Improving Detection Rates Using Misuse Detection and Machine Learning

Thesis submitted in partial fulfilment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Applications

Submitted By
Ramandeep Kaur
(Roll No. 651203006)

Under the supervision of:
Dr. Sanmeet Kaur
Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Improving Detection Rates Using Misuse Detection and Machine Learning*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Applications* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Sanmeet Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Ramandeep Kaur)

651203006

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Sanmeet Kaur)

Assistant Professor, CSED

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University

Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor and coordinator **Dr. Sanmeet Kaur**. I thank my supervisor for her time, patience, discussions and valuable comments. Her enthusiasm and optimism made this experience both rewarding and enjoyable.

I am also thankful to **Dr. Deepak Garg,** Head, Computer Science and Engineering Department.

I will be failing in my duty if I don't express my gratitude to **Dr. S. S. Bhatia**, Professor and Dean of Academic Affairs of University, for making provisions of infrastructure such as library facilities, immensely useful for the learners to equip themselves with the latest field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my parents for their wonderful love and encouragement, without their blessings none of this would have been possible. I would also like to thank my close friends for their constant support.

ABSTRACT

Network Security is becoming a crucial issue for all the firms and companies and with the increase in knowledge of intruders and hackers they have made many prosperous attempts to bring down web services and high-profile company networks. Internet has changed and significantly enhanced the way we do business, this massive network have opened the ways to an growing number of security attacks from which corporations must protect them.

Network security is the provision made in an underlying computer network or rules made by the administrator to protect the network and its resources from unauthorized access. With the recent advances in the field of network security a technique called Intrusion Detection System are develop to further enhance and make your network secure. It is a way by which we can protect our internal network from outside attack, and can take appropriate action if needed.

The thesis starts with the introductive study of various kinds of attacks in the network and then different tools to protect network from various malicious activities are studied. On the broader level, there are two techniques that are for detecting Intrusions viz. misuse detection and anomaly detection. Misuse detection detects intrusions by matching the network traffic with database of stored signatures and anomaly detection looks for behaviour deviating from normal or common behaviour for detecting intrusions.

The primary objective of the thesis work is to combine both these techniques. The KDD dataset is used for this purpose. Finally the data is processed on classification algorithms to obtain the results. The results show high percentage of correct classification and accuracy. Experimental evaluation shows that the combined approach of Machine learning and misuse detection gives better performance.

LIST OF FIGURES

LIST OF SNAPSHOTS

LIST OF TABLES

LIST OF ALGORITHMS

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Introduction to Network Security

Today the use of internet has increased greatly. Due to Internet, computer networks are becoming progressively susceptible to various kinds of attacks. Organizations usually wish to preserve the confidentiality and integrity of their useful data which is very vital to an Organization. It is necessary to protect useful data and resources from attacks. The attacks on the Internet have become both more bountiful and easier to implement because of the ubiquity of the Internet. Network security is very problematic since computers have been linked and networked together. Due to the development of the internet, there has been an increasing demand for network security systems. A successful network security plan is developed with the understanding of security issues; needed level of security, potential attackers and factors which make a network vulnerable to attack. Network security components consists of

- Anti-virus and anti-spyware.
- Firewall: used to block unauthorized access to the network system.
- Intrusion Detection System (IDS): used to identify fast-spreading threats.
- Virtual Private Networks (VPN): used to provide secure remote access.

The security of network includes the authorization which allow user to access data in a network, which is being controlled by administrator of network. Network security covers various computer networks, public and private which are used in everyday jobs including transactions and communications among government agencies, individuals and businesses. Networks may be private, such as within an organization and others which may be open to public access or use. Network security secures not only the network, as well as protecting data and overseeing operations being done. Detection of different attacks in the network traffic is one of the major intent of security. The three major goals of network security are:

- **Confidentiality:** Sensitive data needs to be hidden from unauthorized access.

- **Integrity:** Sensitive data needs to be protected from unauthorized change.

- **Availability:** Sensitive data should be available to an authorized entity when it is required.

1.2 Need For Network Security

Network security is much significant as technology is growing faster, the computer systems are becoming more user friendly and very easy to use. Due to advancement in technology, the ways to attacking the system has also increased to a large extent. Today, it becomes very easy for attacker to attack the system. The network security is required against hackers and attackers. A network system should be protected and secure from viruses, Trojan Horses and worms which leak out all the sensitive data in the system [1]. Some network security services are described below:

- **Authentication**

It is the assurance process which identifies that the communicating party is the one that it claims to be. In authentication process the identity of user is confirmed first. User authentication is very important to ensure proper authorization and access to system services and resources. Authenticity is kind of assurance that participants in communication are genuine. There are various authentication measures such as single sign on systems, user login, digital certificates and biometrics [2].In case of a single message, the main purpose of the authentication service is to ensure the recipient that the message is from the same source that it claims to be from. There are two specific authentication services which are defined in X.800.

a) **Peer entity authentication:** It provides the evidence for the identity of an entity in communication and in an association. This service aims to provide confidence that an entity is not performing either an unauthorized replay of a previous connection or a masquerade.
b) **Data origin authentication:** It provides the evidence for the source of a data unit. It does not provide protection against the modification or duplication of data units. This service supports applications such as electronic mail where there are no prior interactions between the communicating entities.

- **Confidentiality**

Confidentiality means to limit the access to information. It means those who have been authorized can only access the information. It ensures that information is hidden from unauthorized user. Confidentiality is a mechanism which provides protection to transmitted data from attacks. On the other hand, confidentiality is the protection of traffic from analysis. Confidentiality refers to a process which ensures that information or data send by authenticated user does not get leaked or disclosed to any unauthorized user. For example, on the internet credit card transaction needs credit card number to be transmitted from buyer to merchant and from merchant to a transaction executing network. Confidentiality is achieved by encrypting the card number during transaction and limits the places where it may appear i.e. log files, printed receipts, backups and databases etc.

- **Data integrity**

Data integrity is the process of giving assurance that information can only be accessed and modified by a user who have authority to do so. There are some measures which are taken to ensure integrity include servers that restrict access to information, maintaining precise authentication policies and controlling the physical environment of networked terminals and ends. Integrity provides us guarantee that a message which is being transferred never gets corrupted. It guarantees that messages are received in same form or format as sent without duplication, modification, insertion or replays.

- **Non-repudiation**

Non repudiation is a process which ensures that the sender and the receiver of message cannot deny that they have ever sent or received such a message. It prevents user from denying a transmitted message. When a message is sent, the receiver of message can prove that the alleged sender has sent the message. On the other hand, when message is being received, the sender can prove that the alleged receiver has received the message.

1.3 Security Attacks in Network

Security attacks are classified into various categories. These are described as follows:

- **Passive Attack**

Passive attacks are also termed as network attacks in which attacker only eavesdrops. Intruder does not alter or modify the message. In this attack, system is scanned and monitored only. The purpose of attacker is to gain information about target system. There are two types of passive attacks, traffic analysis and release of message content [3].

- **Active Attack**

In active attacks, intruder disrupts communication over the internet. An active attack is also called network exploit in which intruder tries to make changes to data on the network [3]. These attacks involve alteration or modification of data. These are divided into following categories:

i. **Replay**: Replay attack is also called playback attack in which transmission of data is maliciously delayed or repeated.
ii. **Masquerade*:*** In this attack, hacker gains unauthorized access to data by using fake identity.
iii. **Denial of service**: In this attack, an attacker tries to halt the service of the system.

- **Insider Attack**

Insider attack is also termed as insider threat. This attack involves a person such as employee from inside the company or organization who has authorized access to system. These attacks are malicious as attacker can affect all components of network security.

- **Close-in Attack**

This type of attack involves a person who is physically close to the system. In this way, attacker can easily exploit the security of the system. It becomes very easy for attacker to gain the access to information by making social interaction with another person. This is called social engineering attack.

- **Phishing Attack**

Phishing attack involves a fake website or email that seems to be original. An attacker creates this web site or email to gain the personal and financial information like usernames, passwords and credit card etc. An attacker gains the information when user attempts to log on with their account.

- **Hijack Attack**

In hijack attack, an attacker intercepts or disconnects the communication and takes the control over the network. Sender still believes that he is sending messages to the intended receiver.

- **Spoof Attack**

In this type of attack, the source address of the packet is changed or modified by an attacker so that packets seems to be receiving from someone else. There are various types of spoofing attack such as ARP spoofing attacks, IP address spoofing attacks and DNS server spoofing attacks.

- **Buffer Overflow Attack**

In buffer overflow attack, the attacker attempts to send more data than expected. Buffer overflow occurs when user stores more data than it was intended to hold. This results into incorrect results, memory crash and breach of information security.

- **Exploit Attack**

In exploit attack, the attacker attempts to make use of data, commands and software to exploit the weakness in a computer system. There attacks are described as follows:

i. **Viruses:** These are small programs that are loaded onto computer without knowledge of user. These spread from one computer to another and replicates themselves.
ii. **Worms:** These are used to send copies of original code from one computer to another computer and these are also self-replicating programs.
iii. **Trojan Horses:** These are designed to provide unauthorized access to user's computer. These do not replicate themselves.

- **Password Attack**

In password attack, the attacker tries to hack or crack the password that is stored in network's database. Three types of password attack are brute force attack, hybrid attack and dictionary attack.

1.4 Mechanisms in Network Security

X.800 recommended following security mechanisms to provide network security:

- **Encipherment**

Encipherment refers to covering or hiding the data. Steganography and cryptography are two techniques which are used for encipherment. Encipherment provides data confidentiality. Encipherment transforms the data into unreadable form. The purpose of this mechanism is to provide privacy by hiding information from unauthorized entity.

- **Data integrity**

Checkvalue is used to check the preservation of data integrity. The sender sends checkvalue with data to receiver. On the other hand, the receiver also creates checkvalue and compares with the checkvalue sent by sender to ensure the message's integrity.

- **Traffic Padding**

Traffic padding involves the addition of bits into gap of data to protect it from attackers or intruder.

- **Routing Control**

Routing control is a mechanism which aims to improve the connectivity of network and changes the path of communication between sender and receiver to protect the data from attackers.

- **Notarization**

To prevent repudiation, notarization is used. It involves third party to control communication between sender and receiver.

- **Access Control**

To allow user to access data, PINs and passwords are used as access control methods.

- **Digital Signature**

Digital signature is used to validate integrity of the message, authenticity, software and digital document [4].

1.5 Available Security Tools

There are various security mechanisms available to prevent any penetration to the network. Some of these tools are described below:

1.5.1 Firewalls

Network firewalls are systems or devices which control the flow of information between networks. Firewall policy is used to control network traffic. Firewall polices are defined by network administrator. There are different rules for different type of network traffic. Firewalls are used to protect the network from outside attack. Network traffic arrived at firewalls must be checked against the defined rules. Firewalls allow some packets to pass and block the others. Firewalls act as a bridge between an internal network and external network which is not secure. So firewalls are used to protect information from threats. Firewalls offer data integrity, confidentiality and data availability. The main function of firewalls is to protect network from unauthorized access.

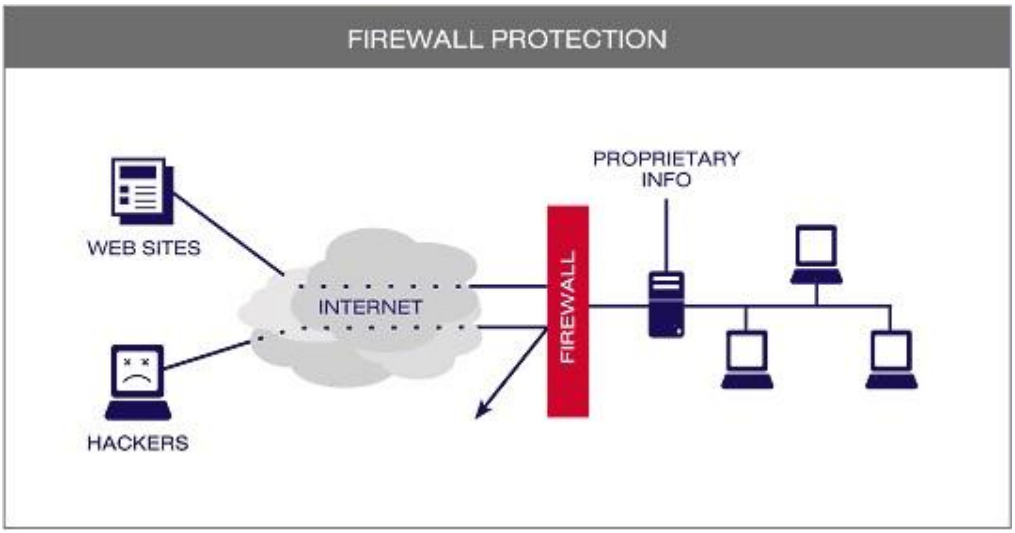


Figure 1.1 Firewall Protection

Following are the types of firewalls:

- **Packet filters:** Packet filtering is a process that controls the outgoing and incoming of packets from a network by filtering or examining the headers of packets. The headers of packets include destination address, source address, type of network traffic like UDP, ICMP, TCP, IP protocol interface. The working of packet filters is based on matching the rules in TCP or IP header. If there is a match then it forwards the packet otherwise packet will be discarded.

