

**HYBRID INTELLIGENT APPROACH FOR NETWORK
INTRUSION DETECTION**

Wael Hasan Ali Al-Mohammed

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

SCHOOL OF COMPUTING

COLLEGE OF ARTS AND SCIENCES

UNIVERSITY UTARA MALAYSIA

2015

PERMISSION OF USE

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

ABSTRAK

Sejak kebelakangan ini, rangkaian komputer telah meluas dan sangat rumit. Banyak informasi sensitif disalurkan kepada pelbagai jenis peranti komputer, dari komputer mini ke pelayan dan juga dari komputer mini ke peranti mudah alih. Perubahan ini menyebabkan serangan ke atas maklumat penting ke atas sistem rangkaian semakin bertambah setiap tahun. Pencerobohan adalah ancaman utama terhadap rangkaian. Ia ditakrifkan sebagai satu siri aktiviti yang bertujuan untuk menjejaskan keselamatan sistem rangkaian dari segi kerahsiaan, integriti dan ketersediaan. Oleh itu, pengesanan pencerobohan adalah sangat penting sebagai sebahagian daripada pertahanan. Oleh itu, ia perlu meningkatkan teknik pengesanan pencerobohan rangkaian dan sistem. Disebabkan pendekatan pengesanan pencerobohan sebelum ini terlalu terhad, kami mencadangkan pendekatan hibrid pintar untuk mengesan pencerobohan rangkaian berdasarkan pengelompokan k-means algoritma dan mesin sokongan vektor algoritma. Tujuan penyelidikan adalah untuk mengurangkan kadar penggera palsu dan juga untuk meningkatkan kadar pengesanan untuk dibandingkan dengan pendekatan pengesanan pencerobohan yang sedia ada. Pencerobohan dataset NSL-KDD telah digunakan untuk latihan dan menguji pendekatan yang dicadangkan. Beberapa langkah telah dilakukan sebelum tujuan pengelasan untuk meningkatkan prestasi pengelasan. Pertama, menyatukan jenis dan menapis dataset melalui data transformasi. Kemudian, pemilihan ciri algoritma telah digunakan untuk membuang ciri yang tidak relevan dalam tujuan pencerobohan. Ciri-ciri yang terpilih telah mengurangkan 41 ciri kepada 21 ciri untuk mengesan pencerobohan dan kemudian kaedah-kaedah yang biasa digunakan untuk dilaksanakan serta mengurangkan perbezaan di antara maklumat. Pengelompokan adalah langkah terakhir pemprosesan sebelum pengelasan dijalankan menggunakan k-means algoritma. Klasifikasi telah dilakukan dengan menggunakan mesin sokongan vektor. Selepas latihan dan menguji pendekatan pintar hibrid yang telah dicadangkan, keputusan penilaian prestasi telah menunjukkan bahawa ia mencapai ketepatan yang tinggi dan kadar pengesanan palsu yang rendah. Ketepatannya ialah 96.025% dan penggera palsu adalah 3.715%.

Kata Kunci: Rangkaian Pengesanan Pencerobohan, Pendekatan Pintar Hibrid, Rangkaian Serangan, Pengelompokan, Klasifikasi, Pencerobohan dataset NSL-KDD, K-Means algoritma, Mesin Sokongan Vektor algoritma.

ABSTRACT

In recent years, computer networks are broadly used, and they have become very complicated. A lot of sensitive information passes through various kinds of computer devices, ranging from minicomputers to servers and mobile devices. These occurring changes have led to draw the conclusion that the number of attacks on important information over the network systems is increasing with every year. Intrusion is the main threat to the network. It is defined as a series of activities aimed for exposing the security of network systems in terms of confidentiality, integrity and availability, as a result; intrusion detection is extremely important as a part of the defense. Hence, there must be substantial improvement in network intrusion detection techniques and systems. Due to the prevailing limitations of finding novel attacks, high false detection, and accuracy in previous intrusion detection approaches, this study has proposed a hybrid intelligent approach for network intrusion detection based on k-means clustering algorithm and support vector machine classification algorithm. The aim of this study is to reduce the rate of false alarm and also to improve the detection rate, comparing with the existing intrusion detection approaches. In the present study, NSL-KDD intrusion dataset has been used for training and testing the proposed approach. In order to improve classification performance, some steps have been taken beforehand. The first one is about unifying the types and filtering the dataset by data transformation. Then, a features selection algorithm is applied to remove irrelevant and noisy features for the purpose of intrusion. Feature selection has decreased the features from 41 to 21 features for intrusion detection and later normalization method is employed to perform and reduce the differences among the data. Clustering is the last step of processing before classification has been performed, using k-means algorithm. Under the purpose of classification, support vector machine have been used. After training and testing the proposed hybrid intelligent approach, the results of performance evaluation have shown that the proposed network intrusion detection has achieved high accuracy and low false detection rate. The accuracy is 96.025 percent and the false alarm is 3.715 percent.

