labeling: Data labeling is considered as the security mechanism in such a sense the data may be protected from being tampered with if the data is labeled by keeping its level of sensitivity. By labeling of data, it becomes easy to categorize it. Usually the data is classified as:

- Top Secret
- Secret
- Unclassified
- Confidential

Data labeling can help a user to identify the data according to its characteristics. This labeling of data has greatly evolved with the other information security techniques. Data labeling may also relate to access control mechanisms in such a way that to access a highly confidential data, subject having high level of trust must be allowed to access it. Similarly, only trusted users must be given access to read secret and top-secret documents. The whole process of data labeling is bit complex but when implemented with a great consideration of correctness of implementation. In data labeling, there should be a check on over classification of data. Over classification of data may result in the poor handling of data, because if everything is treated as a secret or a confidential data, then it would be very difficult to secure and implement such system that will categorize the data based on their labeling [39].

Authentication: Authentication is a mechanism to identify someone's identity based on some parameters. To authenticate a subject for any object, the user or a subject must be in possession of authentication factors. Factors like login passwords and use of digital certificates are the examples of authentication factors. Two factor authentication is a type of authentication that requires two factors to authenticate a subject. If in case,

hacker has bypassed a single factor authentication mechanism then the second factor in two factor authentication will be another challenge for a hacker to bypass that system. Authentication factor is dependent upon the identity management system access control mechanisms [41].

Data Masking: It is a technique to protect sensitive information by changing certain elements of data, but the overall structure remains the same. The main goal in data masking is to replicate the data in such a way that if someone re-engineers the data, the main characteristics of data revealing the critical information won't be exposed in any way. Data masking targets [39] to ensure the deliverance of the privacy of data. In order to fulfill the purpose of data masking, it should be done very carefully so that the available data should not reveal the sensitive information.

Secure Deletion: To delete the confidential data in cloud, it must be clearly understood how that data is disposed from the cloud. The deleted data can be accessed from the archives and from the place where the backup of the files has been created. To recover the deleted files, IT skills are required. If a user deletes data from cloud on its end but a cloud service provider has mentioned in its policy for providing data backup facility for at-least three months, that means even after deletion that data could be restored from a CSP end anytime. So, there is a need to read data backup policy before signing up for the services [39]. The US Department of Defense (DoD) has provided the two keys aspects of data deletion in cloud:

- **Sanitization:** Sanitization is the way toward expelling the information totally from media before reusing the media in a domain that doesn't give an adequate level of insurance for the information that was in the media previously before sanitizing it. Data Security resources ought to be cleaned before discharging them from the ordered controls of a data or when decreasing the level of classifying them.

- **Clearing:** Clearing is the way toward deletion of the information on media before reusing the media in a domain the gives a satisfactory level of security for

the information that was on the media previously clearing. All inner memory, support, or other reusable memory will be cleared to adequately deny access to beforehand available data.

It has been proven that erased data can be recovered. It is because, the data on a disk is stored in a form of electric charges or in magnetic form. Advancements in the field of computing have introduced techniques to recover the deleted data very efficiently.

4.7 Summary

The security concerns that revolve around securing the data in cloud computing framework has been discussed in this chapter. After analysis, conclusion has been made that those concerns aren't inherently unique compared to data that is stored within the premises of an organization. First section defined the scope of data security in cloud computing. Next to it, the concepts related to data in transit and data at rest were covered. Further, data encryption is covered as a part of this chapter followed by mistakes of data encryption in cloud computing. At last, the chapter was concluded at the discussion of categorization of cloud data security. Different techniques to achieve the data security in cloud computing are discussed.

Chapter 5

Proposed Framework and Implementation of Results

5.1 Introduction

This chapter covers the results and the experimentation of this thesis. The implementation of the proposed solution will be described more than in an approach. There will be used few diagrams and figures to explain examples. The explanation will be kept as simple as possible to allow a reader to good understanding of our proposed solution.

5.2 Proposed Solution

Knowing that, in almost every field of science; artificial intelligence has proven its advantages like starting from the very basic concepts of human birth and moving towards the rocket science concepts of space everywhere artificial intelligence has played a very significant role. Similarly, in solution of our proposed problem AI techniques have been used that could detect the intrusions in federated cloud computing environment. Artificial Intelligence techniques are much accurate than any other techniques in designing an intrusion detection and prevention systems. In today's era, every bit of information is being encoded to binary digits that have created huge amount of data in computing world that is difficult to manage. With the increase in amount of data, the types of at-

tacks have also increased. Hackers are trying their best to find new ways to play with the data and to bypass the security mechanisms. So, keeping in mind the pros and cons of AI techniques that focuses IDS/IPS have been designed that first looked into the data, checked the behavior, and after that managed the data according to its behavior. Due to the flexibility and adaptability to the computing system, AI techniques are highly likable as compared to the other techniques used in intrusion detection and prevention systems.

Artificial Intelligence techniques are highly adaptable to different environments in computing but for the proposed problem after analysis, it was concluded that Artificial Neural Network based AI technique is the best option. Using artificial neural network, we can detect the intrusions based on anomalies in federated cloud computing paradigm.

5.3 Proposed Framework

Proposed framework using ANN is that an IDS will be integrated into the FCC environment where it will keep on monitoring the real time traffic going through the network. Neurons of input layer will take the input from network and will forward it to the hidden layer where multiple neurons will perform some mathematical functions and after computing the result will declare that either the traffic from an input is a normal traffic or malicious. The computed result will then be transferred to the output layer. In case of a normal traffic, IDS will allow the traffic to pass through it but if any intrusion is detected it will forward it to an IPS where after further examination it will be decided either to allow or to deny the network traffic to proceed into the network. Using the work, our results have shown that we can stop a malicious user to gain an access of the sensitive data stored on the cloud. Fig. 5.1 shows the visual representation of our proposed framework.

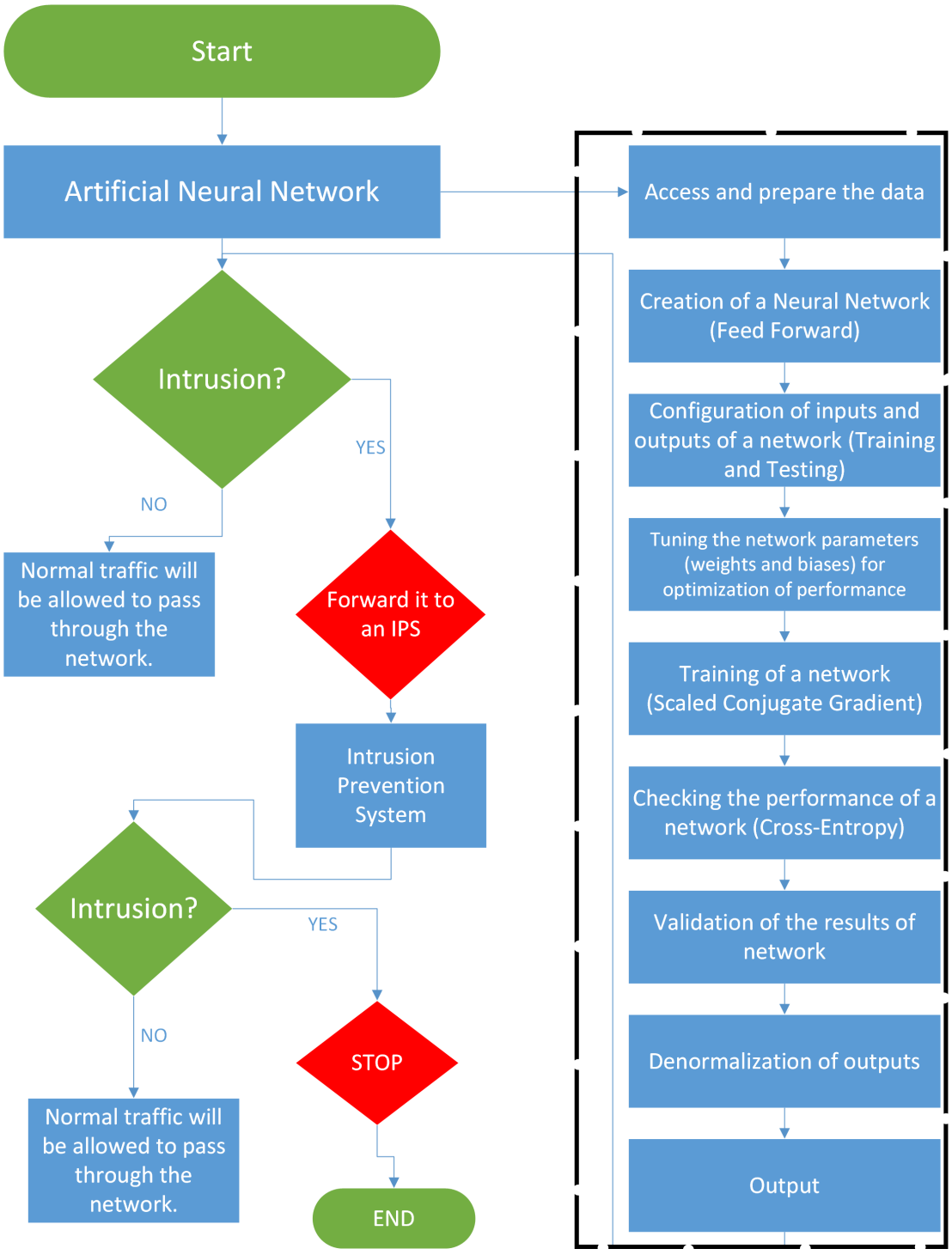

Figure 5.1: Proposed Framework

5.4 Environment and Data Set

To perform the experiments, MATLAB® has been used because of its advantages over other tools (advantages will be discussed later). For this purpose, UNSW-NB15 data

is used. In the previous works, researchers have been using KDD'99 for intrusion detection systems. But after analysis it was concluded that, data set UNSW-NB15 is having few advantages over KDD'99. The comparison of UNSW-NB15 and KDD'99 is covered in the table 5.1.