- **Application layer:** The working of application layer firewall is based on application level of IP/TCP stack. These are divided into two types: active or passive. Active application layer firewall works by examining all the incoming traffic whereas passive application layer only inspect but do not actively discard the message or packet.

- **Proxies:** A proxy firewall is also known as gateway firewall. It provides security to network resources by filtering the packets at application level. Proxying inspects source port, source address, destination port, destination address and content of packets.

- **Stateful inspection:** Stateful inspection firewall is used to keep the record of network connections. It maintains UDP "pseudo" sessions and TCP sessions. It is more secure than packet filter firewall.

1.5.2 Honeypot

Honeypots are virtual machines that are designed to collect information about an intruder or attacker into our system. Honeypot systems are like computer systems which contain directors, files similar to real computers. The purpose of honeypot system is to fascinate the attackers or hackers to record their activities [5]. The activities of attacker provide useful information. Honeypots act like a fake system. Honeypots can be subdivided into two kinds:

a) **Production honeypots:** Production honeypots are used to protect an organization or company from intruders or attacks. These honeypots are used to enhance or improve the security of an organization or company. These are very easy to implement and deploy. Production honeypot gives less information about an attacker.

b) **Research honeypot:** These are very complex honeypots. These are designed to gather information about hackers and their activities. These are not valuable to company or an organization. These are very useful for military, government and research organizations. Research honeypots are capable of capturing large amount of data. The main purpose of research honeypots is to discover new attacks or threats, new techniques and motives of Black hat community.

1.5.3 Network-based Antivirus systems

These are the solutions which are installed on gateway among networks to prevent threats over network. Network based antivirus systems which capture message and then compare the content of messages to a database of known virus signatures [6].

1.5.4 Intrusion Detection System

Intrusion Detection System is a mechanism which monitors the events that occur in a network or computer system and analyses them for signs of possible intrusions or incidents [7]. The main function of Intrusion Detection System is to detect both external threats or attacks and internal misuse of network resources. Intrusion detection Systems can be software or hardware devices which analyze the network

traffic in order to detect malicious traffic. Intrusion detection Systems perform the following functions:

i. Monitor and analyze the activity of user and the system
ii. Auditing of vulnerabilities and system configuration.
iii. Assessing the integrity of data files.
iv. Analyze the abnormal activity.
v. Operate system audit [8].

- **Advantages**

❖ Intrusion Detection Systems are very easy to deploy.
❖ These can be easily installed and cost of installation is very low.
❖ Network based intrusion detection systems are very useful in detecting many attacks by examining the headers of packets.
❖ IDS are capable of detecting attacks or threats very easily.

- **Disadvantages**

❖ Intrusion Detection System is not the solution to all security problems.
❖ Sometimes IDS raise false alarms.
❖ IDS overload the work by capturing large number of alerts.
❖ Intrusion detection system does not perform functions itself.

1.6 Architecture of Intrusion Detection System

The main components of Intrusion Detection System are Information Collection, Detection and Response. Figure 1.2 shows components of Intrusion detection system.

- **Data gathering device:** Data gathering device is used to gather information from network system that is going to be monitored. It performs like an agent which continuously watches or monitors the network in real time. Examples of input to a sensor are log files, network packets and system call traces. Sensors collect and forward this information to the analyzer [9].

- **Detector:** Detector is used to process the data that is gathered from sensors to identify threats and raise alarms based on defined rules.

- **Knowledge base:** Knowledge base stores the information in processed format that is captured by sensors (e.g. knowledge base of intrusions and their signatures, data profiles, filtered data etc.). This information is usually provided by network experts. All the information about the previously detected attack signatures or pattern should be present in the database. When the sensor detects some kind of malicious activity or signature it matches with the content database and report to attack response component.

- **Configuration device:** It gives useful information about the current state of the intrusion detection system [9].

- **Response component:** It raises alarm whenever an intrusion or attack is detected. Response component can either send an alarm or an email notification about the intrusion detected to the administrator depending on the type of configuration [9].

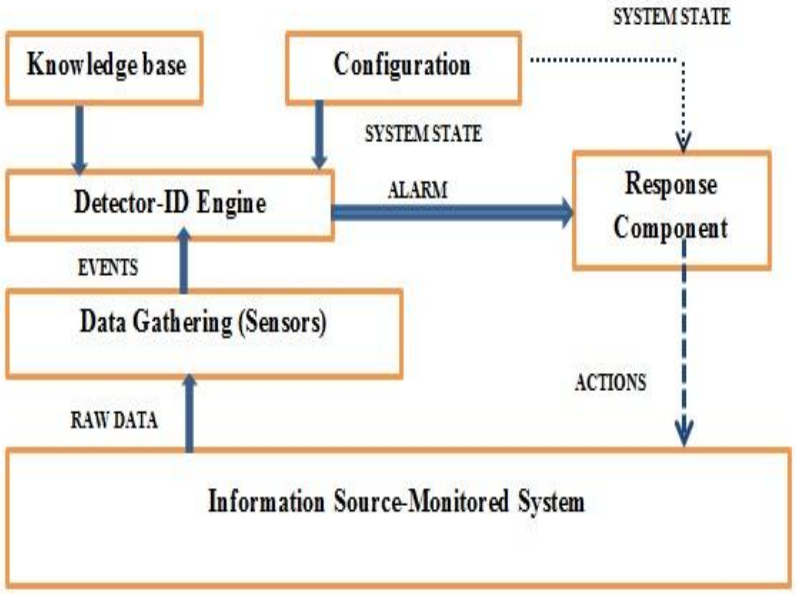


Figure1.2 Architecture of Intrusion Detection System

1.6.1 Categories of Intrusion Detection System

The intrusion detection system has divided into two categories described as follows**.**

i. Network based Intrusion Detection Systems

Network based intrusion detection systems is an intrusion detection system which uses network based approach to find the intrusions in the network system. Network

based IDS include a network appliance an operating Network Interface card. The IDS is implemented along boundary or a network segment and examine all the traffic. Network intrusion detection systems are used to prevent an attack before it causes any damage to network resources. They collect data from network and compare data with known signatures. Every packet is compared against the signatures in order to detect intrusion. NIDS are effective and efficient for monitoring incoming and outgoing network traffic.

- **Advantages**

❖ Network based IDS are very effective in detecting unauthorized access and denial of service attacks.
❖ Network based IDS are very easy to implement in the network.
❖ Network based IDS does not affect network speed.

- **Disadvantages**

❖ Sometimes NIDS gets overloaded due to high network traffic and starts dropping the packets**.**
❖ Network based intrusion detection systems are susceptible to slow and low attacks.
❖ NIDS are not capable of analyzing encrypted data.

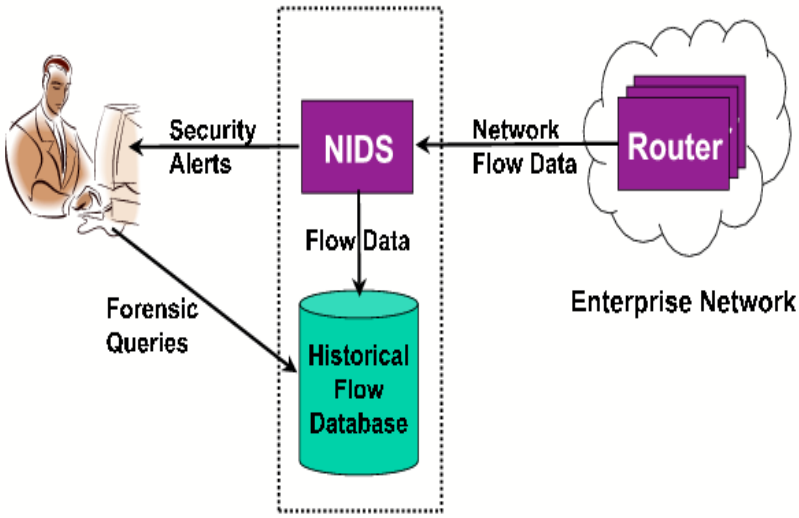


Figure 1.3 Network based Intrusion Detection System

ii. Host based Intrusion detection System

Host based Intrusion Detection System examines logs, registry settings, specific logs on a personal computer. HIDS raises alarm when there is any detection, modification, access to object that is being monitored. Host based IDS have components which watch login by user and processes and parse system logs. Host based IDS are very easy to install on host machine and makes the system versatile. HIDS can be installed on different machine such as workstations, notebook computers and servers.

- **Advantages**

❖ Host based intrusion detection systems capable of using things such as system services, registry events, log files etc.
❖ HIDS are able to detect those attacks or intrusions that are ignored by NIDS.
❖ Host based IDS can detect Trojan Horse or other intrusions which can cause damage to system.

- **Disadvantages**

❖ Host based IDS takes too much time to detect an intrusion or threat.
❖ Host-based IDS are dependent on host operating system.
❖ HIDS consumes extra memory from the host on which it resides.

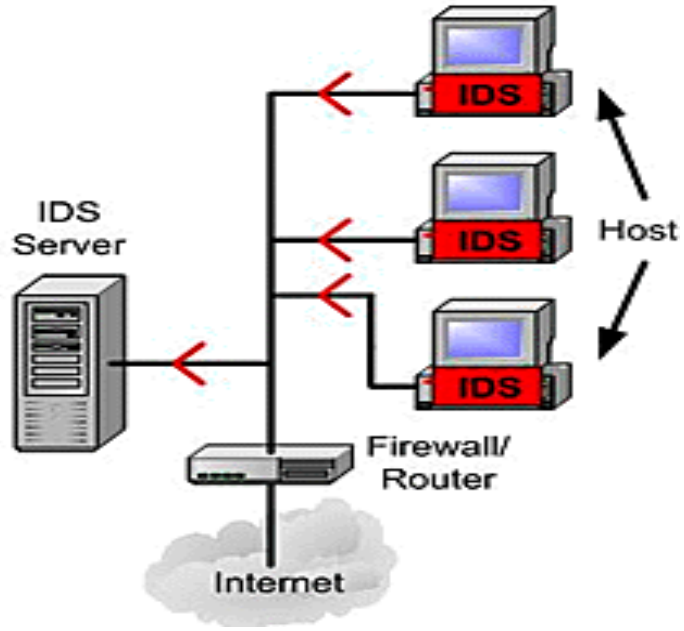


Figure 1.4: Host Based Intrusion Detection System

iii. Hybrid Intrusion Detection System

In hybrid IDS, both the Network Based Intrusion Detection System and Host Based Intrusion Systems are combined together to perform the task of detecting intrusions.

1.6.2 Types of Intrusion Detection System

There are two types of intrusion detection systems described below:

- **Misuse Detection**

The main objective of misuse detection is to use an expert system to detect intrusions based on previously defined rules or patterns or knowledge base. There various methods that are used by misuse detection [10]. These are described below:

i. **Signature based:** In this method the data which is collected from network is match with available signatures in the database.
ii. **Rule based:** In this method implication rules like set of "if-then" are used to characterize intrusions.
iii. **State transition:** In this method, a finite state machine is used to identify intrusions or attacks.

- **Anomaly Detection**

An intrusion detection system that examines network traffic and detects data which is not valid, not correct or abnormal is known as anomaly based approach. This approach detects unwanted data [10]. These are divided into categories mentioned below:

i. **Statistical based method:** Statistical method monitors the behavior of user or system by using measurements of variables over time like login or logout time of every session.
ii. **Distance based method:** Distance based methods are used to overcome limitations of anomaly detection approaches and statistical outlier. They compute distance between points to detect outliers.
iii. **Rule based system:** Rule based systems are used to characterize networks, computer systems and behavior of user by set of rules.

iv. **Profiling method:** In this method, profiles of normal behavior are used for various types of network traffic; programs, users etc. and deviations are termed as intrusions.
v. **Model based method:** In this method, anomalies are examined as deviations for models which represent the normal behavior.

1.7 Organization of Thesis

Chapter 1 describes the brief introduction of network security and need for network security. This chapter also discusses about attacks in network security, mechanisms in network security and various available security tools.

Chapter 2 discusses about brief introduction of machine learning, WEKA and KDD dataset. This chapter also provides the survey of various techniques used for machine learning

Chapter 3 includes problem statement of thesis, goals and objectives to be achieved to carry out the thesis work.

Chapter 4 presents the methods that are used to achieve the goals, objectives, experimental setup and implementation details of our approach i.e. misuse detection using machine learning.

Chapter 5 provides testing and results of thesis work.

Chapter 6 states conclusion of results carried out and future scope of work done.

Chapter 2

Literature Survey

In this chapter, we have discussed about machine learning, algorithms for machine learning, introduction to WEKA and KDD dataset. This chapter also involves literature survey of techniques used to detect intrusions using misuse detection and machine learning algorithm. Machine learning techniques have been discussed by the following authors in their research work.

2.1 Introduction to Machine Learning

Machine Learning (ML) is a major part of Artificial Intelligence (AI). Machine Learning is a algorithmic mechanism that permits machines to learn from examples, analogy and experience. The outcome of this learning procedure is actionable information or knowledge that can be applied to resolve specific problem. In the area of Intrusion Detection, learning requires finding patterns of intrusive behavior or normal behavior by examining ideal data of such activities [44]. Machine Learning is subdivided into three types. These three types are explained below:

- **Predictive or Supervised Learning:** In this learning technique, an algorithm creates a function that is used to map input values to desired output values. Examples of supervised learning are Back Propagation Neural Network and Logistic Regression.