Keywords: Network Intrusion Detection, Hybrid Intelligent Approach, Network Attacks, Clustering, Classification, NSL-KDD intrusion dataset, K-Means algorithm, Support Vector Machine algorithm.

DEDICATION

Every challenging work needs self-efforts as well as guidance and support of others, especially those who are very close to our heart.

Therefore, I dedicate this humble work

To my sweet and beloved

FAMILY

Whose love, support, and pray of day and night for making me able to reach such success.

To my lovely homeland

IRAQ

Which opening my eyes to this world. I hope it will get the peace soon.

To the marvelous land

MALAYSIA

Which granted me the opportunity to complete my study.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Merciful, the Most Compassionate all praise be to Allah, the Lord of the worlds; and prayers and peace is being upon Mohamed His servant and messenger.

First and foremost, I acknowledge my unlimited thanks to Allah Almighty, the Ever-Magnificent and the Ever-Thankful for his help and blessings for which this thesis would not have been possible to achievement without his help. I would like to thank and appreciate our first teacher, prophet Mohammed, and his family who taught us the struggling, and patience to achieve the success.

To finally accomplish this journey which began in February, 6, 2013 (06:00 AM BGD it is the time when I left my home, heading to Malaysia), I have to thank many persons who deserve my gratitude. I was guided, supported and encouraged by them.

I convey a profound appreciation to my supervisor, Associate Professor Dr. Hatim Mohammad Tahir, for his guidance, advice, assistance, and oversight. I have been very fortunate to have been able to work with him since undertaking my master degree. I thank the kind dissertation's examiners Dr. Mohd Nizam Omer and Dr. Nur Haryani Zakaria for their comments and suggestions. I extend my appreciation to the head of department, coordinators and all staff of the school of computing.

My deepest and heartfelt gratitude, loves, thanks and appreciation for my dearest parents and my beloved siblings who are a part of my happiness, success, and the inspiration that led me for the quest for knowledge and self-empowerment through night and day. I hope I can put a smile on their faces for giving back their tremendous support and encouragement, patience, unconditional love, and prayers for me. Thank you for giving me the strength to chase and reach my dreams.

I owe a huge debt of gratitude and thanks to my close friend Dr. Hayder Mohammed Ali, The person who had a main role in my master's candidature. I am forever grateful for him. I wish all the best for him and I am praying to Allah Almighty to ease his life, especially his PhD journey.

I would like to express my wholehearted appreciation to my soulmates, closest and best friends (Hussein Abdulkhaliq – Mohammed Zuhair – Ahmed Shakir – Tammar Hayder – Hayder Kurdi – Rasoul Faik) for their gorgeous support and encouragement along the way in countless ways. I want to exploit this opportunity to thank them for our pure and wonderful friendship over twelve years.

I extend my appreciation to Iraqi friends who met them in UUM (especially Abbass, Maitham, Mohammed, Samer, Wadhah, and Zaid), all my friends in my country and UUM, bachelor's lecturers, bachelor's colleagues, master's lecturers, master's colleagues, UUM staff, the people who are praying, supported, helped, guided and wished the best for me, Malaysian people who are very gentle with me.

Thank You All.

“This Thesis is only the beginning of my journey.”

Wael Hasan Ali Al-Zuwainy

Northern University of Malaysia, Kedah, Malaysia

Monday, October 20, 2014

TABLE OF CONTENTS

PERMISSION OF USE	i
ABSTRAK.....	ii
ABSTRACT.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS.....	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
LIST OF ABBREVIATIONS.....	xiii
CHAPTER ONE : INTRODUCTION	1
1.1 Introduction	1
1.2 Background of study	1
1.3 Problem Statement	8
1.4 Research Questions	9
1.5 Research Objectives	10
1.6 Significance of research	11
1.7 Contributions of Research.....	12
1.8 Scope of Research	12
1.9 Thesis Organization.....	12
1.10 Summary	13
CHAPTER TWO : LITERATURE REVIEW.....	14
2.1 Introduction	14
2.2 Network Security Overview.....	14