Table 5.1: Comparison of UNSW-NB15 and KDD'99

Parameters	UNSW-NB15 [36]	KDD'99 [37]
Number of IP addresses	45	11
Time of Data Collection	16 hours	840 hours
Attack Families	9	4
Format of Collected Data	Pcap files	Dump files, tcpdump and BSM
Total Number of Networks	3	2
Simulation	Yes	Yes
Extracted Features	49	42

5.4.1 UNSW-NB15 Data Set

UNSW-NB15 [36] dataset was made utilizing an IXIA Perfect Storm apparatus in the digital laboratory of the Australian Center for Cyber Security to create a hybrid of the practical latest trends of ordinary activities and the engineered contemporary attack patterns from system activity. A tcpdump instrument was utilized to catch 100 GB of raw traffic of a network. Argus, Bro-IDS apparatuses were utilized, and 12 models were created for separating the features. The UNSW-NB15 dataset is categorized as normal and attack. Further attack records are classified into nine families based on their nature of attack.

- Fuzzers
- Analysis
- Backdoor
- DoS

- Exploit
- Generic
- Reconnaissance
- Shellcode
- Worm

Due to the large set of attack types and other advantages of UNSW-NB15 it was decided to be used in the experiments.

5.5 Terminologies

In the experimentation phase on each network, the following plots were created:

Confusion Matrix: A confusion matrix is a procedure for outlining the execution of a characterization calculation. Accuracy of classification alone can be misleading in the event that you have in excess of two classes in your dataset. The complete confusion matrix is shown in figure 5.2.

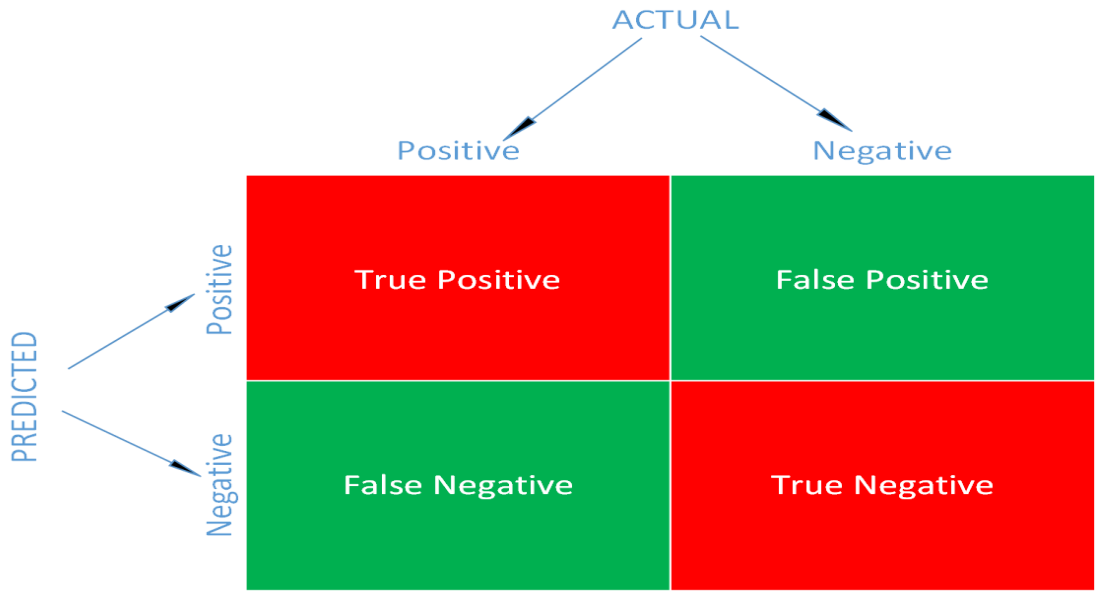

Figure 5.2: Confusion Matrix

Figuring a confusion matrix can give you a superior thought of what your characterization demonstrate is getting right and what are the redundancies in achieving your target. A confusion matrix is a rundown of expectation results on a classification issue. The quantity of right and wrong expectations is abridged with count values and separated by each class. This is the way to confusion matrix. Confusion matrix gives you the errors of your designed model as well as the kind of errors too. Fig. 5.3 shows the detailed confusion matrix.

To understand a confusion matrix, following are the relevant terms that must needs to be understood.

1. False Positive: Incorrectly predicted event values.
2. True Positive: Correctly predicted event values.
3. False Negative: Incorrectly predicted no-event values.
4. True Negative: Correctly predicted no-event values.

To calculate the values for this confusion matrix, following are the important formulas.

$$Accuracy = \frac{True\ Positive\ +\ True\ Negative}{Total\ Samples} \tag{5.5.1}$$

Accuracy is the percentage of how much correctly the classifier algorithm is working. Equation 5.5.1 is used to calculate the accuracy of an algorithm.

$$Misclassification\ Rate = \frac{False\ Positive\ +\ False\ Negative}{Total\ Samples} \tag{5.5.2}$$

Misclassification rate is the percentage of how much the classifier went wrong. Misclassification rate can be calculated using equation 5.5.2.

$$True\ Positive\ Rate = \frac{True\ Positive}{Actual\ Yes} \tag{5.5.3}$$

TPR is also known as sensitivity or recall and can be calculated using equation 5.5.3. Further, few important values (False Positive Rate, Specificity, Precision, and Prevalance)

can be calculated using equations 5.5.4, 5.5.5, 5.5.6, 5.5.7.

$$False\ Positive\ Rate = \frac{False\ Positive}{Actual\ No} \tag{5.5.4}$$

$$Specificity = \frac{True\ Negative}{Actual\ No} \tag{5.5.5}$$

$$Precision = \frac{True\ Positive}{Predicted\ Yes} \tag{5.5.6}$$

$$Prevalance = \frac{Actual\ Yes}{Total\ Samples} \tag{5.5.7}$$

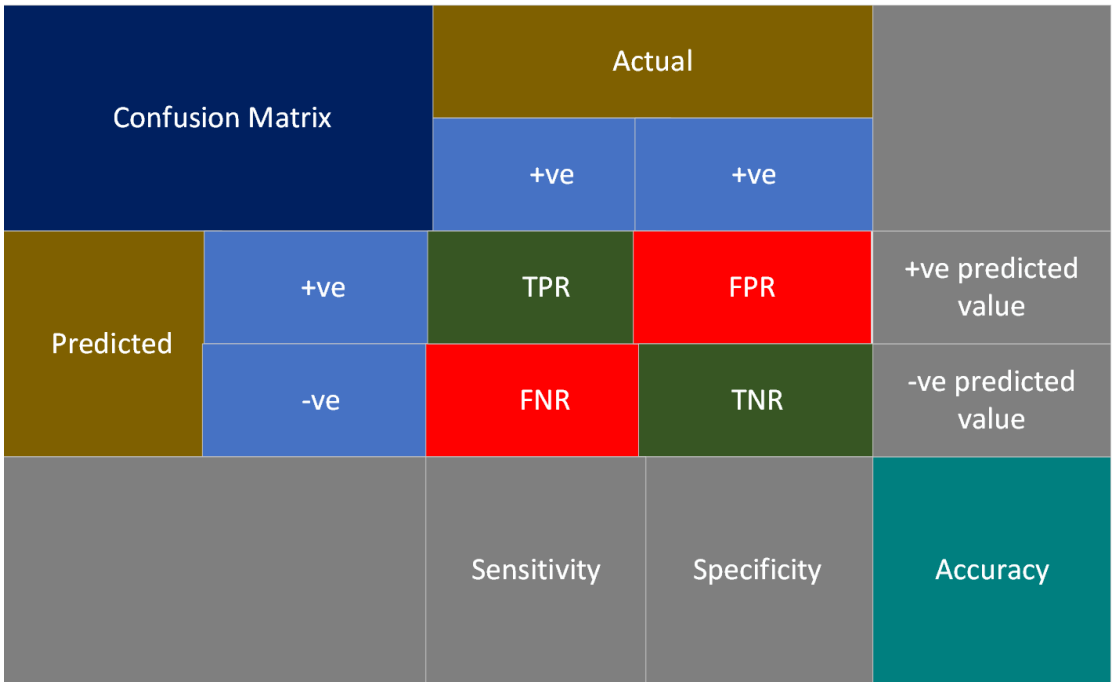

Figure 5.3: Detailed Confusion Matrix

Receiver Operation Characteristics (ROC): Receiver Operation Characteristics enlightens us regarding how best the model can recognize two things. Better models can precisely recognize the two. Though, poor models will experience issues in recognizing the two. To plot ROC, we use sensitivity and specificity. Both these variables, sensitivity and specificity, are in inverse relationship to each other. As sensitivity decreases, specificity increases and vice versa.