- **Descriptive or Unsupervised Learning:** In this technique, only input data is provided. The main purpose is to figure out the regularities in the input values. Unsupervised learning is very helpful in discovering hidden patterns in data.

- **Reinforcement Learning:** In this technique, an algorithm learns a policy that how to perform given an observation of the real world. Each and every task has some effect in the environment and environment gives feedback that directs the learning algorithm.

2.2 Algorithms for Machine Learning

There are various types of machine learning algorithms. These algorithms are briefly described as follows:

- **Artificial Neural Network:** This algorithm is inspired by human nervous system and it is used for approximating the solutions. Collection of neurons connected together to form biological neural network is used for deriving solutions and is called hidden layer. These systems are largely used in statistics [35].

- **Inductive Logic Programming:** This technique uses logic programming to derive hypothesis about the facts stored in database. The facts and the rules are knowledge base for inductive logic programming. Based on the knowledge base and some positive and negative examples from the facts conclude the hypothesis. Prolog language is used for this technique. This type of machine learning is useful in natural language processing field [36].

- **Decision Tree Learning:** In decision tree learning, decision trees are used for classification of data and it is used in data mining. Decision tree helps to create a model which takes input and predict the value of target variable. All nodes except leaf nodes represent input variables and leaf nodes are targeted values. Any path from root to leaf represents solution of the problem [37].

- **Association Rule Learning:** This method is used to identify relations between large set of databases, based on those relation design some rule. The rule generation in this algorithm took two processes. In first process, minimum support is applied to all attributes in dataset and in second step minimum confidence is maintained to derive rules [38].

- **Support Vector Machine:** Support Vector Machine is supervised machine learning model which is used for classification of data. Suppose we have two categories of data in training dataset and SVM helps to categorize new data. SVM constructs a hyperplane in n dimensional space which can be used for classification of data [39].

- **Clustering:** Clustering is the process of grouping similar objects together in same group while some other objects in other groups. Each group is called cluster. This technique is used in image analysis, machine learning, data mining, statistical data analysis. Some of the clustering techniques are hierarchical clustering, centroid based clustering, distribution based clustering and density based clustering [40].

- **Bayesian Network:** Bayesian Network is a probabilistic graph model that represents set of random variable and their conditional independence with the help of directed acyclic graph. Edges in this network represent conditional dependency and nodes which are not connected are conditionally independent. Each node has probability function which takes input from parent variable and gives output as probability of variable represented as node [41].

- **Genetic Algorithm:** Genetic Algorithm works on principle of natural selection and evolution. Algorithm starts with population of individuals and evolves by selection, crossover and mutation process. Output of this algorithm is highly fit individuals which further reproduce to form offspring. In this way, best solution space is obtained from population of solution space [42].

- **Feature Learning:** Feature Learning is motivated by the fact that machine learning tasks such as classification requires input which is easy to process. So, it is necessary to discover relevant features from raw data. Examples of feature learning techniques are principal component analysis, dictionary learning [43].

2.3 Introduction to WEKA

WEKA is the abbreviation of Waikato Environment for Knowledge Analysis. It is very popular suite of machine learning software which is written in Java. It is software which is freely available under the GNU General Public License. The WEKA toolkit contains a collection of visualization tools and algorithms for data analysis and modeling, associated with graphical user interfaces for easy access to this functionality. WEKA is very portable because it used Java Programming Language that can run in almost all modern platforms that contain modeling technique and a large number of data pre-processing technique and easy to use by beginner with its

easy to use graphical user interface. WEKA supports various standard data mining tasks, more specifically, clustering, data pre-processing, classification, visualization, and regression and feature selection. The techniques of WEKA are predicated on the assumption that the data is available as a relation or single flat file where each and every data point is described by a fixed number of attributes (normally, nominal or numeric attributes, but some other attributes are also supported). The data file normally used by WEKA is in ARFF file format that consists of very special tags to indicate different things in the data file. However csv format and some other formats can also be used. The format contains attribute types, attribute values, attribute names and the data. The Explorer is the main interface in WEKA. It has a set of panels, each of which can be used to perform a certain task. WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query [11].

2.4 Description of KDD Dataset

Since 1999, KDD'99 has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo *et al.* and is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type [12]. There are four major categories of attack. All attacks fall under these four categories.

- **Denial of Service Attack (DoS):** is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests.

- **User to Root Attack (U2R):** is a class of exploit in which the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.

- **Remote to Local Attack (R2L):** occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on

that machine exploits some vulnerability to gain local access as a user of that machine.

- **Probing Attack:** is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

In 1999, the original TCP dump files were pre-processed for utilization in the Intrusion Detection System benchmark of the International Knowledge Discovery and Data Mining Tools Competition. To do so, packet information in the TCP dump file is summarized into connections. Specifically, "a connection is a sequence of TCP packets starting and ending at some well-defined times, between which data flows from a source IP address to a target IP address under some well-defined protocol". This process is completed using the Bro IDS, resulting in 41 features for each connection; Features are grouped into four categories namely, basic features, content features, time based traffic features, host based traffic features.

2.5 Survey of Existing Misuse Detection Systems and Machine Learning Techniques

Arjunwadkar *et al.* (2015), proposed an Intrusion Detection System with Machine Learning model which Combined Hybrid Classifiers i.e. Naive Bayes classifier and C 4.5 classifier to detect intrusions. In this proposed model, a multi-layer Hybrid Classifier was used to estimate whether the action is an attack or normal data. Firstly they have built a misuse detection model based on the C4.5 decision tree algorithm after this the normal training data was decomposed into smaller subsets. Hybrid Classifier was used to reduce the dimensions of features. Results from experiment showed that Naive Bayes algorithm both gave the higher accuracy and low time complexity [13].

Khurram *et al.* (2014), proposed a novel framework which was platform independent behavior-based anomaly detection used for mobile devices. This framework was based on a concept that a unique usage patterns has been assigned to every smart phone user. These patterns were modeled into a profile to uniquely identify a user. They have used K-Means algorithm for creating create baseline usage profile for each smart phone user. They have tried to fit data of every user into

profiles of other users to detect accuracy of profiling and monitored the deviation. Results from this experiment showed 80% of user profiles was having accuracy (measure of uniqueness).This framework was the first to implement anomaly detection on mobile devices and provided means of transparent and means of continuous user authentication on mobile devices. The results of their experimental procedure showed that it is possible to achieve high accuracy with behavior based profiling [14].

Yogita *et al.* (2013), proposed Intrusion Detection System by using data mining technique like Support Vector Machine (SVM). SVM was used to classify the attacks. In this system, experiments were conducted by using NSL-KDD Cup'99 dataset which was improved version of KDD data set. Their proposed model have reduced the time required to build model for classification and increased the intrusion detection accuracy when Gaussian RBF kernel was used. They have conducted the experiment 10 fold cross validation and Gaussian RBF kernel the time required to build model was 77.07seconds and attack detection accuracy was 94.1857%**.**By using proper selection of SVM kernel function such as Gaussian Radial Basis Function, False Positive Rate (FPR) was decreased [15]**.**

Manju *et al.* (2013), proposed a comparison of most commonly used machine learning techniques which are based on artificial neural networks, genetic algorithm and support vector machines to detect intrusion. They have used pattern classification, ensemble classifiers, hybrid classifiers and single classifiers. From the results they have concluded that Genetic algorithms are better than other techniques [16].

Neethu *et al.* (2012), proposed a framework based on network intrusion detection system where they have used the combination of and Principal Component Analysis algorithm and Naive Bayes. They have used the PCA to reduce the dimensions of dataset and Naive Bayes classifiers were used to classify the attacks. The dimensionality reduction increased the speed of execution and hence increased the overall performance. The classifier separated the input dataset into two classes which are Normal and Anomaly. They carried out their experiment by using KDD dataset. The dataset represented attribute value of a class in the network data flow, and each class is labeled either as normal or as an attack. This approach resulted into higher

detection rate, low time consuming and has low cost factor. This approach was faster than some other existing systems [17].

Heba *et al.* (2012) [18], proposed multi-layer intrusion detection model which was designed and developed to improve the detection and to achieve high efficiency and classification accuracy rate. They applied various Machine learning techniques such as C5 decision tree, Multilayer Perception (MLP) neural network and Naive Bayes to gain ratio for selecting the best features. This proposed system was layered-model approach. It was divided into two stages. First stage detected normal attacks. Second stage classified the attacks detected by stage first. Stage second consisted four layers and each layer was examined using different machine learning models. Results from experiment showed that multi-layer model which uses C5 decision tree achieves higher classification rate of accuracy than MLP and naive Bayes [18].

Annie *et al.* (2012), proposed a framework for network anomaly detection by using machine learning techniques. KDD99 benchmark dataset was taken to evaluate the performance of the system. They have examined an anomaly detection system by using machine learning algorithms such as Support Vector Machine and Principal Component Analysis. PCA was used to reduce to the dimensions of features resulted into increased classification accuracy and decreased execution time for classification. SVM was used to classify data as normal or an anomaly connection. This framework gave better classification results. Results from the experiment showed that classification using dimensionality reduction was more accurate as the number of misclassification decreases [19].

Ahmed *et al*. (2011), proposed two engines which are back-propagation neural network intrusion detection system and the radial basis function neural network intrusion detection system. These two engines were tested against traditional and other machine learning algorithms. They have used KDD dataset to perform the experiment. They focused on two classifications first is a single class (normal, or attack) and second is multi class (normal, DoS, PRB, R2L, U2R). They have investigated supervised learning approach to intrusion detection and used iterative reduction algorithm to reduce the training time. From the experiment, they have concluded that back-propagation neural networks produced better results than radial basis function [20].

Gireesh *et al*. (2010), proposed an intrusion detection method by using SVM system on Rough Set Theory (RST) to detect network intrusion. KDD data set was used in this method. They used RST to reduce the features from 41 to 29. In this method they captured the packets and then by using RST they minimize the data. After this, pre-processed data was sent to SVM to learn and test. In this way, the space density of data is reduced. Reduced number of features resulted in higher accuracy than having full features. They have also compared the results with Principal Component Analysis (PCA). By using RST and SVM, they increased the accuracy and reduced the false positive rate [21].

Hadi *et al.* (2010), proposed a combinatory system. They have combined the techniques of machine learning to detect attacks. This method has also used the KDD dataset to analyze the attacks. They have used various combinations of classifiers. Unbalanced data was used in this method. They combined the KNN and Decision Trees and then combined the 1NN, 2NN, 3NN, SVM and Tree to achieve the better accuracy. From experiment they resulted that accuracy of system will be constant if we increase the number of classifiers [22].

Subbulakshmi *et al.* (2010), proposed a two phase automatic alert classification system which was used to assist the analyst to identify the false positives. During first phase, alerts were collected from one or more sensors were normalized and then similar alerts are then combined to form a meta-alert. An asset database was used to verify these meta-alerts to find out irrelevant alerts. An optional alert generalization was also performed for root cause analysis and hence reduced false positives with human interaction. During second phase, the reduced alerts were labeled and passed to an alert classifier where the machine learning techniques was used for building the classification rules. This resulted into automatic classification of the alerts and reduces the workload of the analyst considerably [23].

Tich *et al*. (2010), proposed an innovative Machine Learning framework to detect different types of intrusions by using different classifiers which included learning algorithms and different attribute selections. Outputs of these different types of classifiers were then combined by using appropriate voting techniques. KDD dataset was used in this technique. This proposed framework worked well in intrusion detection system problem and resulted into higher accuracy rate and made the system

robust while offering “affordable” computation as compared to already existing techniques [24].

Mrutyunjaya *et al*. (2009),proposed machine learning algorithms such as an AdaBoost and Random Forest along with Naive Bayes, to build an efficient and effective intrusion detection model. They have performed their experiment over KDDCup’1999 datasets. They have employ ensemble algorithm to improve the performance of detection. In this experiment, packets were captured then features are constructed. After that dataset was fed into builder module to build patterns and these patterns were used as input to detector module. This module classified connections into normal traffic or different intrusions. Developed algorithms for learning classifiers were success in detecting attacks in network when KDD dataset was used than standard data mining techniques which are based on neural networks [25].

Pavel *et al.* (2007), proposed an alternative approach based on machine learning techniques which enable automatic construction of profiles for normal packet payloads. They have focused on two important components which are feature extraction and anomaly detection methods. They have conducted their experiments on PESIM 2005dataset. They have presented the framework for detection of unknown attacks by using machine learning technique. In this approach, methods are evaluated on mixtures of unseen normal and attack data which contains only 2% to 14% malicious connections. Results from this approach showed remarkable accuracy of 77.100% at zero low false positive rates [26].

Taeshik *et al.* (2007), proposed a new SVM approach called Enhanced SVM which was a combination of one class SVM and unsupervised learning. In this approach performance was increased by a new technique referred as Anomaly Detector using Enhanced SVM. Firstly they created a profile of normal packets with the help of Self-Organized Feature Map (SOFM),for Support vector machine learning without any knowledge. Secondly they used a packet filtering scheme. Thirdly they used a feature selection technique to extract optimized information from raw packets and used flow of packets during pre-processing for considering relationship among inputs. Finally they demonstrated the effectiveness of the Enhanced SVM with these techniques. The overall goal of this technique was to propose a framework to detect and classify novel attacks. Three kinds of training sets were used to perform the

experiment. This approach was compared to real world Network Intrusion Detection Systems (NIDS) [27].