2.3	Network Intrusion Detection	17
2.4	Network Attacks.....	19
2.4.1	Probing (Probe).....	19
2.4.2	Denial of Service (DoS).....	19
2.4.3	Remote to Local (R2L)	20
2.4.4	User to Root (U2R).....	20
2.5	Intrusion Detection System	20
2.5.1	Intrusion Detection System Sites	25
2.5.1.1	Host Based Intrusion Detection System.....	25
2.5.1.2	Network Based Intrusion Detection System	27
2.5.1.3	Hybrid Intrusion Detection System.....	29
2.5.2	Intrusion Detection System Behaviors.....	31
2.5.2.1	Passive Behavior	31
2.5.2.2	Active Behavior.....	32
2.5.3	Intrusion Detection System Approaches.....	33
2.5.3.1	Misuse Detection Approach	33
2.5.3.2	Anomaly Detection Approach.....	35
2.5.3.3	Hybrid Intrusion Detection Approach.....	39
2.6	Artificial intelligence for Intrusion Detection	39
2.6.1	Artificial Immune Systems (AIS)	40
2.6.2	Artificial Neural Networks (ANN)	40
2.6.3	Fuzzy Logic (FL)	41
2.6.4	Genetic Algorithm (GA).....	42
2.6.5	Support Vector Machine (SVM).....	43
2.6.6	Hidden Markov Models	43

2.6.7	Naïve Bayes	44
2.6.8	Data Mining	44
2.6.9	Hybrid Artificial Intelligence Approach	45
2.7	Performance Evaluation	46
2.7.1	Intrusion Detection Dataset.....	46
2.7.2	Evaluation Metric.....	46
2.8	Existing hybrid intelligent approaches	48
2.9	Summary	54
CHAPTER THREE : RESEARCH METHODOLOGY		55
3.1	Introduction	55
3.2	Phase I: Selection of Experiment Dataset	56
3.3	Phase II: Data Pre-Processing	61
3.4	Phase III: Classification	69
3.5	Phase VI: Performance Evaluation	71
3.5.1	Confusion Matrix	72
3.6	Summary	74
CHAPTER FOUR : HYBRID INTELLIGENT APPROACH DESIGN		75
4.1	Introduction	75
4.2	Approach Design.....	75
4.3	Clustering	77
	K-Means Clustering.....	79
4.4	Classification	80
	Support Vector Machine.....	81
4.5	Summary	82
CHAPTER FIVE : EXPERIMENTAL RESULTS AND EVALUATION		83

5.1	Introduction	83
5.2	Preprocessing Results.....	83
5.3	Classification Results	87
5.4	Performance Evaluation	90
5.5	Summary	93
CHAPTER SIX : CONCLUSION AND FUTUREWORK.....		94
6.1	Conclusion.....	94
6.2	Recommendation and Future work	98
REFERENCES		99

LIST OF FIGURES

Figure 2.1: A Generic Intrusion Detection System.....	21
Figure 2.2: Classification of Intrusion Detection Systems.....	24
Figure 2.3: Host Based Intrusion Detection System	26
Figure 2.4: Network Based Intrusion Detection System	28
Figure 2.5: Hybrid Based Intrusion Detection System.....	30
Figure 2.6: Passive Intrusion Detection System	31
Figure 2.7: Active Intrusion Detection System.....	32
Figure 2.8: Misuse Intrusion Detection System.....	34
Figure 2.9: Anomaly Intrusion Detection System	35
Figure 2.10: Classification of Anomaly Based Intrusion Detection Techniques.....	37
Figure 3.1: Research Methodology Phases	55
Figure 3.2: The Original NSL-KDD Dataset Connection	61
Figure 4.1: Workflow of Proposed Hybrid Intelligent Approach.....	76
Figure 5.1: The NSL-KDD Dataset Connection After Transformation.....	84
Figure 5.2: The NSL-KDD Dataset Connection After Normalization.....	86
Figure 5.3: Clustering Results of NSL-KDD Dataset.....	87
Figure 5.4: Detection Rate for Attack Categories.....	89
Figure 5.5: Comparison of Proposed Approach's Detection Rate with Others.....	93

LIST OF TABLES

Table 2.1: Network Based vs. Host Based Intrusion Detection System.....	29
Table 2.2: Misuse vs. Anomaly Intrusion Detection System	38
Table 2.3: Confusion Matrix.....	47
Table 2.4: Existing Hybrid Intelligent Approaches	50
Table 3.1: List of Attributes in NSL-KDD Dataset.....	58
Table 3.2: Attacks Categories.....	60
Table 3.3: Transformations Table	63
Table 3.4: Confusion Matrix.....	73
Table 5.1: The Result of Features Selection Process	85
Table 5.2: Confusion Matrix for Classification (number of connection records).....	88
Table 5.3: Confusion Matrix for Classification	89
Table 5.4: Result of Performance Evaluation	91
Table 5.5: Comparison Existing Approaches with the Proposed Hybrid Approach..	92