But that is not how we plot ROC curve. To plot ROC curve, we use (1-Specifity) instead of just using specificity. So, now when the specificity increases, (1-Specifity) also

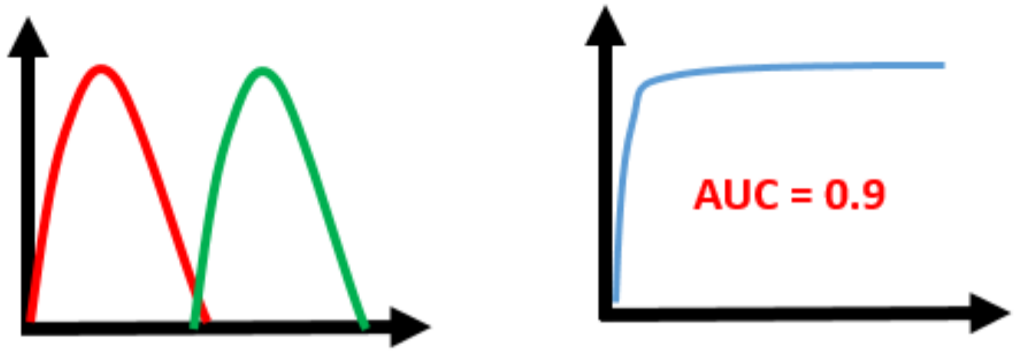

Figure 5.4: Case1: AUC ROC

increases. The plot drawn again sensitivity and (1-Specifity) is known as a ROC curve. The area under the ROC curve is AUC. AUC gives us the good idea of how well the model performs.

In figure 5.4 case 1 illustrates an ideal case of AUC ROC where the model does quite a good job of distinguishing the positive and negative values. Therefore, there the AUC value is 0.9 as of the large area under the ROC curve.

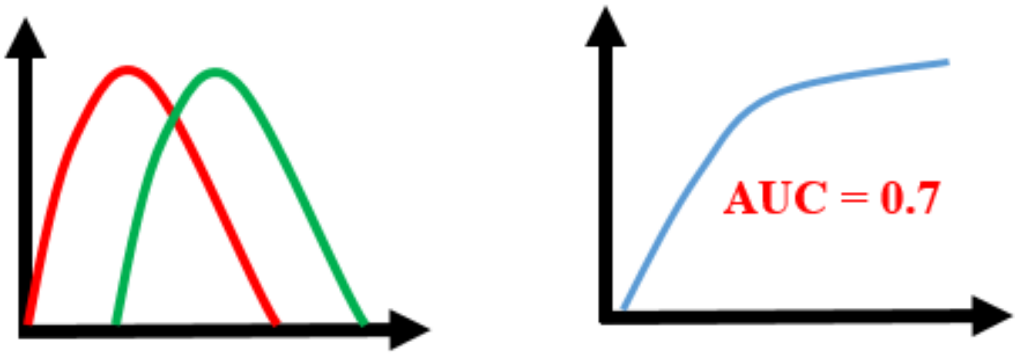

Figure 5.5: Case 2: AUC ROC

In figure 5.5 it can be seen that there are few positive and negative values overlapping each other due to which the AUC has decreased its value to 0.7.

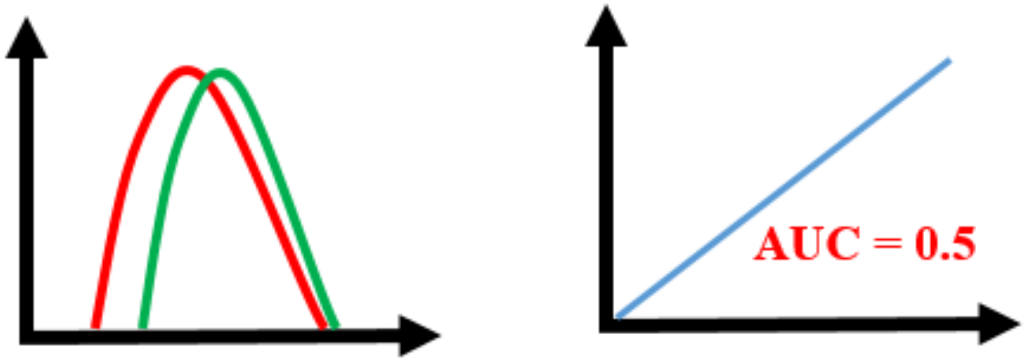

Figure 5.6: Case 3: AUC ROC

In the above figure 5.6 it can be seen that the last model predictions are overlapping each other and we get the AUC value of 0.5. This means that the model is performing very poorly. AUC ROC curve assesses how well the probabilities from the positive classes are isolated from the negative classes.

Error Histogram Plot: Histogram can give an indication of outliers. An outlier is a perception that lies a strange separation from different qualities in an irregular example from a population. It might be said, this definition surrenders it over to the examiner to choose what entity will be considered as anomalous. To differentiate between abnormal and normal entities, first normal entities should be separated out.
It is a good idea to have a look on the outliers to determine either the data is good or bad. If the outliers are the data points that are valid, but aren't like rest of the data, then it can be said that network is extrapolating for these data points. In case of many outliers, you should collect more data that looks like the outlier data points and retrain the network to get the good results.

Performance Plot: To compute the overall performance of a neural network model, a performance plot is created. To plot the performance of a neural network there are different options:

- Mean Squared Error (MSE)

- Cross Entropy Error (CE)

To train neural network you require some proportion of mistake between processed outputs and the coveted target outputs of the training data. You can consider artificial neural network (ANN) as a complex function that acknowledges numeric sources of information and creates numeric outputs. The output values for an ANN are controlled by its internal structure and by the estimations of an arrangement of numeric weights and biases. The primary challenge when working with an ANN is to prepare the system, or, in other words of finding the values for the weights and biases so that, for an arrangement of training data with known inputs and outputs, when given the training inputs, the computed outputs intently coordinate the known training outputs. When utilizing a neural network to perform classification and prediction, it is generally proposed to use cross entropy error as opposed to utilizing mean squared error. Its reason will leave the extent of this thesis work. In classification tasks with neural networks, for example to classify dog breeds based on images of dogs, a very common type of loss function to use is Cross Entropy loss. It is defined as

$$H(p,q) = \mathscr{E}[-log(q)] = H(p) + D(p||q) \tag{5.5.8}$$

In an equation 5.5.8 p is the true distribution and q is the model distribution. H(p) is the entropy and D(p||q) is the KL-divergence.
Minimizing this loss is the same as maximizing the negative loss, it can be seen in equation 5.5.9.

$$min\ \mathscr{E}[-log(q)] = max\ \mathscr{E}[log(q)] \tag{5.5.9}$$

In order to update towards this objective we compute gradients in a neural network w.r.t. theta (These are used to update the weights).

$$\nabla \mathscr{E}[log(q)] \tag{5.5.10}$$

(5.5.10) is the gradient of the cross entropy.
In this expression the expectation is taken only w.r.t. p and the gradient is a function theta. This allows us to (under some mild assumptions) rewrite the gradient in the

following equation 5.5.11.

$$\mathscr{E}[\nabla log(q)] \tag{5.5.11}$$

In practice this expression is approximated by computing the average gradient over a batch of input data to a neural network. In other words, we end up with the score function. So, when we use cross entropy loss while training neural networks, we calculate the score function every time when compute gradients for the weights in the network. To interpret the plots, there are few things to keep in mind to understand the desired output results:

1. As the epochs increases, CE error must decrease.
2. A well-trained ANN should have a very low CE error at the end of the training phase.
3. The meaning of CE error being very small (nearest to zero) is that the desired outputs and ANN's output for the training set have become very close to each other.
4. Calculating CE error is only half of the judgement process of checking the overall performance of ANN.

Training State Plot: Training state plot shows some training statistics.

- Gradient is a value of backpropagation gradient on each iteration on a logarithmic scale. X could be any value of the gradient (computed by the model). The value of gradient (X) means that you reached the bottom of the local minimum of your goal function.
- Validation fails are iterations when validation CE increased its value. A lot of fails means overtraining of a model. The model automatically stops after 6 fails in a row.

Sigmoid Function: The Sigmoid curve resembles a S-shape. The primary motivation behind why we utilize sigmoid function is on account of it exists between a range from

0 to 1. In this manner, it is particularly utilized for models where we need to predict probability as an output. Since probability of anything exists between the scope of 0 and 1, sigmoid is the correct decision. It is possible to differentiate the function.That implies, we can calculate the incline of the sigmoid curve at any two points. Figure 5.7 shows the sigmoid curve.