Latifur *et al.* (2007), proposed a method Clustering Trees to increase training time of Support Vector Machine (SVM) and to detect network based anomalies. They have used the combination of Dynamically Growing Self-Organizing Tree and Support Vector machine. In this paper, KDD dataset was used to classify the attacks. They selected only 14% of training data. Even they compared this method with Rocchio Bundling method which is used to reduce data points. With the help of reduction technique like clustering analysis, they enhanced the speed of training process of SVM. They have improved the accuracy of classifiers by using Support vectors [28].

Srinivas *et al.* (2002), proposed intrusion detection systems using support vector machines and neural networks. They have carried out their performance on a set of benchmark DARPA data. Results from this experiment showed that both the neural networks and support vector machines delivered highly accurate result such as greater than 99% accuracy on testing set and showed compatible level of performance. The training time for SVMs was significantly shorter which was17.77 sec. The running time of SVMs was also notably shorter. On the other hand, neural networks have already proven to be useful in many intrusion detection systems, and are especially suited for multi-category classifications [29].

Chris *et al.* (1999), proposed an application which enhanced the domain knowledge using machine learning techniques for creating rules for an intrusion detection system. They have employed decision trees and genetic algorithm for automatically generating rules for classifying connections in network. Genetic algorithms were used to create optimal solution to the problem and Decision trees were used to classify data with common attributes. They have developed genetic algorithm package which is a generalization of the classic genetic algorithm. Rules deployed and developed in this application differentiate anomalous connections from normal network connections. The main result of this method was the production of rules for compilation into the expert system. The goal of this application was to produce a dynamic rule base capable of detecting new attack signatures [30].

Table 2.1 Comparison of approaches using machine learning by various researchers

S.No.	Year of publication	Authors	Approach	Results
1.	2015	Arjunwadkar *et al.*	Naive Byes classifier + C 4.5 classifier	Higher accuracy and low time complexity of Naïve Bayes.
2.	2014	Khurram *et al.*	K Means algorithm	Higher accuracy rate.
3.	2013	Yogita *et al.*	IDS + Support Vector Machine (SVM)	Decreased false positive rate.
4.	2013	Manju *et al.*	ANN + Genetic algorithm + SVM	Better results of Genetic algorithm.
5.	2012	Neethu *et al.*	Principal Component Analysis + Naive Bayes	Higher detection rate + low cost factor.
6.	2012	Heba *et al.*	C5 decision tree + MLP neural network + Naive Bayes	Higher classification rate of C5 decision tree.
7.	2012	Annie *et al.*	Support Vector Machine + principal component Analysis	Better classification of attacks.
8.	2011	Ahmed *et al.*	Back-Propagation neural network IDS + radial basis function neural network IDS	Reduced the training time.
9.	2010	Gireesh *et al.*	SVM + Rough Set Theory	Reduced Space Density.
10.	2010	Hadi *et al.*	KNN + Decision tree	Constant accuracy of system.
11.	2010	Subbulakshmi *et al.*	Two phase automatic alert classifier system	Alerts automatically classified.
12.	2010	Tich *et al.*	Machine Learning + different classifiers	Higher accuracy rate.
13.	2009	Mrutyunjaya *et al.*	AdaBoost + RandomForest + Naive Bayes	Detected new attacks.
14.	2007	Pavel *et al.*	Alternative Approach Based on Machine	Higher detection rate.

			Learning	
15.	2007	Taeshik *et al.*	One Class SVM + Unsupervised Learning	Classified novel attacks.
16.	2007	Latifur *et al.*	Dynamically Growing Self-Organizing Tree + SVM	Enhanced training time of SVM.
17.	2002	Srinivas *et al.*	Support Vector Machine + Neural Network	Highly accurate results.
18.	1999	Chris *et al.*	Decision Tree + Genetic Algorithm	Produced dynamic results.

Chapter Summary: In this chapter, we have discussed brief introduction about machine learning. Also, machine learning algorithms are presented in this chapter. The survey of the existing machine learning techniques by various researchers is presented here and we have used a tool named WEKA in this thesis work.

Chapter 3
Problem Statement

With the increasing demand of network and advances in the field of network, now days each organization wants to have their own network and furthermore they want to connect or interact with each other in a reliable way. So network security is becoming more and more important and also getting more complicated issue with recent advances and with increasing demand.

To protect the enterprise, security managers have deployed a variety of technologies including honeypots, firewalls, and intrusion detection/prevention systems. Intrusion detection systems are divided into two types: anomaly detection and misuse detection. Misuse detection systems work by detecting attacks which are based on previously defined signatures whereas anomaly detection systems only analyze normal network traffic and profile network traffic patterns.

Various researchers are working on improving accuracy rates of network attacks. Techniques used by them are combining one or more machine learning algorithms. The problem here is to improve accuracy rate of intrusions.

Mostly all machine learning algorithms generate false positives. Any Machine learning classifier is considered to be good if it generate less false positive rate. The problem here is to minimize false positive rates while improving the accuracy rate of network attacks.

The major concern in high speed networks is to reduce time taken for detecting intrusions. So, the problem here is to reduce the processing time taken by various classifiers. Another problem is to enhance true positive rates of attacks detected by machine learning classifiers

In the thesis the topic of concern is Machine Learning and Misuse Detection Systems. There are various techniques and machine learning algorithms used by various researchers to enhance security of systems in terms of accuracy rate, increasing true positives rate and minimizing false positive rates.

The objectives of the thesis include the following:

- To design a hybrid method which is a combination of Misuse Detection and Machine Learning techniques.
- To implement the proposed method resulting in high Accuracy Rate, True Positive Rate, minimizing Processing Time and False Rate.
- To verify and validate the proposed method.

Chapter 4
Implementation Details

The solution to the problem discussed in previous chapter is to develop a hybrid intrusion detection system which is based on combined approach of misuse detection and Machine learning techniques. This chapter provides implementation details of the hybrid approach. The methodology used to implement the objectives and experiments performed are explained in this chapter.

4.1 Methodology

This section presents the methods used to carry out the thesis work. Figure 4.1 describes the schematic diagram of the system combining misuse detection and machine learning techniques. The proposed approach contains two stages namely, misuse detection and machine learning technique. The datasets used in this approach are taken from http://nsl.cs.unb.ca/NSL-KDD/. In first stage, i.e. misuse detection, the training data attributes are compared with testing data attributes and results are stored in database. In second stage, these results are applied to machine learning classifiers such as NaiveBayes, Decision Table, MultilayerPerceptron, J48 and RandomForest. Moreover we have also applied classifiers on dataset without performing misuse detection and results from both the datasets are compared. The methodology consists of six major steps namely, Pre-processing, Attribute selection, Pattern Matching and results Interpretation. Each process is explained in detail below.

- **Pre-processing:** For both training and testing of connections two datasets, KDD cup'99 training and KDD cup'99 testing datasets are used. In this phase, duplicate data from both the datasets has been removed. Duplicate data will result into few types of attacks. KDD dataset contains attributes which are either in integer or string format. Categorical data is also converted into numerical data. In Table 4.1, we have assigned integer values to string attributes. The processed data is used as input to the system. Like class attribute has twenty one different string values, so we converted them to integer values ranging from 1 to 21.We can assign any value, it is not mandatory to assign sequentially, it gives same result with any value but two strings can't have same value.

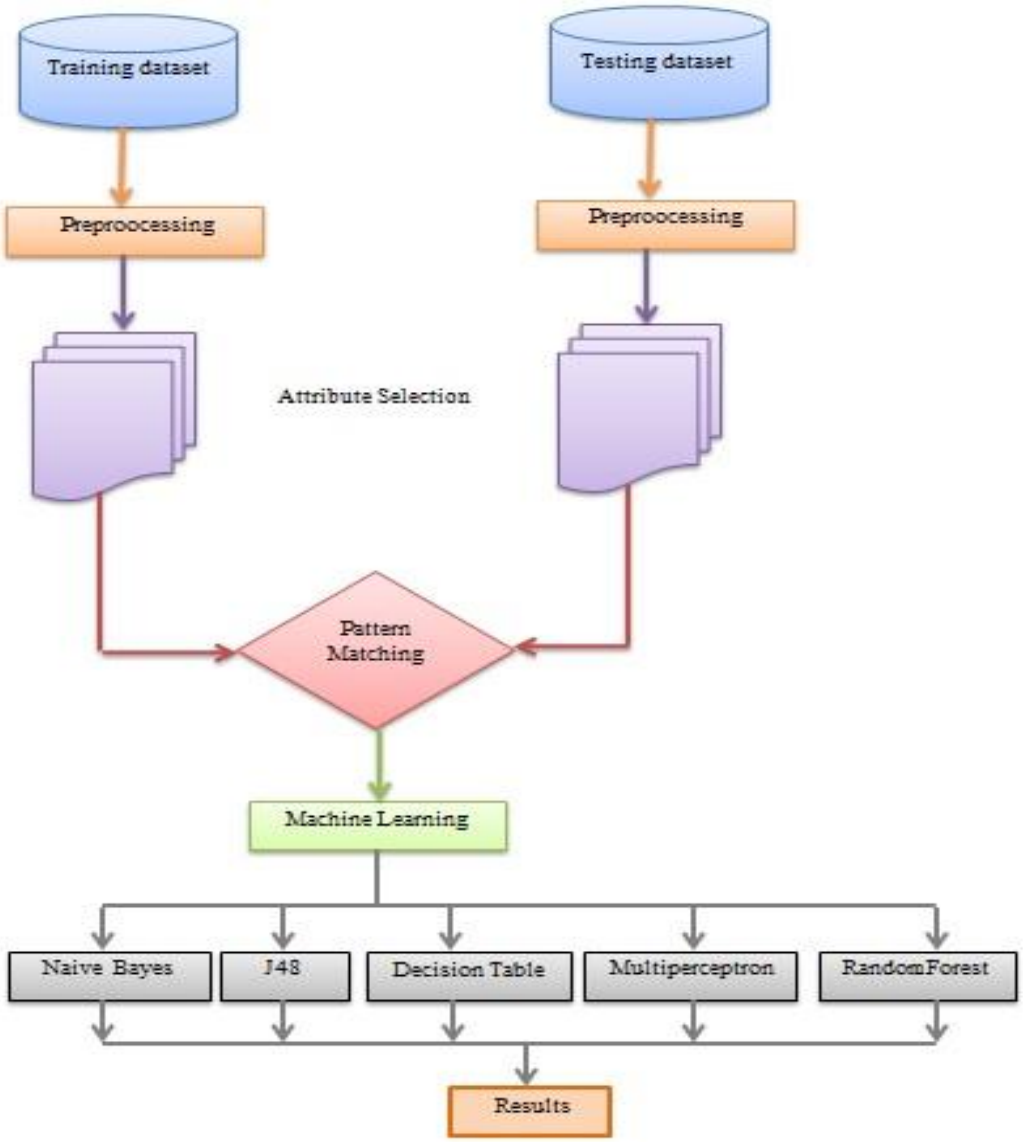
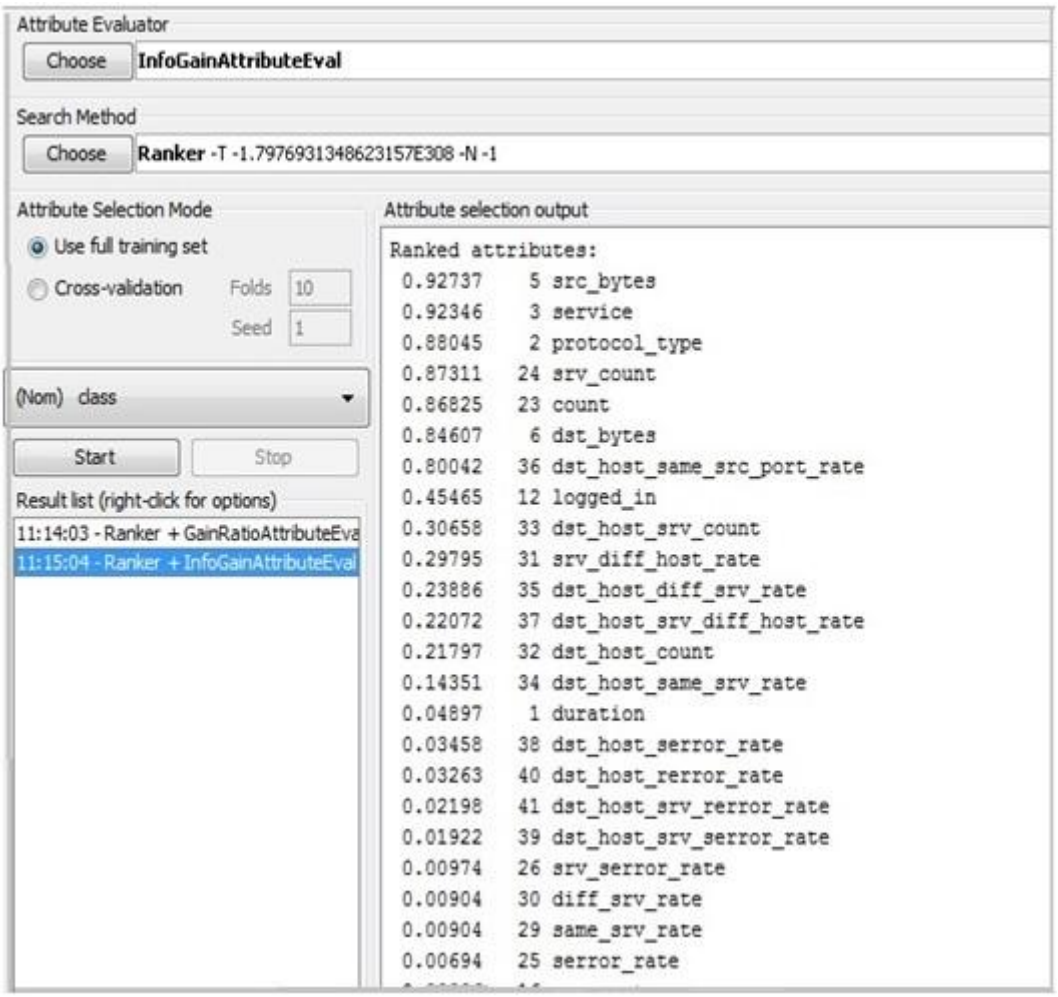


Figure 4.1 Diagram of the Proposed system

- **Attribute selection:** Attributes are selected with the help of a technique that is described in WEKA (Waikato Environment for Knowledge Analysis) i.e. InfoGainAttributeEval. The package weka.attributeSelection includes InfoGainAttributeEval class. It works by evaluating the worth of an attribute by measuring information gain with respect to the class. InfoGainAttributeEval supports Binary class, Missing class values, Numeric values. KDD data set consists of 41 features. We have selected only three attributes to classify the attacks. These are service, protocol_type, src_bytes. These three attributes consume less time as compared to 41 attributes. Snapshot 4.1 shows the selected attributes.