LIST OF ABBREVIATIONS

A	Accuracy
AI	Artificial Intelligent
ANN	Artificial Neural Network
DoS	Daniel of Services Attack
DR	Detection Rate
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
HIDS	Host-based Intrusion Detection System
IDS	Intrusion Detection System
NID	Network Intrusion Detection
NIDS	Network-based Intrusion Detection System
R2L	Remote to Local Attack
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
U2R	User to Root Attack

CHAPTER ONE

INTRODUCTION

1.1 Introduction

This chapter has discussed briefly the background of network security impacts, network intrusion problems and its solutions. On the other hand, it presents amply the statement of the problem in this study. This chapter defines the research questions, objectives of this study, the scope of research, research's significance and contributions of the study as well.

1.2 Background of study

In recent years, computer networks are broadly omnipresent and have become very complicated. Almost everybody with a computer or mobile device, is linked with the Internet in order to have access to data or send messages. A lot of sensitive information passes through various kinds of computer devices, ranging from minicomputers to servers and mobile devices (Elbasiony et al., 2013; Upadhyaya & Jain, 2013). In a wide scale, all governments, higher education organizations and different organizations depend on the network computer systems for the daily processes to perform, and network computer system play an essential role for the processes (Shanmugam & Idris, 2011).

As the utilization of the Internet and computers by the people worldwide has grown, more critical data are being saved and handled online. These changes have led to draw the conclusion that the number of attacks on important information over the network systems is increasing in every year. Hence, the network computer system security has become a significant issue to prevent dangerous threats on services and to protect sensitive data over the network (Jaisankar & Kannan, 2011; Wankhade et al., 2013).

Ensuring the confidentiality, integrity, and availability of information saved on computer network systems is an issue of growing concerns. Therefore, network systems security has become the main concern in the globe because of the attacks over the Internet. Network security must not be seen as a subset of network design, and network design must be taken into account at the last phase of design. Security has the purpose related to protecting assets. Network security deals with securing the transport of data. To achieve the security for connection, network is alienated into a number of security fields, and each security domain has similar security requirements (Chen et al., 2008; Eric, 2009).

In spite of having many types of security methods, like access control, encryption, and the use of firewalls, network security breaches are increasing daily (Elbasiony et al., 2013). Lately, the risk of attacking information over the system or Internet has also increased simultaneously. In order to prevent these attacks on information, there must be an appropriate solution for this problem (Jaisankar & Kannan, 2011).

Moreover, for the broad Internet usage and widespread of its application in all areas of everyone's life, the number of unauthorized access has increased from both types of attacks, external and internal intruders. Internal intruders are the indignant employees or the persons who are abusing their privileges to get personal and private gain. There are two broad types of threats in computer security; one of them is intrusion and the other one is a virus. In general, intrusion refers to either as hacker or a cracker. It is defined as a series of activities aimed at compromising the security of computer and network components in terms of confidentiality, integrity and availability. Intrusion can be done by an inside or outside agent to gain unauthorized entry and control of the security mechanism (Bhuyan et al., 2013; Chimphlee, 2007).

Hackers and attackers have done a lot of successful attempts to drop down web services and high-level organization's networks. As a consequences, the process of intrusion detection for computer network is becoming a critical and an emerging field, and it is one of the essential trend of the present research regarding the computers and network security (Jain et al., 2011; Panda et al., 2012).

Intrusion behavior can be categorized in to several attack types. Generally, there are four categories of attacks: firstly, probing where attackers begin gathering information about victim's system prior to initiating their attacks. Secondly, DoS: Denial of Service (DoS) which is the preventing of legal access for system source by consuming the bandwidth or overloading network computational resources, like memory and processor. Thirdly, User to Root (U2R): for this case of attacks, intruders have authorized profile on the

network system with specific privileges, and they try to use the network vulnerabilities for getting high privileges and root access; it is known as an internal attacks. And the last one is Remote to Local (R2L): On the contrary of previous category, an attacker does not has authorized profile on a victim computer system, therefore; he sends message for that remote machine through network system and exploits the vulnerabilities for earning access as a normal user on that device (Kulhare & Singh, 2013; Panda et al., 2012; Shanmugam & Idris, 2011).

Network intrusion detection is defined as a process of protecting network systems from compromised intrusion. The main operation of intrusion detection is to monitor users' activities happening over computer and network system and analyzing these actions to identify the intrusion. Many mechanisms have emerged for detecting network intrusion since the past twenty years or more, but they are still not fully grown yet (Jain et al., 2011; Kong & Xiao, 2009).