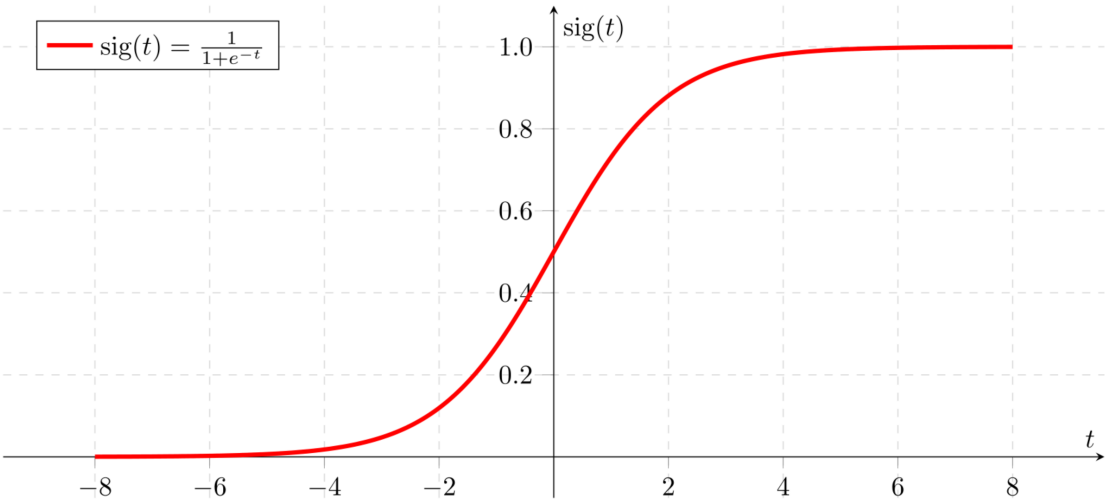

Figure 5.7: Sigmoid Curve [35]

5.6 The Experimentation

To achieve the desired output various experiments were carried out. UNSW-NB15 data set was first divided into following features:

- Basic Features (BFs)
- Content Features (CFs)
- Time Features (TFs)
- Additional Generated Features (AGFs)

AGFs are created using the basic, content and time features. The purpose of creating AGFs was to create such features that particularly will give the accurate results of an intrusions. AGFs are basically the part of payload, which are used by the intruders to

exploit any network.

After the division of features, group of each feature was passed through an experimentation phase. The results were concluded to make a comparison among the features that will help researchers in future.

Deep Learning has been used in this approach to get the best output. Multiple experiments were carried out against the features of UNSW-NB15 data set. Experiments varies with the increase of hidden layers in our network. Starting from the 10 hidden layers, maximum of 100 hidden layers have been created. The desired results were achieved at 100 hidden layer network, So, the experimentation was stopped.

To compile the results, all the features of UNSW-NB15 data set are combined. The results of the additional generated features and all features is then compared. A very minor variation in the output have been seen that made the concept clear that the features generated additionally will be good enough to train ANN for IDS to get the good results. But as a part of this work only the results obtained from all features are discussed.

5.7 Experiment 1

In the first experiment of this project, the artificial neural network was decided to be consisted of 42 input layers, 10 hidden layers and 1 output layer. Figure 5.8 shows the network diagram of this experiment.

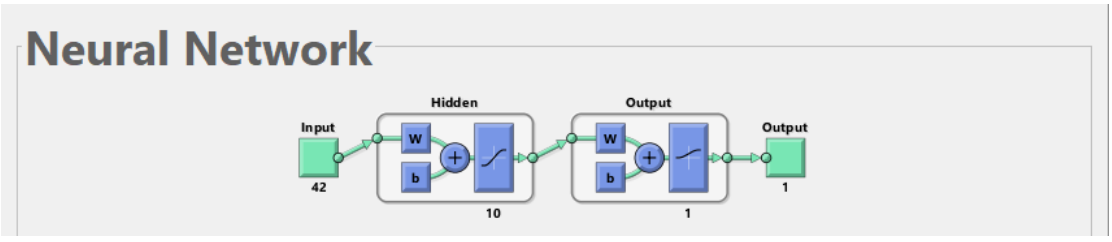

Figure 5.8: Network Diagram

Forty-two layers of input were given to 10 hidden layers, where after adjusting their weights and biases they were passed through a sigmoid function. The output of hidden layers will be treated as an input to the output layer, where again the same process will

be done as of the hidden layers and output taken from that layer will be the final output of the network i.e. 0 or 1.

$$\begin{cases} 1 & \text{if } attack packet \\ 0 & \text{if } normal packet \end{cases}$$

For the training of this model different algorithms are used that are:

- Training: Scaled Conjugate Gradient
- Performance: Cross Entropy

The data was divided in a proportion that is shown in the table 5.3.

Table 5.3: Division of Dataset

Samples	Percentage	Number of Samples
Training	60%	105205 samples
Validation	20%	35068 samples
Testing	20%	35068 samples

Training: Training samples are presented to the network during training, and the network is adjusted according to its error.

Validation: These are used to measure network generalization, and to stop training when generalization stops improving.

Testing: These samples have no effect on training and so provide an independent measure of network performance during and after training.
After all the requirements to create an artificial neural network, it was trained which took approx. 4-5 minutes. Later, after training of a network few plots were created to demonstrate the results.

5.7.1 Confusion Matrix

We have already discussed about this confusion matrix in detail. Fig. 5.9 shows the overall confusion matrix of this experiment.

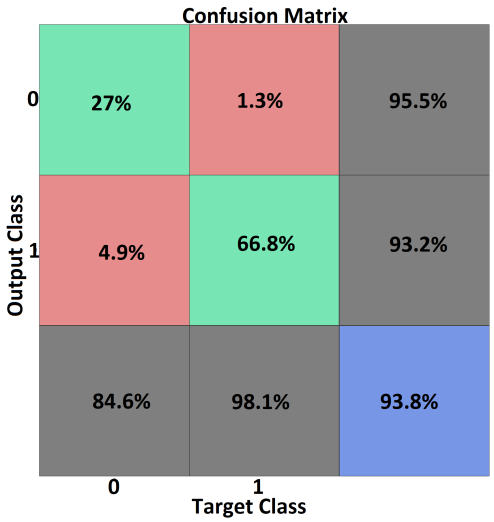

Figure 5.9: All Confusion Matrix Plot

In the above matrix of confusion, there are few target values that have been obtained from this experiment. Those values are represented in the form of success rate versus failure rate in table 5.5.

Table 5.5: Experiment 1: Success versus Failure Rate of Confusion Matrix

Characteristics	Success Rate	Failure Rate
Accuracy	93.8%	6.2%
Sensitivity	84.6%	15.4%
Specificity	98.1%	1.9%
Positive Predictive Value	95.5%	4.5%
Negative Predictive Value	93.2%	6.8%
False Positive Rate	98.7%	1.3%
False Negative Rate	95.1%	4.9%

5.7.2 Receiver Operation Characteristics (ROC) plot

In ANN, ROC is drawn against False Positive Rate (FPR) and True Positive Rate (TPR). Quality of ANN model to predict the values can be analyzed from the area under the curve (AUC) of ROC plot. Greater the area, better the results.

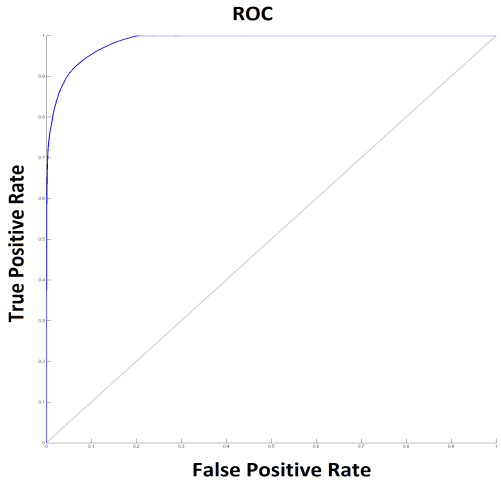

Figure 5.10: Receiver Operation Characteristics Plot

In figure 5.10 it can clearly be seen that the value of false positive rate on the graph is less than 2%. In chapter 2, we studied that in an anomaly-based IDS the major issue is with the FPR value. This experiment has resulted in such a small value of FPR. So, from the result it can be said that this ANN model is good for detecting anomaly-based intrusions.

5.7.3 Error Histogram Plot

The graphical representation of the visual impression of the dispersion of errors is appeared in this plot. It comprises of tabular bars appeared as contiguous rectangles raised over discrete bins. Bins are the number of vertical bars on the graph. The total error from neural network ranges from -0.95 (the leftmost bin) to 0.95 (the rightmost bin). This error range is divided into 20 bins. Each vertical bar represents the number of samples from your data set, which lies in a bin.

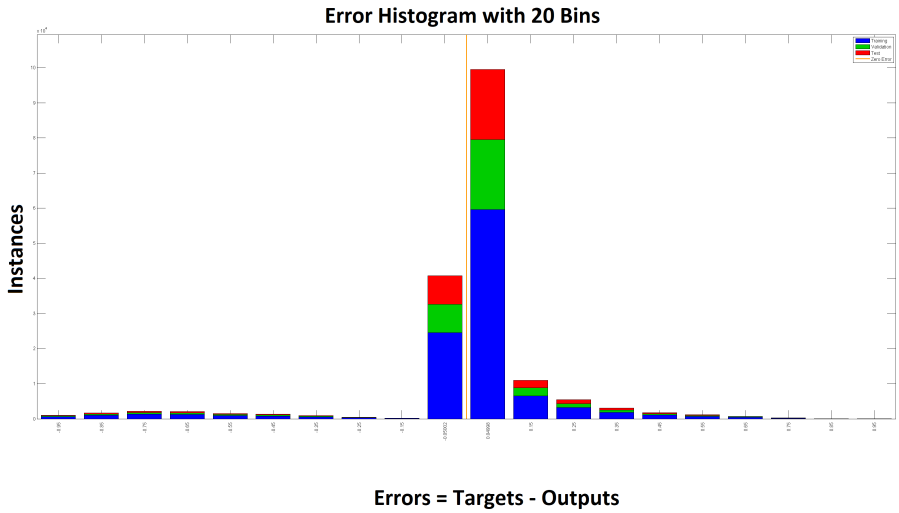

Figure 5.11: Error Histogram Plot

Fig. 5.11 shows the error histogram plot of our experiment. This graph shows that the maximum error is 0.04998. More the error closer to zero is, the more perfectly designed ANN model is.