Table 4.1 Conversion of string values to integer value

Class name	**Value**
smurf	1
normal	2
neptune	3
snmpgetattack	4
portsweep	5
ipsweep	6
nmap	7
xlock	8
multihop	9
worm	10
xterm	11
teardrop	12
sqlattack	13
apache2	14
satan	15
pod	16
warezclient	17
buffer_overflow	18
guess_passwd	19
warezmaster	20
back	21



Snapshot 4.1 Attribute Selection using InfoGainAttributeEval in WEKA.

- **Pattern matching:** Pattern matching is also termed as misuse detection. This mechanism uses a predefined set of intrusions so that the system already knows the possibility of abuse or misuse. In this phase, attributes that are selected from both testing dataset and training dataset are matched. If any match is found, then system will find a class corresponds to training data that is already stored in database and system will assign this class to testing data set.

- **Machine learning:** In machine learning, systems learn from given data. It is a type of artificial intelligence. The purpose of the machine learning is to focus on evolution of computer programs that can teach themselves to proceed and modify when exposed to new data. WEKA is used to carry out the experiment as WEKA is a workbench for machine learning. WEKA contains different types of classifiers to classify the attacks. The following classifiers are used in this methodology and are described below:

a) **NaiveBayes:** Naive Bayes technique is based on Bayesian Theorem. This is very simple classifier that comes under the category of supervised learning. Naive Bayes classifiers assume that value of a specific attribute does not depend on values of another attributes, given the class variable. Naive Bayes classifiers work well in complex situations [31].

b) **Decision Table:** Decision Table algorithm classifier is used to summarize the dataset with a decision table which includes the same kind of attributes as the real dataset. After this a category is assigned to new data item by looking the line in the decision table which compares the non-class values of the data item. Decision Table classifier is a very simple algorithm [32].

c) **MultilayerPerceptron:** Multilayer perceptron classifiers are based on artificial neural network model which works by mapping sets of input values onto appropriate output values. Multilayerperceptron consists of more than three layers of nodes with non-linear activation function. This is very powerful classifier [11].

d) **RandomForest:** This classifier is the best suited algorithm capable of classify very large amount of data with higher accuracy rate. These classifiers are very easy to use and learn. Random Forest classifiers provide effective and efficient methods to estimate the missing data. Trees generated can also be stored for future use [33].

e) **J48:** J48 classifier is a very simple C4.5 decision tree which is used for classification purpose. This classifier works by creating binary trees. These trees are used to model the classification process. J48 classifier ignores the missing values when it is building a tree. J48 classifier has higher rate of accuracy [34].

- **Results:** Results obtained from misuse detection are applied to machine learning classifiers. This technique resulted into better accuracy of detection rate. Also the time taken by our approach is comparatively less.

4.2 Experimentation

We have conducted two experiments to carry out the evaluation of accuracy rate with two approaches. Subset of KDD Cup'99 dataset is used for both training and testing of data. In this experiment, two data sets are used namely Testing and Testing^{+}. Testing consists of selected records from KDD data set whereas Testing^{+} includes the records obtained from misuse detection. Below two sections explain both the experiments in detail.

4.2.1 Experiment 1: Using Machine Learning

In this experiment, machine learning classifiers are applied to dataset named Testing. These are five classifiers namely NaiveBayes, Decision Table, RandomForest, J48 and MultilayerPerceptron. These five classifiers resulted into accuracy rate, true positive rate and false positive rate.

4.2.2 Experiment 2: Using Misuse Detection and Machine Learning

In this experiment, training data attributes are matched with testing data attributes and if there is a complete match, means all attributes which are selected for intrusion detection from training dataset are matched with same attributes of testing data, and then training data class is assigned to testing connection. Three attributes are selected based on their info gain value which is described in Section 4.1. Selected attributes are src_bytes, service, and protocol_type. The algorithm used in this thesis for misuse detection is explained below.

Algorithm 4.1: Misuse Detection:

```
Step 1 :  Subset of KDD cup'99 training and testing
          dataset.

Step 2 :  Attribute selection using
          InfoGainAttributeEval.

Step 3 :  Loading testing data.

Step 4 :  For each connection in testing data.
```

Step 5 : Match attribute of training and testing data.

Step 6 : If (attributes are identical)

6.1 : Assign training class attribute value to testing connection.

Step 7 : Else

7.1 : Quit

Step 8 : Result of testing data is stored.

Step 9 : Applied on Machine learning classifiers.

Step 10 : Interpretation of results.

Chapter Summary: In this chapter, methodology of the proposed system is discussed in detail. Two experimentations are used to describe the methods. In first experiment, Machine learning classifiers are applied to Testing data. In second experiment, both misuse detection and machine learning technique is used to carry out the results.

Chapter 5

Results and Discussion

In this chapter results achieved from both the data sets namely Testing and Testing^{+} are explained. Firstly, results obtained from Testing and Testing^{+} are analyzed. Secondly, results from both data sets are compared on the basis of accuracy rate, time taken, true positive and false positive rates. Also we have compared the results obtained by our approach with existing techniques.

5.1 Analysis of Testing Dataset

In this section, we analyze the accuracy rate, true positive rate and false positive rates of dataset termed as Testing.

5.1.1 Accuracy Rate of Testing Dataset

Table 5.1 shows accuracy rates of dataset named Testing. From Table 5.1, we can analyze that accuracy rate of RandomForest is higher than other classifiers. NaiveBayes classifier has least accuracy rate among all these classifiers. Accuracy rate of Decision Table classifier and MultilayerPerceptron is almost same. Accuracy rate of J48 classifier is slightly less than accuracy rate of RandomForest classifier. Figure 5.1 shows the accuracy rates of RandomForest, J48, Decision Table, MultilayerPerceptron and NaiveBayes.

Table 5.1 Accuracy rate of Testing Dataset

Classifiers	Testing
RandomForest	92.335
J48	91.690
Decision Table	88.968
Multilayerperceptron	88.467
NaiveBayes	77.793

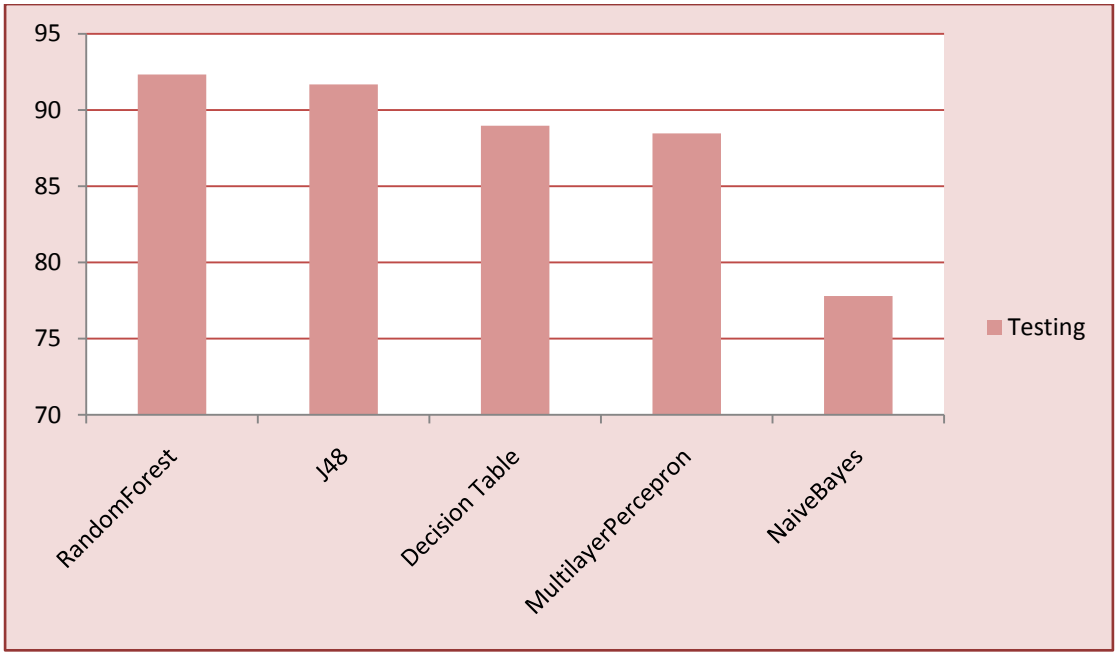


Figure 5.1 Accuracy Rate of Testing Dataset

5.1.2 True Positive Rates of Testing Dataset

Table 5.2 shows the true positive rates of dataset named Testing. We have concluded from Table 5.2 that RandomForest classifier has higher value of true positive rate i.e. 0.92 whereas NaiveBayes has least value of true positive among these five classifiers. MultilayerPerceptron and Decision Table classifier have almost similar rates of true positive. J48 classifier has true positive rates less than RandomForest i.e. 0.918. Figure 5.2 shows true positive rates of Testing dataset.

Table 5.2 True Positive Rates of Testing Dataset

Classifiers	True Positive Rate of Testing
RandomForest	0.921
Multilayerperceptron	0.845
Decision Table	0.876
Naivebayes	0.704
J48	0.918

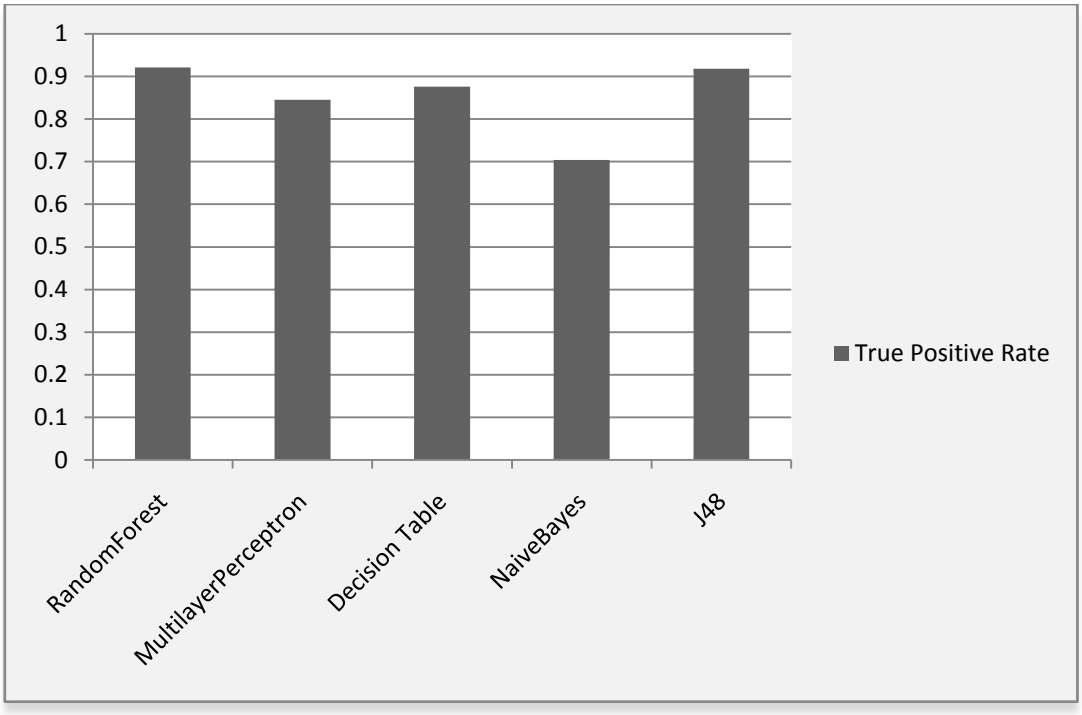


Figure 5.2 True Positive Rates of Testing Dataset

5.1.3 False Positive Rates of Testing Dataset

Table 5.3 shows false positive rates of Testing dataset. Naïve Bayes classifier has higher false positive rate whereas RandomForest has least false positive rate. False positive rate of Decision Table is 0.035; J48 and Multilayerperceptron have 0.025 and 0.028 false positive rate respectively. Figure 5.3 shows false positive rates of Testing dataset.