In general, IDS is a dynamic security system which can provide effective defense to the information stored in the network systems and over Internet. It is developed to identify and capture unauthorized use of, access to, or to find out malicious activities in a computer network system. It monitors network packets and tries to notify and detect the users' action as normal or abnormal (or attack) through comparing the current network behavior to the known attack signatures (Neethu, 2013; Panda et al., 2012).

The main objective of deploying the IDS is to recognize abuse, illegitimate use and misuse of network system (internal and external) attacks and to prevent them from carrying out their attacks (Jaisankar & Kannan, 2011). IDS emits alarms proportionate to abnormal or illegitimate behavior in the network, and it also sends an alert to administrator to inform him (Mohammad et al., 2011; Panda et al., 2012).

Accordance with their deployment in real time, there are two main categories of IDSs: namely, host-based IDS (HIDS) and network-based IDS (NIDS). HIDSs are designed to monitor and analyze the events on each individual computer system, host, or state, while NIDSs monitor and analyze individual package which is transferred through network for multiple devices. Both of these two types of IDSs have been developed to apply all intrusion detection methods (Kumar et al., 2012; Upadhyaya & Jain, 2013).

There are two major methods to solve the problem of intrusion detection, namely the anomaly detection and signature detection (misuse). Misuse detection is an ability for detecting attacks, which depend on recognized signatures for illegitimate behavior. The system stores pattern for known threats and use them later for comparing with the actual activities or captured data. The main advantage for this approach is that it can recognize known attacks accurately and with very low false alarm ratio, but it cannot detect novel attacks whose signatures are unknown (Powers & He, 2012; Upadhyaya & Jain, 2013).

On other hand, anomaly detection approach depends on defining the normal network behavior, and it tries to recognize packet on deviation. This method compares the captured data (current activity) with the stored normal profile. The key advantage over misuse detection is that it can recognize novel attacks where a pattern does not known if it drops out of the normal network behavior. But the drawbacks are these it has low accuracy and generates a large number of false positive alarm (Elbasiony et al., 2013; Panda et al., 2012). Therefore, many hybrid approaches have combined various techniques for enhancing accuracy of intrusion detection (Govindarajan, 2014; Powers & He, 2012).

In order to recognize the attacks with high accuracy, different techniques have been applied and suggested over the last few years. Most recent methods for processing of detecting network system attack and exploit some of AI techniques. AI methods perform major role in intrusion detection via decreasing and categorizing the data which is utilized according to the needs, and it is also used in both of detection approaches (signature and anomaly detection). AI techniques have been utilized for automating the detection process, such as neural networks, evolutionary computation, fuzzy logic, and machine learning (Shanmugam & Idris, 2011; Wang et al., 2010).

Lately, machine learning algorithms, such as clustering and classification algorithms are used to overcome the intrusion detection problem and improve its solutions. Clustering is the one of techniques for intrusion detection. Currently, clustering analysis is the most popular research. Cluster analysis involves the task of dividing data points into homogeneous classes or clusters so that items in the same class can be as similar as much as possible and items in different classes are as dissimilar as possible. It identifies such grouping (or clusters) in an unsupervised manners (Esh Narayan, 2012; Panda & Patra, 2008).

While, classification is a supervised machine learning techniques that deals with every instance of a data set and classifies it to a specific class, it extracts the models for defining important data classes. Such types of classes are called as classifiers. A classification based IDS will classify all the network traffic into natural or intrusion behavior (Wankhade et al., 2013; Xiang et al., 2014).

Recently, many researchers have proposed different hybrid intrusion detection methods which combine both misuse and anomaly detection such as combining of FCM and BRF, HMM and BN, K-means and GA, PCA and FCM and so on. But there are some main drawbacks in these existing approaches, such as adaptive ability weakness, sensitive to the order of the input data set and failure of detecting some new invasion, detection rate, false positives rate and nonresponse rates. All of these problems remain to be further improved (Bhuyan et al., 2013; Elbasiony et al., 2013; Govindarajan, 2014; Jaisankar & Kannan, 2011; Liu et al., 2014; Panda et al., 2012; Upadhyaya & Jain,

2013). In this study, it is proposed another hybrid intelligent approach for network intrusion detection which combines two artificial intelligent techniques to improve network intrusion.

1.3 Problem Statement

The main limitations of existing intrusion detection approaches are accuracy and the failure to reduce and prevent false positive alarm and false negative alarm rate. Machine learning techniques have provided important features for solving various problems related to intrusion detection (Mohammad et al., 2011; Panda et al., 2012).