5.7.4 Training State Plot

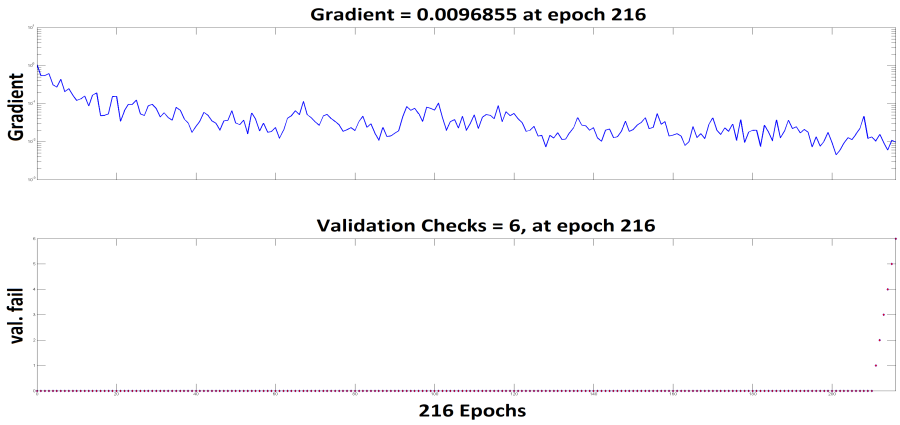

Figure 5.12: Training State Plot

Fig. 5.12 shows the training state plot of artificial neural network. It shows validation check at epoch 216. That means the training of this ANN stopped after 216 iterations.

5.7.5 Performance Plot

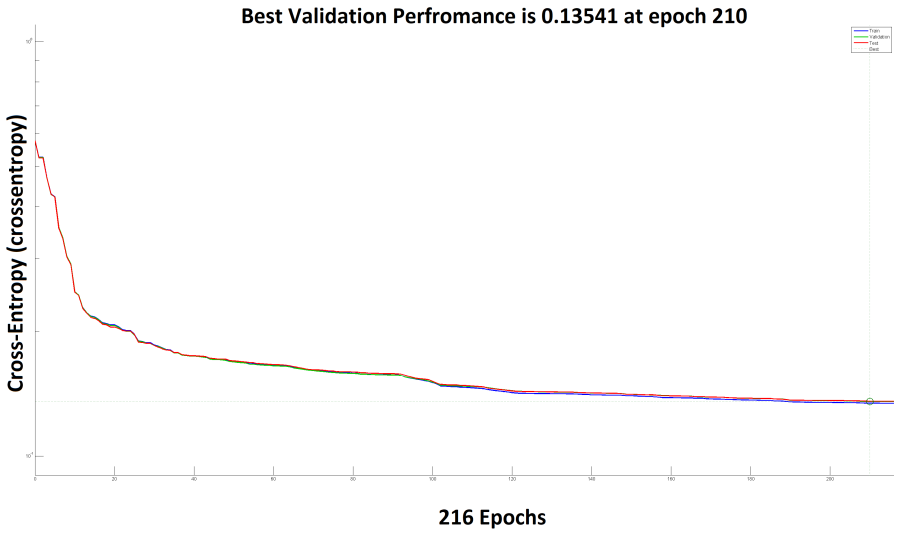

Figure 5.13: Performance Plot

Fig. 5.13 demonstrates the execution of the framework considering training, testing and validation. It demonstrates that the best validation execution is 0.13541 at epoch 210.

Results

	Samples	CE	%E
Training:	105205	9.18868e-1	6.20217e-0
Validation:	35068	1.70295e-0	6.08246e-0
Testing:	35068	1.70739e-0	6.18512e-0

Figure 5.14: Additional Results

Fig. 5.14 shows the additional outcomes of the ANN model utilized in this examination. CE shows the cross entropy. Limiting cross-entropy results in great characterization. The lower estimations of CE are better. Zero estimation of CE implies there is no mistake. %E is the percentage error. It shows the part of tests which are misclassified. Esteem 0 implies no misclassification, esteem 100 shows greatest misclassifications. In figure 5.14 it can clearly be seen that the output results (CE and %E) for training, validation and testing data sets are closer to lower esteem.

5.8 Experiment 2

In the second experiment of this project, the artificial network was decided to be consisted of 42 input layers, 50 hidden layers and 1 output layer. Figure 5.15 shows the network diagram of this experiment.

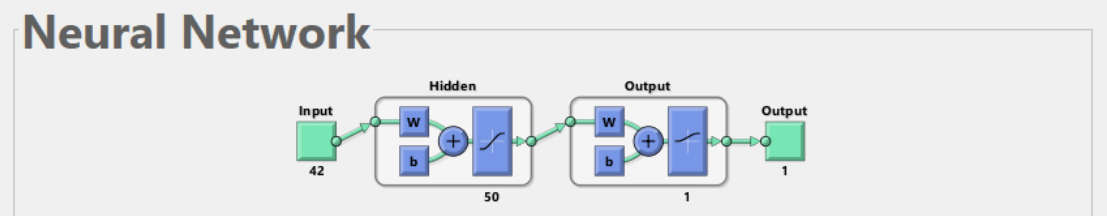

Figure 5.15: Network Diagram

Forty-two layers of input were given to 50 hidden layers, where after adjusting their weights and biases they were passed through a sigmoid function. The output of hidden layers will be treated as an input to the output layer, where again the same process will be done as of the hidden layers and output taken from that layer will be the final output of the network i.e. 0 or 1.

After all the requirements to create an artificial neural network, it was trained which took approx. 4-5 minutes. Later, after training of a network few plots were created to demonstrate the results.

5.8.1 Confusion Matrix

We have already discussed about this confusion matrix in detail. Fig. 5.16 shows the overall confusion matrix of this experiment.

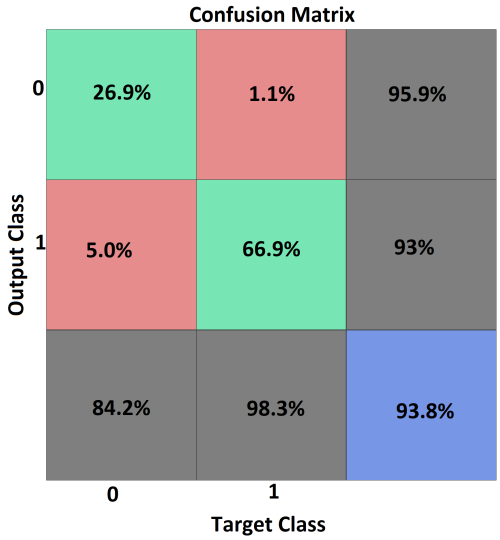

Figure 5.16: All Confusion Matrix Plot

In the above matrix of confusion, there are few target values that have been obtained from this experiment. Those values are represented in the form of success rate versus failure rate in table 5.7.

Table 5.7: Experiment 2: Success versus Failure Rate of Confusion Matrix

Characteristics	Success Rate	Failure Rate
Accuracy	93.8%	6.2%
Sensitivity	84.2%	15.8%
Specificity	98.3%	1.7%
Positive Predictive Value	95.9%	4.1%
Negative Predictive Value	93%	7%
False Positive Rate	98.9%	1.1%
False Negative Rate	95%	5%

5.8.2 Receiver Operation Characteristics (ROC) plot

In ANN, ROC is drawn against False Positive Rate (FPR) and True Positive Rate (TPR). Quality of ANN model to predict the values can be analyzed from the area under the curve (AUC) of ROC plot. Greater the area, better the results.

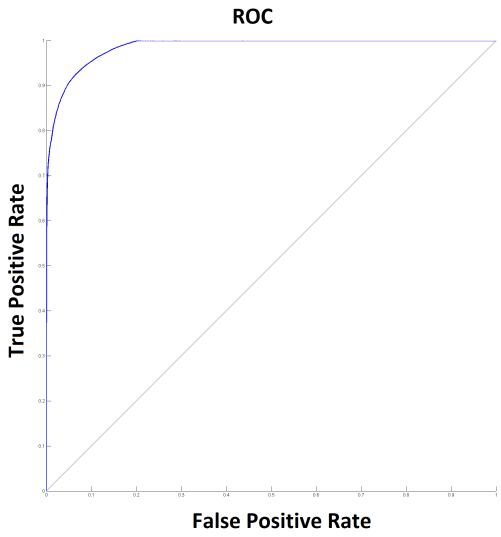

Figure 5.17: Receiver Operation Characteristics Plot

In figure 5.17 it can clearly be seen that the value of false positive rate on the graph is less than 2%. In chapter 2, we studied that in an anomaly-based IDS the major issue is with the FPR value. This experiment has resulted in a small value of FPR. So, from the result it can be said that this ANN model is good for detecting anomaly-based intrusions.

5.8.3 Error Histogram Plot

The graphical representation of the visual impression of the dispersion of errors is appeared in this plot. It comprises of tabular bars appeared as contiguous rectangles raised over discrete bins. Bins are the number of vertical bars on the graph. Each vertical bar represents the number of samples from your data set, which lies in a bin.