Table 5.3 False Positive Rates of Testing Dataset

Classifiers	False Positive Rate of Testing
RandomForest	0.023
Multilayerperceptron	0.028
Decision Table	0.035
Naivebayes	0.055
J48	0.025

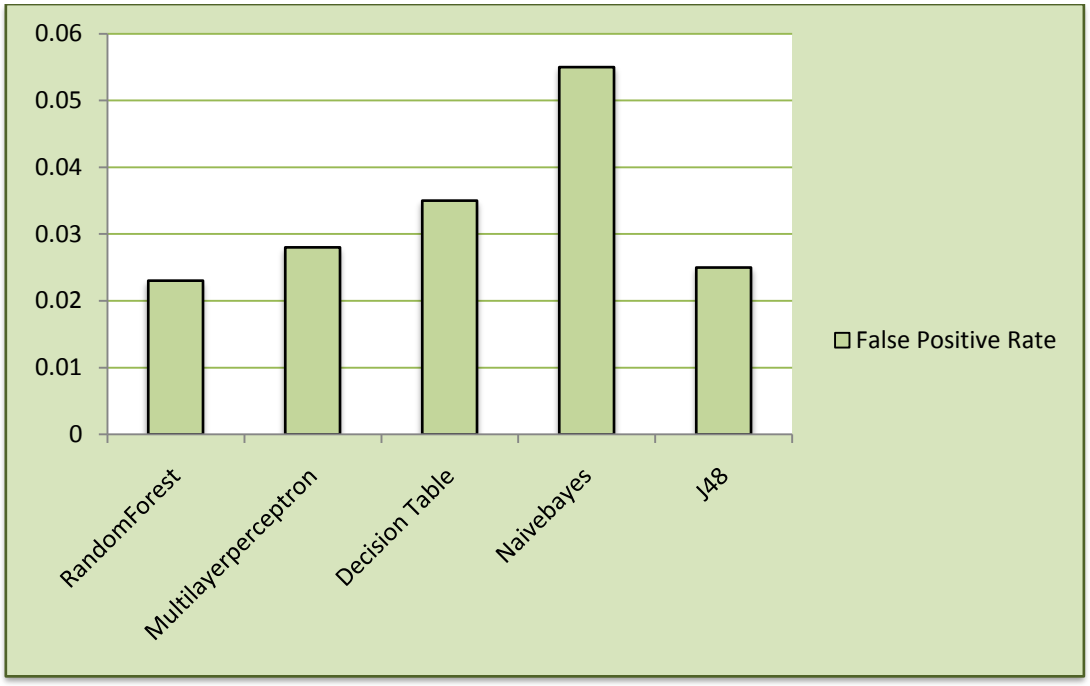


Figure 5.3 False Positive Rates of Testing Dataset

5.2 Analysis of Testing^{+} Dataset

In this section, we analyze the accuracy rate, true positive rate and false positive rates of dataset termed as Testing^{+}. From this analysis, we concluded that our approach has higher detection rates and better performance than the testing data set.

5.2.1 Accuracy Rate of Testing^{+} Dataset

Table 5.4 shows that our approach gives better accuracy rate. RandomForest, J48, Decision Table have higher accuracy rate. Accuracy rate of Naïve Bayes is least. Figure 5.4 shows accuracy rates of Testing^{+} dataset.

Table 5.4 Accuracy rate of Testing^{+} Dataset

Classifiers	Accuracy of Testing+ Dataset
RandomForest	97.4946
J48	97.3515
Decision Table	97.0651
Multilayerperceptron	94.8461
Naivebayes	86.8289

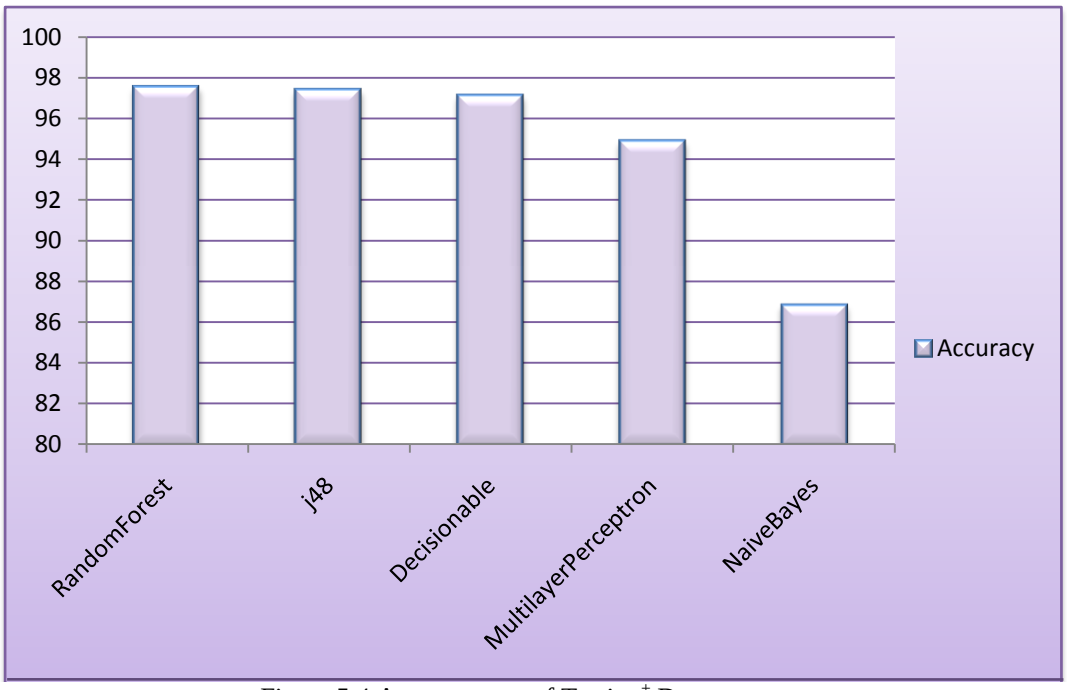


Figure 5.4 Accuracy rate of Testing^{+} Dataset

5.2.2 True Positive Rates of Testing^{+} Dataset

Table 5.5 shows true positive rate of Testing^{+} data. True positive rates of RandomForest, J48 and MultilayerPerceptron are higher i.e. 0.948, 0.942, 0.925 respectively. Decision table classifier has highest rate of true positive i.e. 0.955 whereas NaiveBayes has least true positive rate i.e. 0.610. Figure 5.5 shows true positive rates of Testing^{+} dataset.

Table 5.5 True Positive Rates of Testing^{+} Dataset

Classifiers	True Positive Rate of Testing^{+}
RandomForest	0.948
MultilayerPerceptron	0.925
Decision Table	0.955
NaiveBayes	0.610
J48	0.942

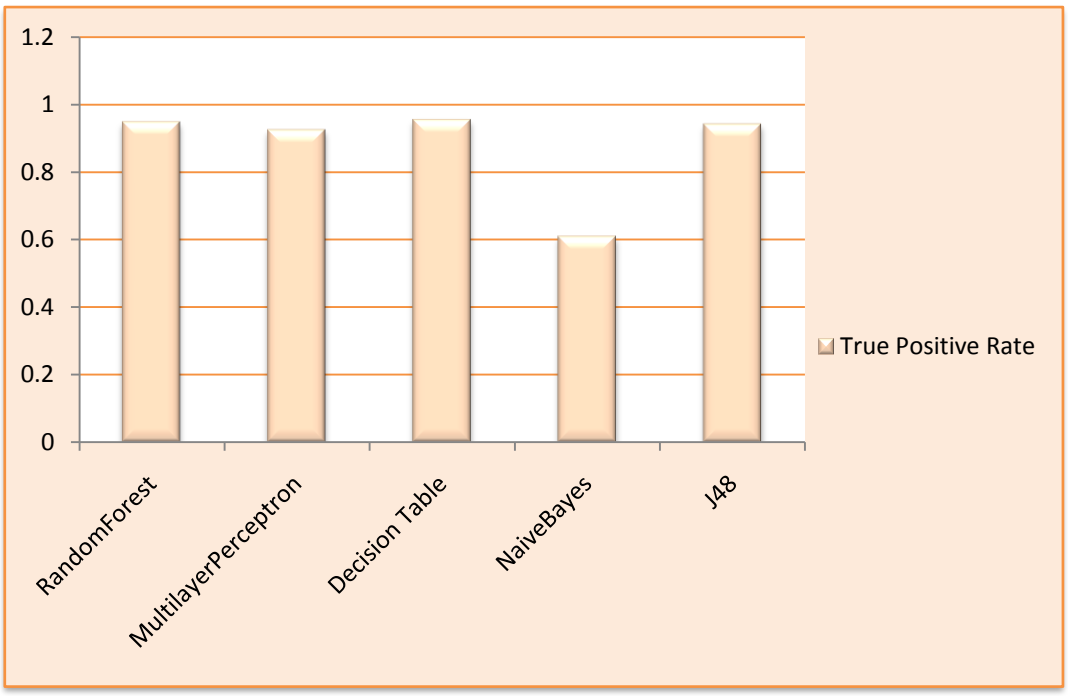


Figure 5.5 True Positive Rates of Testing^{+} Dataset

5.2.3 False Positive Rates of Testing^{+} Dataset

Table 5.6 shows false positive rates of Testing^{+} dataset. NaiveBayes classifier has higher false positive rate i.e. 0.033whereas Decision Table has least false positive rate i.e. 0.009. False positive rate of MultilayerPerceptron is 0.028, J48 and RandomForest have almost same false positive rate i.e. 0.011. Figure 5.6 shows false positive rates of Testing^{+} dataset.

Table 5.6 False Positive Rates of Testing^{+} Dataset

Classifiers	False Positive Rate of Testing^{+}
RandomForest	0.010
MultilayerPerceptron	0.028
Decision Table	0.009
NaiveBayes	0.033
J48	0.011

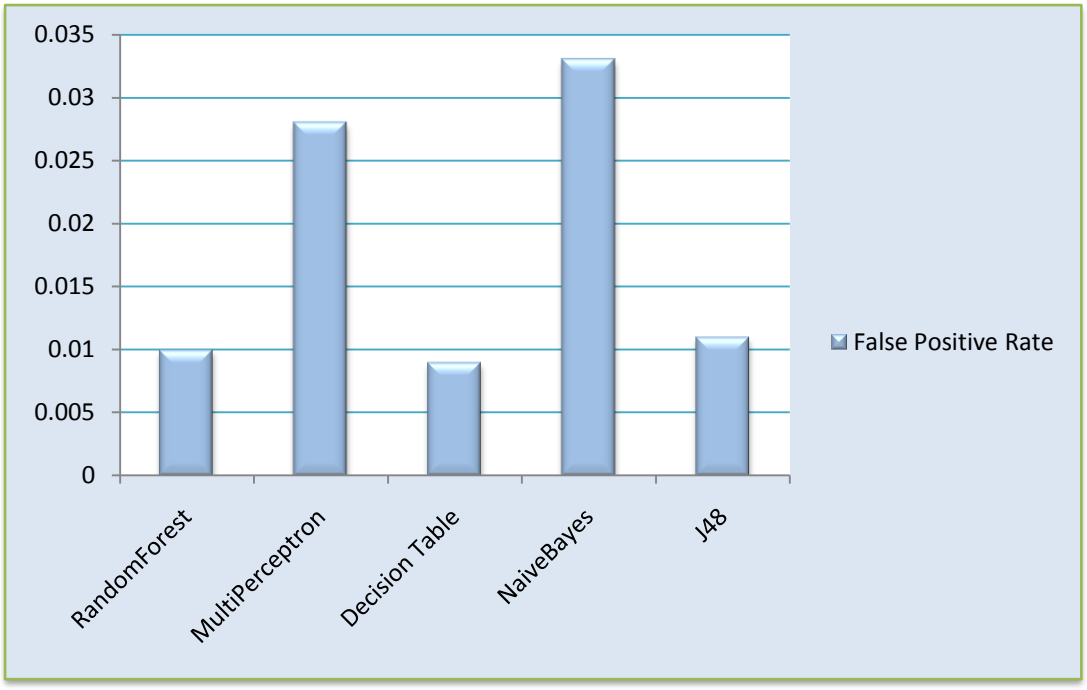


Figure 5.6 False Positive Rates of Testing^{+} Dataset

5.3 Evaluation of Proposed Method

In this approach, we have evaluated the results that are obtained from both the data sets i.e. Testing and Testing+. From the experiment we have concluded that outcome of Testing^{+} gives better results as compared to Testing. Moreover performance of Testing^{+} is better than Testing.