Recently, some researchers have proposed hybrid machine learning intrusion detection approaches which combine clustering or classification techniques such as combining of FCM and BRF and random forests and weighted k-means. Clustering is the process of labeling the data objects as groups called cluster. Every cluster has a similar objects as much as possible and different from other objects in another clusters as large as possible. Whilst, a classification based IDS will classify all the network traffic into natural or intrusion behavior and assign each attack to its category (Gao & Wang, 2014; Govindarajan, 2014).

These proposed approaches have faces impediments like the sensitivity to amount of network data and big overlapping in these data. Such these difficulties have led to reduce the accuracy and cause big amount of false alarm rate in the proposed approaches. The difficulty of recognizing among normal and abnormal behavior in

network systems is a major problem because of the big overlapping in data monitoring. The intrusion detection processing causes a large false alarms rate resulting from the huge overlap in network's data (Gan et al., 2007; Jawhar & Mehrotra, 2010; Panda & Patra, 2008).

On other side, the main problem with classification techniques is that it cannot handle unlabeled data (Teng et al., 2010). In addition, the manual labeling is not useful, tedious, expensive, time consuming and inaccurate because of the huge amount of network data available (Jawhar & Mehrotra, 2010; Wankhade et al., 2013).

Hence, this study proposes a hybrid intelligent approach for intrusion detection which combines both of clustering technique that is K-Means and classification technique is support vector machines (SVM) to recognize and detect intrusion with high accuracy overcome the mentioned problems in existing hybrid intelligent approaches. The usage of clustering will grouping the data to solve the data overlap problem. Whereas, the applying of clustering before classification in the proposed approach will solve the unlabeled network data problem.

1.4 Research Questions

The following research questions have been investigated in order to achieve the aim of this study. They are as follows:

- i. How to design the hybrid intelligent approach for network intrusion detection to improve intrusion detection rate and reduce false alarm rate?

- ii. How to develop the proposed hybrid intelligent approach for network intrusion detection?
- iii. How to evaluate the proposed hybrid intelligent approach for network intrusion detection?

1.5 Research Objectives

The main goal for this study is to propose hybrid intelligent approach for network intrusion detection with low false rate and high detection. Keeping the aim of this study into consideration, three objectives are determined and specified; the specific objectives of this study are:

- i. To design a hybrid intelligent approach for network intrusion detection by combining the K-Means clustering and Support Vector Machine classification techniques to reduce false alarm rate and increase accuracy of intrusion detection rate.
- ii. To develop the proposed hybrid intelligent approach for network intrusion detection by intrusion dataset and Waikato Environment for Knowledge Analysis (WEKA).
- iii. To evaluate the results by applying mathematic equations to calculate accuracy, detection rate and false alarm rate, and compare the proposed hybrid intelligent approach results with existing hybrid intelligent approaches.

1.6 Significance of research

Recently, computer networks have broadly used and become complex. In the world, governments, higher education institutions and different organizations are fully dependent on the network systems that perform a key role in a daily processes. To keep services and sensitive data from such dangerous threats, the intrusion detection has become a significant issue.

Therefore, the recent hybrid intrusion detection approaches combine classification or clustering techniques or integrating both of these types to improve the accuracy of intrusion detection and avoiding the network's attacks. These hybrid intelligent approaches have some main disadvantages such as sensitivity to amount of input network data, the big overlapping in the network data, adaptive ability weakness, handling of an unlabeled data and so on. In order to overcome the mentioned drawbacks in existing approaches, the proposed hybrid intelligent approach integrated the k-means and SVM techniques. Hence, the importance of a proposed approach comes from solving the existing weakness in hybrid intrusion detection approaches. Both of proposed techniques insensitive comparatively to amount of input data and high execution speed. Moreover, the applying of clustering before classification in a proposed approach has led to resolve the overlapping and unlabeled network data problems in current approaches. Furthermore, the universality of mining process for SVM makes the proposed approach can be construct to network intrusion detection system in various computing environment.

1.7 Contributions of Research

The key contribution of this study is to propose a hybrid intelligent approach for network intrusion detection, which combines two artificial intelligence techniques, and they are clustering and classification. The specific contributions of this study are in the following:

- i. Design a hybrid intelligent approach for network intrusion detection with high accuracy and low false alarm rate and its workflow.
- ii. Reducing the manual work through categorizing and labeling the network data by performing the clustering before classification.

1.8 Scope of Research

The scope of this study is limited to wired network. Also, the proposed intelligent hybrid approach is combination of k-means and support vector machine (SVM). The proposed hybrid intelligent approach will be tested by applying the k-means and SVM techniques using Waikato Environment for Knowledge Analysis (WEKA). One of network intrusion dataset types will be used for testing the proposed hybrid intelligent approach.