Figure 5.18: Error Histogram Plot

Fig. 5.18 shows the error histogram plot of our experiment. This graph shows that the maximum error is 0.04999. More the error closer to zero is, the more perfectly designed ANN model is.

5.8.4 Training State Plot

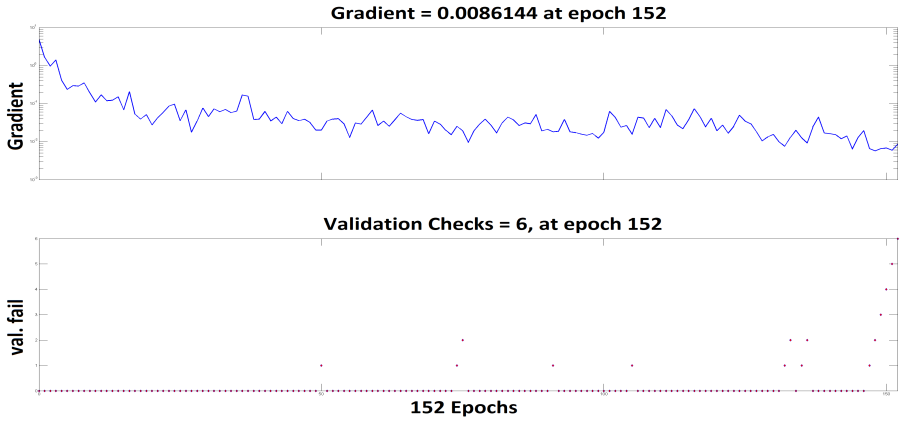

Figure 5.19: Training State Plot

Fig. 5.19 shows the training state plot of artificial neural network. It shows validation check at epoch 152. That means the training of this ANN stopped after 152 iterations.

5.8.5 Performance Plot

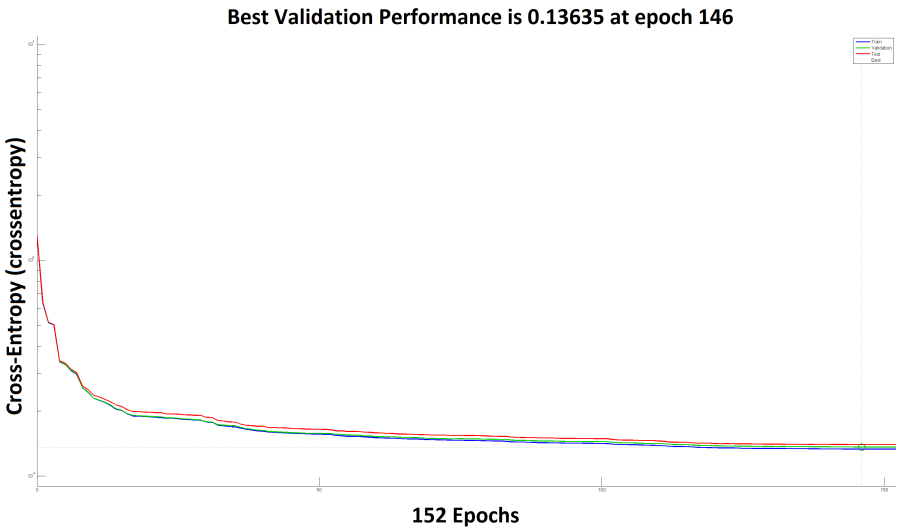

Figure 5.20: Performance Plot

Fig. 5.20 demonstrates the execution of the framework considering training, testing and validation. It demonstrates that the best validation execution is 0.13635 at epoch 146.

Results

	Samples	CE	%E
Training:	105205	1.11002e-0	6.08146e-0
Validation:	35068	2.07183e-0	6.20508e-0
Testing:	35068	2.08436e-0	6.47884e-0

Figure 5.21: Additional Results

Fig. 5.21 shows the additional outcomes of the ANN model utilized in this examination. CE shows the cross entropy. Limiting cross-entropy results in great characterization. The lower estimations of CE are better. Zero estimation of CE implies there is no mistake.

%E is the percentage error. It shows the part of tests which are misclassified. Esteem 0 implies no misclassification, esteem 100 shows greatest misclassifications. In figure 5.21 it can clearly be seen that the output results (CE and %E) for training, validation and testing data sets are closer to lower esteem.

5.9 Experiment 3

In the second experiment of this project, the artificial network was decided to be consisted of 42 input layers, 100 hidden layers and 1 output layer. Figure 5.22 shows the network diagram of this experiment.

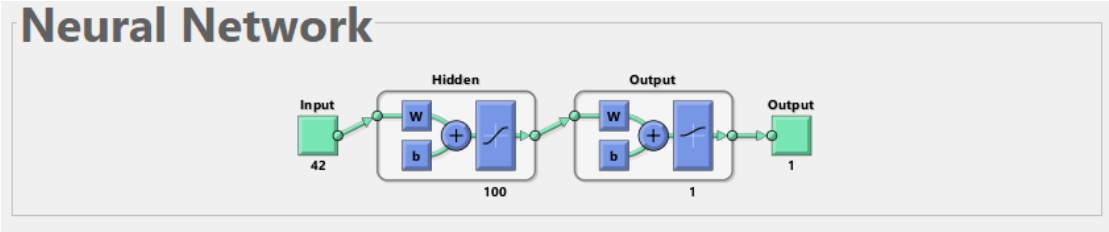

Figure 5.22: Network Diagram

Forty-two layers of input were given to 100 hidden layers, where after adjusting their weights and biases they were passed through a sigmoid function. The output of hidden layers will be treated as an input to the output layer, where again the same process will be done as of the hidden layers and output taken from that layer will be the final output of the network i.e. 0 or 1.

After all the requirements to create an artificial neural network, it was trained which took approx. 4-5 minutes. Later, after training of a network few plots were created to demonstrate the results.

5.9.1 Confusion Matrix

We have already discussed about this confusion matrix in detail. Fig. 5.23 shows the overall confusion matrix of this experiment.

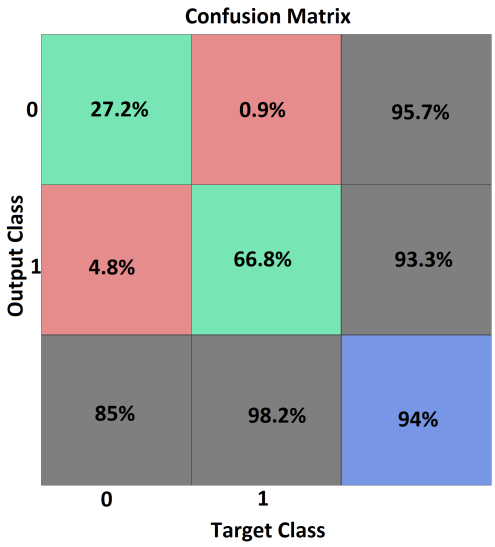

Figure 5.23: All Confusion Matrix Plot

In the above matrix of confusion, there are few target values that have been obtained from this experiment. Those values are represented in the form of success rate versus failure rate in table 5.9.

Table 5.9: Experiment 3: Success versus Failure Rate of Confusion Matrix

Characteristics	Success Rate	Failure Rate
Accuracy	94%	6%
Sensitivity	85%	5%
Specificity	98.2%	1.8%
Positive Predictive Value	95.7%	4.3%
Negative Predictive Value	93.3%	6.7%
False Positive Rate	99.1%	0.9%
False Negative Rate	95.2%	4.8%

5.9.2 Receiver Operation Characteristics (ROC) plot

In ANN, ROC is drawn against False Positive Rate (FPR) and True Positive Rate (TPR). Quality of ANN model to predict the values can be analyzed from the area under the curve (AUC) of ROC plot. Greater the area, better the results.

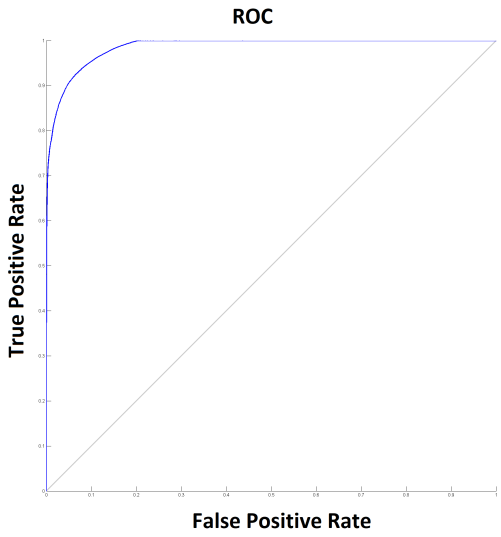

Figure 5.24: Receiver Operation Characteristics Plot

In figure 5.24 it can clearly be seen that the value of false positive rate on the graph is less than 2%. In chapter 2, we studied that in an anomaly-based IDS the major issue is with the FPR value. The experiment has resulted in a very small value of FPR. So, from the result it can be said that this ANN model is perfect for detecting anomaly-based intrusions.

5.9.3 Error Histogram Plot

The graphical representation of the visual impression of the dispersion of errors is appeared in this plot. It comprises of tabular bars appeared as contiguous rectangles raised over discrete bins. Bins are the number of vertical bars on the graph. Each vertical bar represents the number of samples from your data set, which lies in a bin.