5.3.1 Comparison of accuracy rates of Testing and Testing^{+}

Five classifiers namely RandomForest, J48, NaiveBayes, Decision Table and Multilayerperceptron (MLP) are used to perform this experiment. These five classifiers are applied to both Testing and Testing^{+}. It is concluded from the results that detection rates of Testing^{+} are higher that Testing. Out of these five classifiers, RandomForest classifier achieves better accuracy rate. Accuracy rate is defined by number of correct assessments to the number of all assessments. Mathematical formula of accuracy is given below:

Accuracy = (True Negatives + True Positives) / (True Negative + True Positive +False Negative +False Positive)

Table 5.7 shows a comparison of Testing dataset and Testing^{+} dataset. From this comparison we concluded that accuracy rate of our approach i.e. Testing^{+} is effectively higher that Testing dataset. Results from our approach shows that accuracy rate of RandomForest, J48 and Decision table is approximately same i.e. 97.4946%, 97.3515%, 97.0651% which is comparatively higher than the accuracy rate of Testing dataset. Naive Bayes classifier has less accuracy rate in both the cases whereas RandomForest classifier achieves better accuracy rate in both Testing and Testing^{+} dataset. Figure 5.7 shows comparison of accuracy rates of Testing and Testing^{+}.

Table 5.7 Comparison of Accuracy Rates of Testing and Testing^{+}

Classifiers	Testing	Testing^{+}
RandomForest	92.3352	97.4946
J48	91.6905	97.3515
Decision Table	88.9685	97.0651
Multilayerperceptron	88.467	94.8461
Naivebayes	77.7937	86.8289

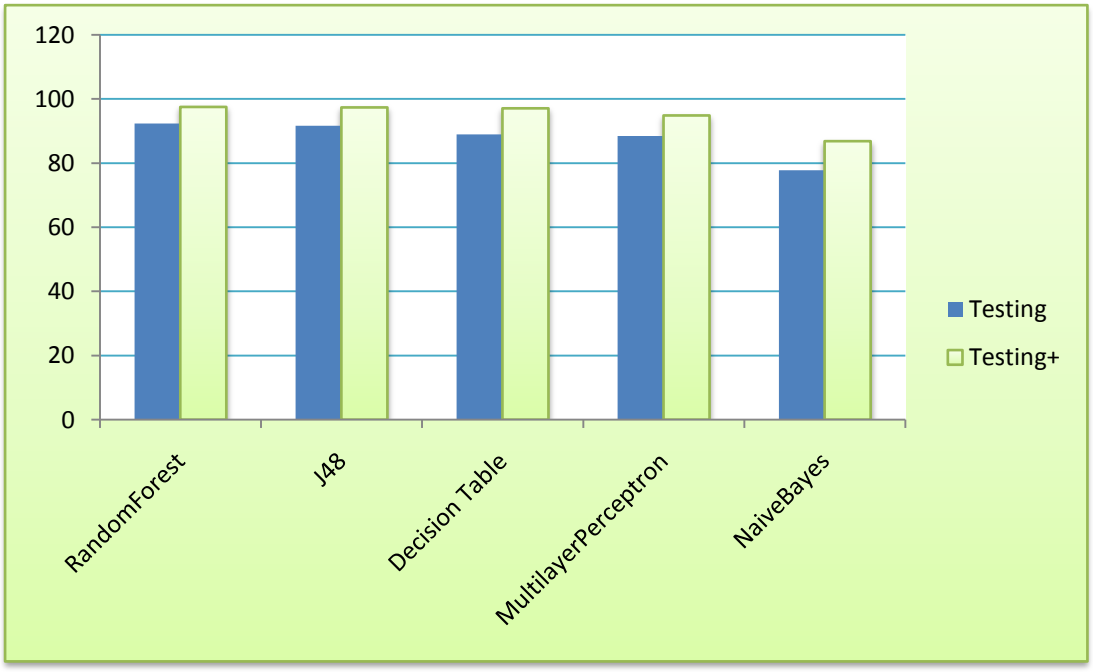


Figure 5.7 Comparison of Accuracy Rates of Testing and Testing^{+}

5.3.2 Comparison of Time Taken by Testing and Testing+

Table 5.8 shows time taken by our approach is comparatively very less than time taken by Testing dataset to evaluate the results. In this way, Time complexity of our approach is very low. Testing dataset has resulted into better performance by taking less time to produce the outcome. Figure 5.8 shows the comparison of time taken by both the datasets.

Table 5.8 Comparison of time taken by Testing and Testing+

Classifiers	Time taken by Testing(in sec)	Time taken by Testing+in sec)
RandomForest	1.31	0.52
Multilayerperceptron	10.62	5.51
Decision Table	0.31	0.08
Naivebayes	0.05	0.02
J48	0.25	0.03

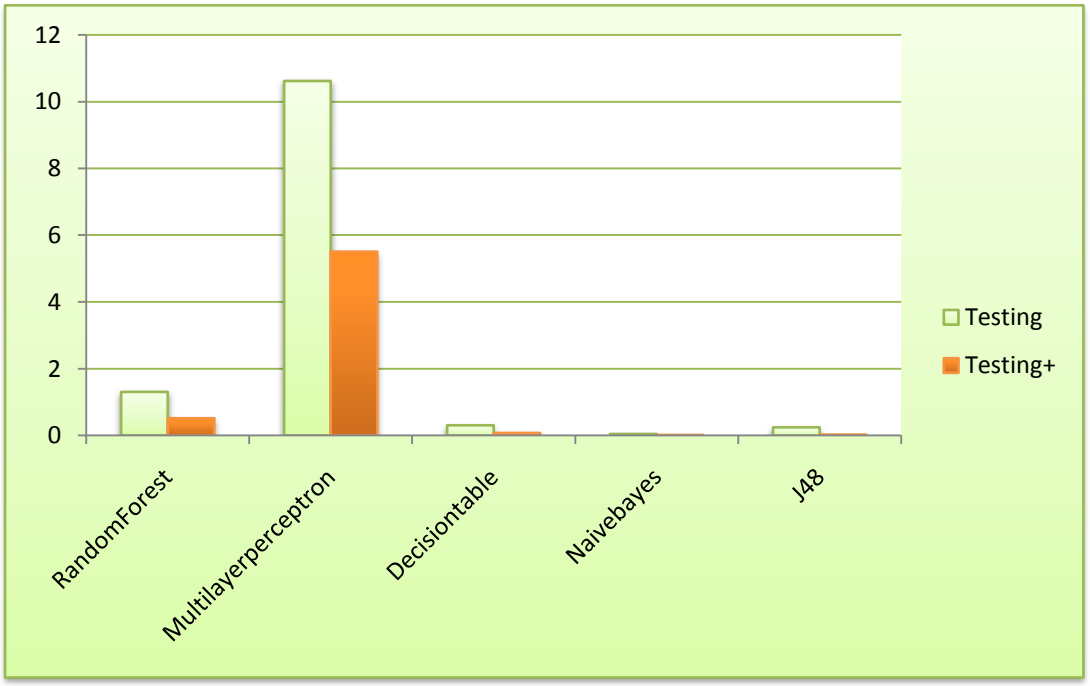


Figure 5.8 Comparison of Time Taken by Testing and Testing+

5.3.3 Comparison of True Positive Rates of Testing and Testing+

Table 5.9 shows comparison between true positive rates of both Testing and Testing+ dataset. We can analyze from Table 5.9 that all the classifiers have higher rate of true positives of our approach except Naive Bayes. Naive Bayes Classifier of Testing data set has high rate of true positive than Testing+ dataset. The term true positive is defined as an attack that directs system to raise an alarm. In this way, our approach is very effective in detecting attacks in network and immediately raises alarm. Figure 5.9 shows comparison of true positive rates of Testing and Testing+ datasets.

Table 5.9 Comparison of True Positive Rates of Testing and Testing+

Classifiers	True Positive(TP) Rate of Testing	True Positive(TP) Rate of Testing+
RandomForest	0.921	0.948
Multilayerperceptron	0.845	0.925
Decision Table	0.876	0.955
Naivebayes	0.704	0.610
J48	0.918	0.942

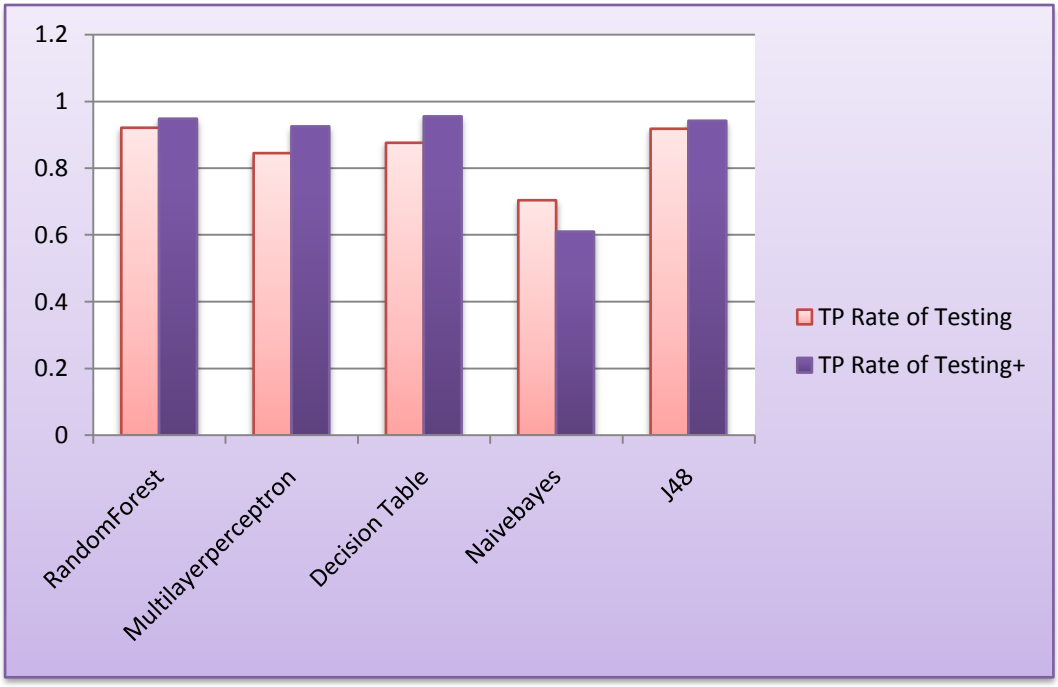


Figure 5.9 Comparison of True Positive Rates of Testing and Testing+

5.3.4 Comparison of False Positive Rates of Testing and Testing+

Table 5.10 shows comparison between false positive rates of both Testing and Testing+ dataset. We can analyze from Table 5.10 that all the classifiers have lesser rate of false positives of our approach. The term false positive is defined as an event that gives signal to system to raise an alarm when there is no attack. So our approach resulted into lower rate of false positives. Figure 5.10 shows comparison of both the datasets.

Table 5.10 Comparison of False Positive of Testing and Testing+

Classifiers	False Positive Rate of Testing	False Positive Rate of Testing+
RandomForest	0.023	0.010
Multilayerperceptron	0.028	0.028
Decision Table	0.035	0.009
Naivebayes	0.055	0.033
J48	0.025	0.011

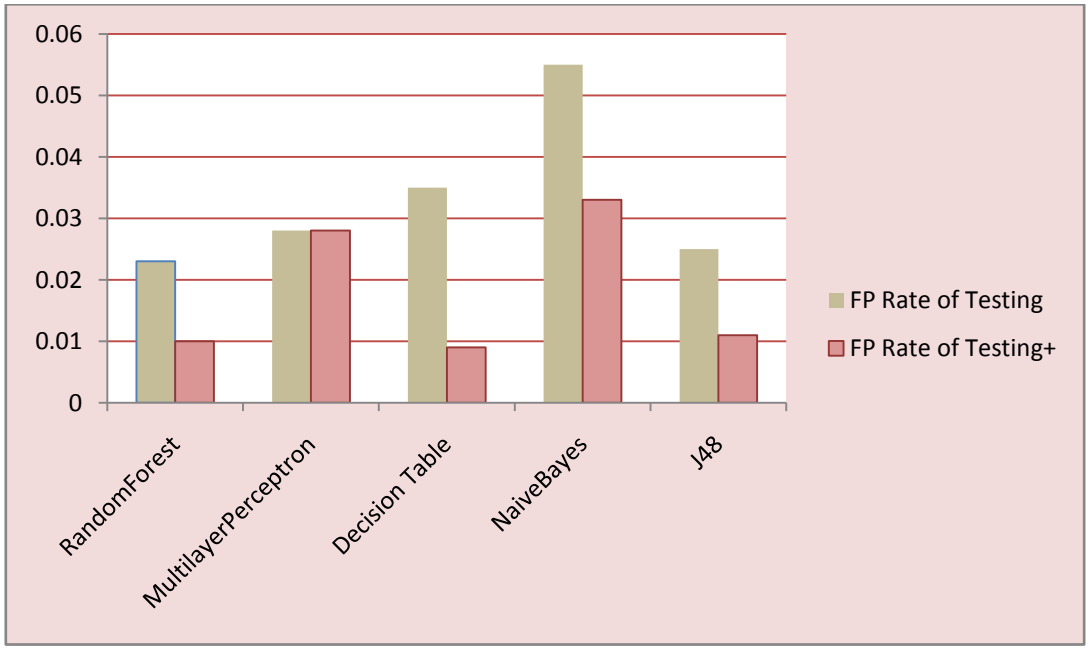


Figure 5.10 Comparison of False Positive Rates of Testing and Testing+

5.4 Comparison of proposed method with existing methods

Various researchers are working on improving accuracy of machine learning classifiers. Figure 5.11 shows the accuracy of techniques used by various researchers and proposed approach. Results of their work and proposed approach are compared. Below graph shows the proposed approach gives higher accuracy than others.