1.9 Thesis Organization

The rest of the thesis is structured as follows:

Chapter One: presents background about study and problem, problem statement, objective of this study, scope, significant and contributions of research.

Chapter Two: explains detailed description of the network intrusion term, network intrusion detection concept and network intrusion detection system (types, approaches and behavior), and also it presents the role of artificial intelligent techniques for solving network intrusion problem. Lastly, it shows some related work done so far about this problem and existing intelligent approaches.

Chapter Three: describes the research methodology phases, shows in details about proposed work, and describe the algorithms used in this approach.

Chapter Four: summarizes and analyzes the results of the experiment. Also, it shows a comparison of the results with existing hybrid intelligent approaches.

Chapter Five: consists of the conclusion of the study, recommendation and the future work needs to be done in this area.

1.10 Summary

In this chapter, we have shown the overview of the network development and its security issue. In addition to it, one of the most important threats on computer system which intrusion and the methods of solve this problem. The role of artificial intelligence in the solution of network intrusion detection. The statement of problem for this study is explained and a hybrid intelligence approach is proposed, which combines k-means and SVM. Moreover, the scope of this study and the contributions are included in this chapter.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, the researcher has focused on the relevant literature that is related to the context of this research. Most of the literature that is highlighted in this chapter is part of the broader area which touches upon and explains the concepts and principles of network security and network intrusion detection problems. Also, the literature relates the significance of network security today and the types of attacks that should be avoided. In addition to it, it explains the main concepts in this study, such as intrusion detection, network attacks, types of attacks, IDSs, the approaches of solving intrusion detection problem, the role of AI in intrusion detection problem and the main techniques which are used in this field. The literature shows the existing related studies for intrusion detection problem as well.

2.2 Network Security Overview

In the last second half of the last century, computer networks started growing with tremendous speed. The data sharing, exchange of messages and enabling using of applications over interconnected devices situated in different places are the main goals of design computer network. This technology has spread quickly and internet has become worldwide an important cornerstone for our daily processes in life (Jaisankar & Kannan, 2011; Selman, 2013).

The importance of computer network and Internet is coming to the fore from adoption of these technologies for several activities, such as business, defense, education and banking. Nowadays, governments, higher education organizations and different organizations are fully dependent on the computer system networks that perform a key role in the daily processes. Almost everybody with a computer or mobile device is linked with the internet in order to have access to data and sending messages, and people share a lot of sensitive information among various kinds of computer devices, from minicomputers to huge servers and mobile devices (Jiang, 2012; Upadhyaya & Jain, 2013).

The development of infrastructure related to cloud computing services, the existence for many websites of social networking and the huge rise of mobile platforms using access services on the web have led to increase in the traffic of internet. Accordance with the increasing in terms of size of Internet traffic, one of the main challenges already faced by computer networks is security (Bansal et al., 2010; Jiang, 2012; Sanyal & Thakur, 2012).

Nowadays, due to various types of threats over internet, security has become the main concern in network technologies. Enterprise systems are suffering from many problems, such as losing of data, illegitimate access for information and malicious usage of system resources (Garzia et al., 2012; Muniyandi et al., 2012).

Network security deals with security concerned with the transport of data. It can be defined as the operation of protecting the main factors for any computer system security. The factors related with computer system security are confidentiality, integrity, and availability. These three concepts are defined as follows: firstly, confidentiality means that computer related assets, such as data is displayed according to the rules only, that is, only those who should have access to anything will actually get that access. Secondly, integrity refers to that data which is not damaged or destroyed through transferring, and the correct performance of system is monitored only by authorized parties and in an authorized way. Finally, availability means that system services are available when they are needed and information is accessible to authorized parties at appropriate time (Eric, 2009; Panda et al., 2012).

Different techniques have been employed to enhance security for data being transmitted over Internet as well as hardening computer network. There are two main threats to security, and they are intrusion and virus (Bhuyan et al., 2013). In general, intrusion refers to either a hacker or a cracker. The techniques, such as access control, firewalls, cryptography, biometric, and new algorithms to evaluate vulnerability into system are being used to reduce risks into systems, on the other hand; the risk of attacking information over the Internet has also increased simultaneously (Idika et al., 2009; Jaisankar & Kannan, 2011). However, all the above solutions cannot alone prevent the possible attacks. The compromised nodes can launch mentionable numbers of attacks in computer network, therefore; intrusion detection is needed, and it is essential to find out an efficient approach to protect it (Daniel et al., 2013).