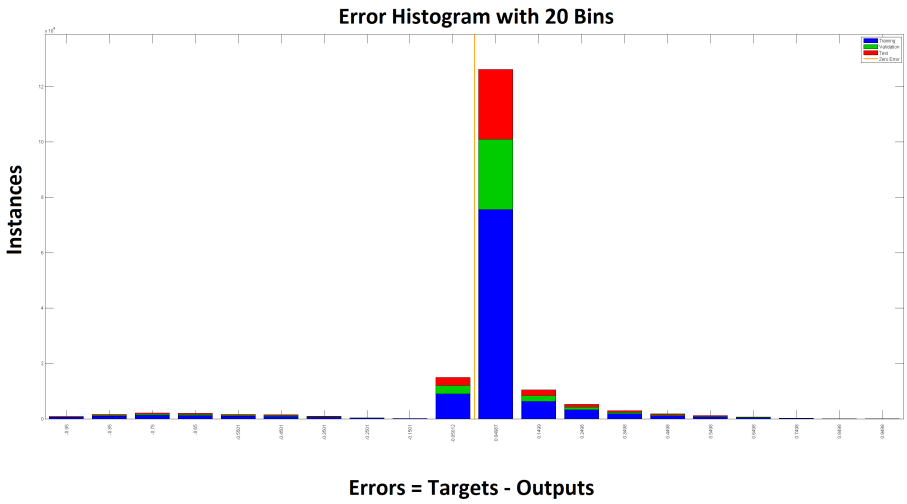

Figure 5.25: Error Histogram Plot

Fig. 5.25 shows the error histogram plot of our experiment. This graph shows that the maximum error is 0.04987. More the error closer to zero is, the more perfectly designed ANN model is.

5.9.4 Training State Plot

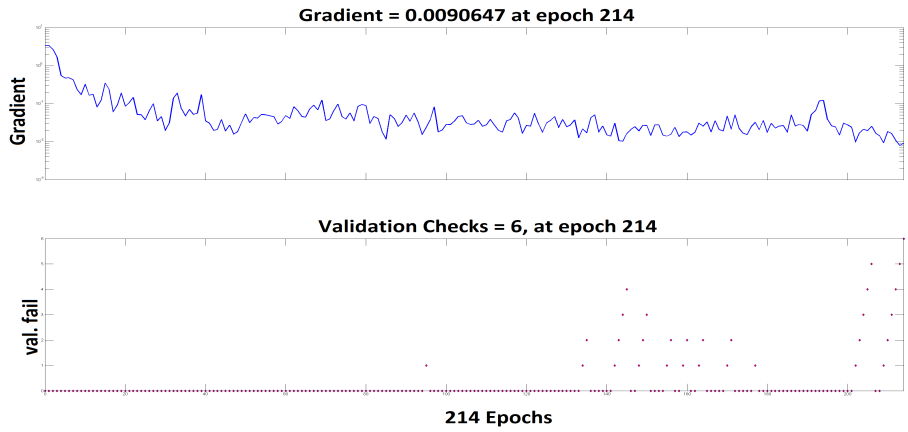

Figure 5.26: Training State Plot

Fig. 5.26 shows the training state plot of artificial neural network. It shows validation check at epoch 214. That means the training of this ANN stopped after 214 iterations.

5.9.5 Performance Plot

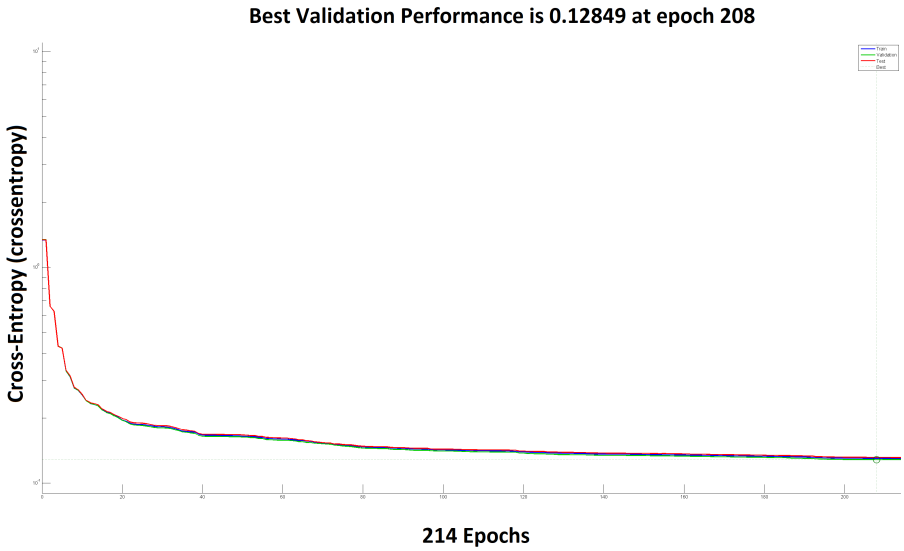

Figure 5.27: Performance Plot

Fig. 5.27 demonstrates the execution of the framework considering training, testing and validation. It demonstrates that the best validation execution is 0.12849 at epoch 208.

Results

	Samples	CE	%E
Training:	105205	9.93658e-1	6.20787e-0
Validation:	35068	1.83594e-0	6.07961e-0
Testing:	35068	1.84366e-0	6.32200e-0

Figure 5.28: Additional Results

Fig. 5.28 shows the additional outcomes of the ANN model utilized in this examination. CE shows the cross entropy. Limiting cross-entropy results in great characterization. The lower estimations of CE are better. Zero estimation of CE implies there is no mistake.

%E is the percentage error. It shows the part of tests which are misclassified. Esteem 0 implies no misclassification, esteem 100 shows greatest misclassifications. In figure 5.28 it can clearly be seen that the output results (CE and %E) for training, validation and testing data sets are closer to lower esteem.

5.10 Discussion of Results

From the above plots that have been obtained from experimentation phase, it can be concluded that more we create a deep neural network the more it will produce better results in the form of accuracy, errors, false alarms, performance etc. In dealing with anomaly based IDSs, a major challenge is to avoid false alarm rates. In our last experiment, we have achieved 99.1% of success rate in avoiding false alarm rates. Though the experimentation was stopped at reaching a certain level of approximation but knowing from a fact that was explained in the definition of machine learning in chapter 2 that a machine can learn from its experience. In deep learning, machines are tend to improve their efficiency of producing better outputs after having experience of the multiple events. So, keeping this point in mind we can get 100% results of our proposed ANN model after a certain level of experience of it.

5.11 Summary

In this chapter, the proposed solution and proposed framework of the problem statement is discussed. After that UNSW-NB15, the dataset used in the experiments of this thesis is discussed briefly. Lastly, the terminologies that will be used this chapters to understand the results and the work of this thesis are covered. Moreover, the experiments and results are discussed in detail followed by the introduction of new terms like cross entropy and percentage error that gives the description of the designed algorithm that how well it can perform. The results are shown using different plots of confusion matrix, ROC, performance, training state and error histogram.

Chapter 6

Discussion, Conclusion and Future Work

6.1 Introduction

In this chapter, we will discuss the results that were produced from the experiments. Later, answer of research questions that were mentioned in the first chapter will be covered. At the end, future work that could be carried out from this research work will be suggested.

6.2 Discussion

Cloud Computing is an emerging technology that has attracted the focus of many computer scientists and engineers in its domain of research for the development of computing systems based on cloud computing framework. Many organizations have adapted the framework of cloud computing for their operations. But other than its demand, many organizations are concerned about the security in cloud computing. In this thesis, work has been done to detect and prevent the intrusions in federated cloud computing paradigm. Detection is based on two methods; either it is done using the signatures of known attacks or by monitoring the deviations in the behaving patterns of user. This work has focused on the second approach, that is anomaly-based detection of an IDS.

In anomaly-based detection, many researchers have been working on the signature-based IDS. Very few researchers have targeted this domain of detecting intrusions using anomalies because of the inherited challenges. In this work, many challenges like reducing the errors, reducing false alarm rates and the overall better performance of a system have been focused. Multiple experiments have been carried out to achieve the expected results and after achieving those expected results a framework has been proposed that will be perfect to be deployed in an environment of cloud federation.

6.3 Conclusion

This study has focused on an environment of cloud federation, detection and prevention of intrusions in cloud computing. This thesis was started with few research questions. So, this conclusion will be done with answering those research questions. The answers of those questions are numbered in the same sequence as the questions were numbered in chapter 1.

1. After an exhaustive research and studying multiple papers, security challenges and concerns has been discussed in detail to give reader a deep insight of all the challenges that could be faced by a user of cloud computing model.

2. Yes, integrating concepts of cloud computing with artificial intelligence has proven to be very effective. This work has adapted artificial neural network technique from the domain of artificial intelligence and from the results it can be concluded that AI techniques are way better than traditional network security techniques to be used in cloud computing frameworks.

3. Detailed analysis has been done to select the best approach to be used to get the desired results. After studying papers and doing the literature review, machine learning techniques were finalized to be used in our work. Among all machine learning techniques, a detailed study of artificial neural network was done. Later, ANN was chosen to be used in our intrusion detection system.