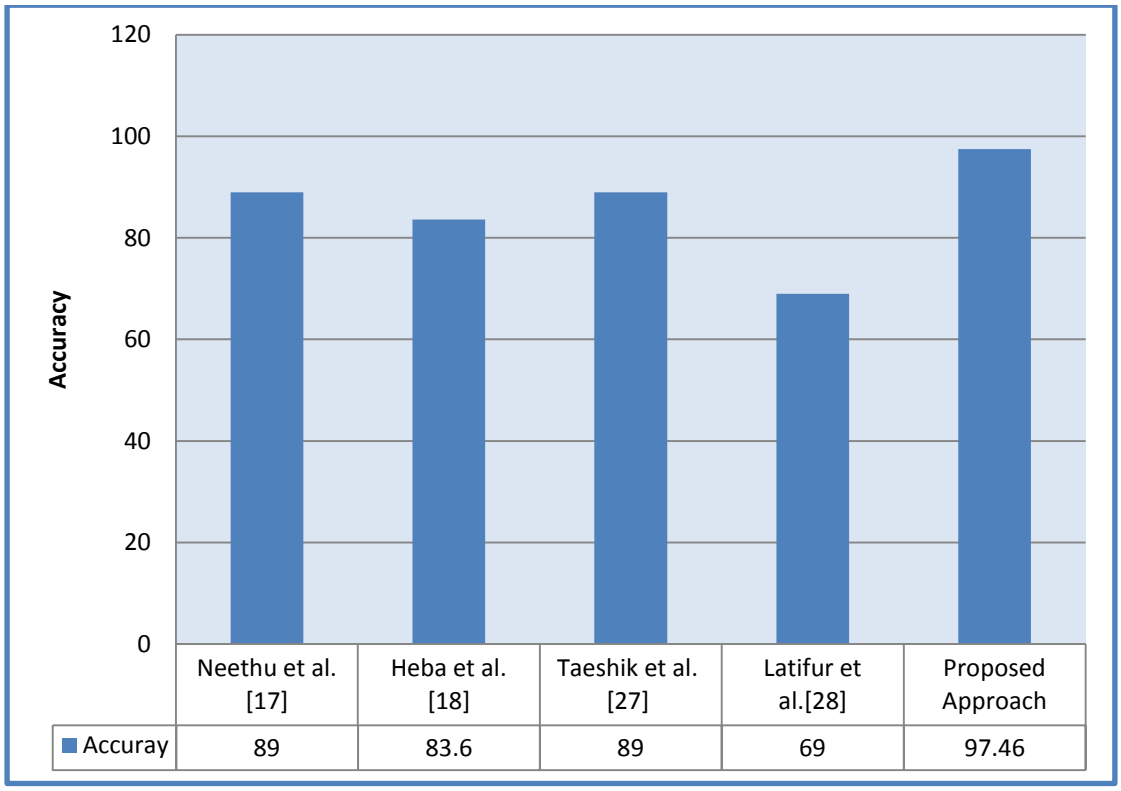


Figure 5.11 Comparison of proposed method with existing methods

Chapter Summary: In this chapter Testing and Testing^{+} dataset is analyzed. We have compared the accuracy rates, true positive rates and false positive rates of Testing and Testing^{+} dataset.

Chapter 6

Conclusion and Future Scope

The research carried out for the thesis mainly emphasize on the need to protect the information systems from various kinds of threats .The increase in the use of Internet and complex architecture of the networks has made the systems vulnerable to intrusions . The networks can be secured by deploying security solutions like Antivirus, Firewall, Honeypot and IDS. The objective of the thesis is to study Machine learning techniques and misuse detection techniques for detecting intrusions. Our proposed system is combination of two techniques misuse detection system and Machine learning techniques. Firstly, we applied subset of KDD dataset to machine learning classifiers directly. Secondly, we use misuse detection system to match training data with testing data. Results from this system are then applied to machine learning classifiers which are Naive Bayes, decision Table, J48, Multiperceptron and RandomForest. This experimentation is carried out by selecting three attributes based on their info gain value. These three attributes are src_bytes, service, and protocol_type. The results of misuse detection are compared with misuse detection + machine learning technique. It is concluded from the results that accuracy rates obtained from with misuse detection + machine learning techniques are better. Moreover, processing time is also very less. True positive rates are also enhanced and true false rates are also reduced by our approach. RandomForest, have higher accuracy rate i.e. 97.4946. Decision table classifier has highest rate of true positive and least rate of false positive i.e. 0.955 and 0.009 respectively. The goal of the thesis is to combine misuse detection with machine learning. From the experimental results, it is evident that the combined misuse detection + machine learning approach are better than individual misuse detection approach.

This combination of misuse detection and machine learning can be improved by selecting attributes more than three. Accuracy rates and true positive rates can be enhanced by using different types of classifiers.

References

[1] Oh S. H. and Lee W.S., “An Anomaly Intrusion Detection Method by Clustering Normal User Behavior”, Computer and Security, vol. 22, no.7, pp. 596-612, October 2003.

[2] Yue X., Chen W. and Wang Y., “The Research of Firewall Technology in Computer Network Security,” in Proceedings of Conference on Computational Intelligence and Industrial Applications, pp. 421-424, 2009.

[3] W. Stallings, Cryptography and Network security: Security Services, 4th Ed., Delhi: Prentice Hall, 2007.

[4] W. Stallings, Cryptography and Network Security Principles and Practices. 4th ed., Prentice Hall, 2005.

[5] V. Visoottiviseth, U. Jaralrungroj, E. Phoomrungraungsuk, P. Kultanon, “Distributed Honeypot log management and visualization of attacker geographical distribution,” in Proc. of Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE), May 2011, pp. 23-28.

[6] Fortinet, Inc. “Improving Network Protection and Performance with Network-Based Antivirus Technology,” White paper, Oct. 2002.

[7] A. S. Ashoor and S. Gore, “Importance of Intrusion Detection system (IDS)”. International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp.1-4, Jan-2011.

[8] D. Rozenblum, “Understanding Intrusion Detection System,”www.sans.org/reading_room/whitepapers/detection/understanding-intrusiondetection-systems_337, October 31, 2003.

[9] Lazarevic et al., "Intrusion detection: A survey," Managing Cyber Threats, vol. 5, pp.19-78, 2005

[10] W. Lee, Salvatore J. Stolfo, and Kui W. Mok, "A Data Mining Framework for Adaptive Intrusion Detection," in Proc. of the IEEE Symposium on Security and Privacy,1999, pp.I20-132.

[11] Desai, Aaditya, and Sunil Rai. "Analysis of Machine Learning Algorithms using WEKA." In International Conference & Workshop on Recent Trends in Technology,(TCET). 2012.

[12] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali-A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set." In Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009. 2009.

[13] Arjunwadkar Narayan, M., and Thaksen J. Parvat. "An Intrusion Detection System,(IDS) with Machine Learning (ML) Model Combining Hybrid Classifiers." connections 1: 3.

[14] Majeed, Khurram, Yanguo Jing, Dusica Novakovic, and Karim Ouazzane. "Behaviour Based Anomaly Detection for Smartphones Using Machine Learning Algorithm."

[15] Bhavsar, Yogita B., and Kalyani C. Waghmare. "Intrusion Detection System Using Data Mining Technique: Support Vector Machine." International Journal of Emerging Technology and Advanced Engineering 3, no. 3 (2013): 581-586.

[16] Khari, Manju, and Anjali Karar. "Analysis on Intrusion Detection By Machine Learning Techniques: A Review." International Journal of Advanced Research in Computer Science & Software Engineering 3, no. 4 (2013).

[17] Neethu, B. "Classification of intrusion detection dataset using machine learning approaches." International Journal of Electronics and Computer Science Engineering (2012): 1044-1051.

[18] Ibrahim, Heba Ezzat, Sherif M. Badr, and Mohamed A. Shaheen. "Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems." arXiv preprint arXiv:1210.7650 (2012).

[19] George, Annie. "Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM." International Journal of Computer Applications (0975–8887) Volume (2012).

[20] Fares, Ahmed R., Mohamed I. Sharawy, and Hala H. Zayed. "Intrusion Detection." Journal of Computing Science and Engineering 5, no. 4 (2011): 305-313.

[21] Kumar, Gireesh. "Network intrusion detection system based on machine learning algorithms." (2010).

[22] Sarvari, Hadi, and Mohamad Mehdi Keikha. "Improving the accuracy of intrusion detection systems by using the combination of machine learning approaches." In Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of, pp. 334-337. IEEE, 2010.

[23] Subbulakshmi, T., George Mathew, and S. Mercy Shalinie. "Real time classification and clustering of ids alerts using machine learning algorithms." International journal of Artificial & Application 1, no. 1 (2010): 20.

[24] Tran, Tich Phuoc, Pohsiang Tsai, Tony Jan, and Xiaoying Kong. Network Intrusion Detection using Machine Learning and Voting techniques. INTECH Open Access Publisher, 2010.

[25] Panda, Mrutyunjaya, and Manas Ranjan Patra. "Ensemble of classifiers for detecting network intrusion." In Proceedings of the International Conference on Advances in Computing, Communication and Control, pp. 510-515. ACM, 2009.

[26] Laskov, Pavel, Konrad Rieck, and Klaus-Robert Müller. "Machine learning for intrusion detection." Mining Massive Data Sets for Security (2007): 366-373.

[27] Shon, Taeshik, and Jongsub Moon. "A hybrid machine learning approach to network anomaly detection." Information Sciences 177, no. 18 (2007): 3799-3821.

[28] Khan, Latifur, Mamoun Awad, and Bhavani Thuraisingham. "A new intrusion detection system using support vector machines and hierarchical clustering." The VLDB Journal—The International Journal on Very Large Data Bases 16, no. 4 (2007): 507-521.

[29] Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung. "Intrusion detection using neural networks and support vector machines." In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, vol. 2, pp. 1702-1707. IEEE, 2002.

[30] Sinclair, Chris, Lyn Pierce, and Sara Matzner. "An application of machine learning to network intrusion detection." In Computer Security Applications Conference, 1999.(ACSAC'99) Proceedings. 15th Annual, pp. 371-377. IEEE, 1999.

[31] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6, no. 2 (2013): 256-261.

[32] Banerji, Geetali, and Kanak Saxena. "An Efficient Classification Algorithm for Real Estate domain."

[33] Livingston, Frederick. "Implementation of Breiman's random forest machine learning algorithm." ECE591Q Machine Learning Journal Paper (2005).

[34] Kaur, Gaganjot, and Amit Chhabra. "Improved J48 Classification Algorithm for the Prediction of Diabetes." International Journal of Computer Applications 98, no. 22 (2014).

[35] Naoum, Reyadh Shaker, Namh Abdula Abid, and Zainab Namh Al-Sultani. "An Enhanced Resilient Backpropagation Artificial Neural Network for Intrusion Detection System." International Journal of Computer Science and Network Security 12, no. 3 (2012): 11-16.

[36] Costantini, Stefania. "Towards active logic programming." arXiv preprint arXiv:1403.5508 (2014).

[37] Kwok, Suk Wah, and Chris Carter. "Multiple decision trees." arXiv preprint arXiv:1304.2363 (2013).

[38] Aher, Sunita B., and L. M. R. J. Lobo. "A comparative study of association rule algorithms for course recommender system in e-learning." International Journal of Computer Applications 39, no. 1 (2012).

[39] Aher, Sunita B., and L. M. R. J. Lobo. "A comparative study of association rule algorithms for course recommender system in e-learning." International Journal of Computer Applications 39, no. 1 (2012).

[40] Turgut, Damla, Sajal K. Das, Ramez Elmasri, and Begumhan Turgut. "Optimizing clustering algorithm in mobile ad hoc networks using genetic algorithmic approach." In Global Telecommunications Conference, 2002. *GLOBECOM'02. IEEE*, vol. 1, pp. 62-66. IEEE, 2002.

[41] Sebyala, Abdallah Abbey, TemitopeOlukemi, Lionel Sacks, and Dr Lionel Sacks. "Active platform security through intrusion detection using naive bayesian network for anomaly detection." In London Communications Symposium. 2002.

[42] Gong, RenHui, Mohammad Zulkernine, and Purang Abolmaesumi. "A software implementation of a genetic algorithm based approach to network intrusion detection." In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on, pp. 246-253. IEEE, 2005.

[43] Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques."Artificial Intelligence Review 26, no. 3 (2006): 159-190.

[44] Damopoulos, Dimitrios, Sofia A. Menesidou, Georgios Kambourakis, Maria Papadaki, Nathan Clarke, and Stefanos Gritzalis. "Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers." Security and Communication Networks 5, no. 1 (2012): 3-14.

Research Publications

- Rohini Rajpal, Ramandeep Kaur and Sanmeet Kaur, “Improving Detection Rate using Misuse Detection and Machine Learning”, IEEE Security & Privacy. [Status-Communicated]