6.4 Future Work

To address the solution of our proposed problem, there are many other machine learning techniques like genetic algorithm (GA), fuzzy logic techniques, and artificial immune systems (AIS) that could be very helpful in detecting and preventing the intrusions done by hackers in federated cloud computing environment. Machine learning is a very large domain that could be highly beneficent if used in the domain of cybersecurity.

Knowing that, a complete security of any computing system can never be guaranteed. There are number of examples of data breaches done to those organizations [42] which were following the best practices of securing their data but because of few vulnerabilities found in their systems, their data got compromised. There are always the possibilities of data leakage, the probability might be very small but still it can't be denied. In that case along with the deployment of intrusion detection and prevention systems a data destruction cryptographic module could be advantageous if deployed in FCC paradigm. That module will keep on sensing data, and in case of data breach it will simply destroy the data available in cloud.

However, most vital future work expresses that there are no solid norms for cloud computing. Some open cloud manifesto benchmarks and couple of endeavors made by the Cloud Security Alliance to institutionalize the procedures in cloud. Cloud providers and clients aren't willing to utilize those measures since they are bit prohibitive. This inability to give solid security guidelines, normal fundamental systems for migration of data and general standards for cloud interoperability, making the leading innovative technology "Cloud computing" a defenseless choice for its clients. To stay away from this, ought to be a legitimate law and requirement system to be adjusted with the goal that the cloud structures could be made more secure for consumers.

6.5 Summary

This chapter covers the discussion and conclusion that answers the research questions of this thesis. Later, future work of this research is covered. Future work includes few suggestions for future researchers who are willing to contribute in this domain.

List of Abbreviations and Symbols

Abbreviations

CC	Cloud Computing
FCC	Federated Cloud Computing
CSC	Cloud Service Consumer
CSP	Cloud Service Provider
MU	Malicious User
TU	Trusted User
DDOS	Distributed Denial of Service
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
IT	Information Technology
IS	Information Security
API	Application Programming Interface
NIST	National Institute of Standards and Technology
SaaS	Software as a Service
PaaS	Platform as a Service

IaaS	Infrastructure as a Service
HTTP	Hyper Text Transfer Protocol
HTTPS	Hyper Text Transfer Protocol Secure
MAC	Mandatory Access Control
DAC	Discretionary Access Control
RBAC	Role Based Access Control
NIDS	Network based Intrusion Detection System
HIDS	Host based Intrusion Detection System
NIPS	Network based Intrusion Prevention System
HIPS	Host based Intrusion Prevention System
WIPS	Wireless based Intrusion Prevention System
OSI	Open Systems Interconnection model
CMS	Central Management Server
LAN	Local Area Network
UTM	Unified Threat Management
AI	Artificial Intelligence
ML	Machine Learning
ANN	Artificial Neural Network
MLP	Multi Layer Perceptron
ABC	Artificial Bee Colony
AIS	Artificial Immune System
CE	Cross Entropy

ROC	Receiver Operation Characteristics
MSE	Mean Squared Error
TPR	True Positive Rate
TNR	True Negative Rate
FPR	False Positive Rate
FNR	False Negative Rate
AUC	Area Under Curve

References

[1] Foster, Ian, et al. "Cloud computing and grid computing 360-degree compared." Grid Computing Environments Workshop, 2008. GCE'08. Ieee, 2008.

[2] Ghazizadeh, Eghbal, Mazdak Zamani, and Abolghasem Pashang. "A survey on security issues of federated identity in the cloud computing." Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on. IEEE, 2012.

[3] Dillon, Tharam, Chen Wu, and Elizabeth Chang. "Cloud computing: issues and challenges." Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on. Ieee, 2010.

[4] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).

[5] Weiss, Aaron. "Computing in the clouds." networker 11.4 (2007): 16-25.

[6] Liu, Fang, et al. "NIST cloud computing reference architecture." NIST special publication 500.2011 (2011): 1-28.

[7] Michael, Armbrust, et al. "Above the clouds: A Berkeley view of cloud computing." EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28 (2009).

[8] Lo, Chi-Chun, Chun-Chieh Huang, and Joy Ku. "A cooperative intrusion detection system framework for cloud computing networks." Parallel processing workshops (ICPPW), 2010 39th international conference on. IEEE, 2010.

[9] MacDermott, Áine, Qi Shi, and Kashif Kifayat. "Collaborative intrusion detection in federated cloud environments." J. Comput. Sci. Appl. Big Data Anal. Intell. Syst 3 (2015): 10-20.

[10] A. Jahan, M. A. Alam, “Intrusion Detection Systems based on Artificial Intelligence,” IJARCS, 2017, pp. 705-708.

[11] IEEE Confluence Report, “Artificial Intelligence and Machine Learning applied to Cyber Security”, 2017. [Online], [Retrieved September 2, 2018], https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/industry/ieee_confluence_report.pdf

[12] Alrajeh, Nabil Ali, and Jaime Lloret. "Intrusion detection systems based on artificial intelligence techniques in wireless sensor networks." International Journal of Distributed Sensor Networks 9.10 (2013): 351047.

[13] Calheiros, Rodrigo N., et al. "A coordinator for scaling elastic applications across multiple clouds." Future Generation Computer Systems 28.8 (2012): 1350-1362.

[14] Patel, Ahmed, et al. "An intrusion detection and prevention system in cloud computing: A systematic review." Journal of network and computer applications 36.1 (2013): 25-41.

[15] Xing, Tianyi, et al. "Snortflow: A openflow-based intrusion prevention system in cloud environment." Research and Educational Experiment Workshop (GREE), 2013 Second GENI. IEEE, 2013.

[16] Scarfone, Karen, and Peter Mell. "Guide to intrusion detection and prevention systems (idps)." NIST special publication800.2007 (2007): 94.

[17] Chen, Zhen, et al. "Cloud computing-based forensic analysis for collaborative network security management system." Tsinghua science and technology 18.1 (2013): 40-50.

[18] Montes, Jesús, et al. "GMonE: A complete approach to cloud monitoring." Future Generation Computer Systems 29.8 (2013): 2026-2040.

[19] Von Neumann, John. "The general and logical theory of automata." Cerebral mechanisms in behavior 1.41 (1951): 1-2.

[20] Kumar, Gulshan, Krishan Kumar, and Monika Sachdeva. "The use of artificial intelligence based techniques for intrusion detection: a review." Artificial Intelligence Review 34.4 (2010): 369-387.

[21] Naik, Nitin. "Fuzzy inference based intrusion detection system: FI-Snort." Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on. IEEE, 2015.

[22] Hajimirzaei, Bahram, and Nima Jafari Navimipour. "Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm." ICT Express (2018).

[23] N. Horn. (2014) Snort (2.9.5.6). [Online]. Available http://slackbuilds.org/repository/14.1/network/snort/.

[24] M. Rouse (2014) Snort. [Online]. Available http://searchmidmarketsecurity.techtarget.com/definition/Snort.

[25] Naik, Nitin, Pan Su, and Qiang Shen. "Integration of interpolation and inference." Computational Intelligence (UKCI), 2012 12th UK Workshop on. IEEE, 2012.

[26] Abdullah, B., et al. "Performance evaluation of a genetic algorithm based approach to network intrusion detection system." Proceedings of the International Conference on Aerospace Sciences and Aviation Technology. 2009.

[27] B. S. Dhak and S. Lade, “An evolutionary approach to intrusion detection systems using genetic algorithm,” International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 2, pp. 632-637, 2012.

[28] Hosseinpour, Farhoud, et al. "Design of a new distributed model for Intrusion Detection System based on Artificial Immune System." Advanced Information Management and Service (IMS), 2010 6th International Conference on. IEEE, 2010.

[29] Fanelli, Robert L. "A hybrid model for immune inspired network intrusion detection." International Conference on Artificial Immune Systems. Springer, Berlin, Heidelberg, 2008.

[30] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).

[31] William “Bill” Meine, in private communication; 2010.

[32] Brunette G, Mogull R. Security Guidance for Critical Areas of Focus in Cloud Computing V2.1; Prepared by the Cloud Security Alliance; 2009.[Online], [Retrieved September 18, 2018], http://www.cloudsecurityalliance.org/guidance/csaguide.v2.1.pdf

[33] Ouedraogo, Moussa, et al. "Security transparency: the next frontier for security research in the cloud." Journal of Cloud Computing 4.1 (2015): 12.

[34] Gruschka, Nils, and Meiko Jensen. "Attack surfaces: A taxonomy for attacks on cloud services." 2010 IEEE 3rd international conference on cloud computing. IEEE, 2010.

[35] https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6. Date Accessed: 24/10/18

[36] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[37] KDDCup1999. Available: http://kdd.ics.uci.edu/databases/kddcup99/KDDCUP99.html, 2007.

[38] McCulloch, Warren S., and Walter Pitts. "The statistical organization of nervous activity." Biometrics 4.2 (1943): 91-99.

[39] Winkler, Vic. "Securing the cloud." Cloud computer security techniques and tactics. Syngress (2011).

[40] Jin, Xin, Ram Krishnan, and Ravi Sandhu. "A unified attribute-based access control model covering DAC, MAC and RBAC." IFIP Annual Conference on Data and Applications Security and Privacy. Springer, Berlin, Heidelberg, 2012.

[41] Maler, Eve, and Drummond Reed. "The venn of identity: Options and issues in federated identity management." IEEE Security & Privacy 2 (2008): 16-23.

[42] https://blog.barkly.com/biggest-data-breaches-2018-so-far. Data Accessed: 5/11/18