2.3 Network Intrusion Detection

Nowadays, the network applications have increased rapidly through Internet. Hence, there is a need to detect the malicious packets, such as virus, intrusion and worm in the network to support these applications. Intrusion is the most publicized threat to security. It is considered as the main part in data assurance. Several techniques are employed to solve intrusion problem in network systems as the initial shield of defense for network security, such as data encryption, firewall, avoiding programming, and authentication, but these traditional techniques still are insufficient for solving intrusion problem. Today, an essential factor in information and communication technologies is intrusion detection (Husagic et al., 2013; Prabha & Sukumaran, 2013).

Due to the advancements in internet technologies and the concomitant rise of a number of network threats, network intrusion detection has become a significant research issue. In spite of remarkable progress and a large number of research work, there remain still many opportunities to advance the state of the art detection systems in detecting and thwarting network-based attacks. Intrusion detection is an area which is growing to achieve data storing and processing over network systems more confidentially. Intrusion also includes a set of different techniques developed for detecting network, malicious activities and providing evidence of attacks (Chowdhary et al., 2014; Joshi & Varsha, 2013).

The definition of network intrusion refers to any malicious activities or illegitimate access from users or another system on network and computer systems. In general, network intrusion detection is the process of protecting network systems from compromised intrusion and hacking. As a result, it can be said that intrusion detection is an operation of monitoring and detecting unauthorized activities, entrance, or data editing on network (Babatunde et al., 2014; Joshi & Varsha, 2013).

Network intrusion detection can also be configured as the detection of outside illegal visitors who do not have authorization to use a network system and inside intruders who are legal users and have authorized account on the network system, however; they abuse their privileges. Thus, it requires an extra shield to protect network in spite of traditional techniques. The successful detecting does not solely benefit the intrusion detection, but also it protects security system from break attempts by monitoring network, which gives necessary information for timely countermeasures. The main operation of intrusion detection is to monitor the users' activities happening in a computer system or network and is followed by analyzing these activities for identifying attacks. Intrusion detection mechanisms has emerged since more than twenty years, but it is still not fully grown yet (Govindarajan, 2014; Jain et al., 2011; Kong & Xiao, 2009).

2.4 Network Attacks

Network threats and an intrusion are referred to as an unauthorized and intentional attempt to get access for data, modify information or deliver unusable system or provide inaccuracy. There are different types of attack on computer network, which can be classified into four main groups based on their intrusion behavior, and also there is introduction regarding several names of attacks that belong to each one of them (Al-Jarrah & Arafat, 2014; Bhuyan et al., 2013).

2.4.1 Probing (Probe)

Unauthorized user attempts to get information about victim machine and discover system's vulnerabilities by monitoring the networks. As example of this type are: ipsweep, Satan, and nmap attacks (Joshi & Varsha, 2013; Selman, 2013).

2.4.2 Denial of Service (DoS)

Under this type of threat, intruders try to prevent legitimate users from accessing and using the services in the target machine by exhausting the bandwidth or overloading network computational resources. For examples Teardrop, Smurf, and Neptune attacks (Hameed et al., 2013; Kulhare & Singh, 2013).

2.4.3 Remote to Local (R2L)

For this category, attackers do not have a legal profile access on the target system, so they attempt to get access from a remote computer by exploiting the holes in local machine. For instance guess password, ftp_write, and imap (Hameed et al., 2013; Husagic et al., 2013).

2.4.4 User to Root (U2R)

For the last category of intrusion, intruder has local access with normal user privileges to the victim machine and attempts to use vulnerabilities of the system, so he can gain an access to a root of the system and super user privileges, such as Perl, eject, and load module attacks (Husagic et al., 2013; Selman, 2013).

2.5 Intrusion Detection System

Typically, intrusion happens as anomalous activities through specific data model techniques by a sequential method and detecting it subsequences. The main cause of those anomaly patterns is attacks which are launched by external attackers who attempts to get illegal access to network system for stealing data or for damaging the network systems. In a normal environment, a network is linked with the whole world through the worldwide network (Pei et al., 2013).

Consequently, the system for intrusion detection has become a necessary part of computer network security. IDS is utilized to support computer networks to set for and transact with any attacks. This target is achieved via gathering data information from various systems and network sources and analyzing it to symptoms of intrusion problems. In several situations, IDS permits the administrators to assign immediate responses for the attacks (Babatunde et al., 2014; Chowdhary et al., 2014).

The main tasks of intrusion detection systems are in the following, and they are monitoring, observing, and analyzing system and users activities. Reviewing of system settings and vulnerabilities and estimating the safety of essential system and information files also come under detection system. The activity patterns recognition detects attacks. Analysis of the malicious and attacking activities patterns are shown statistically. Figure below shows the main process IDS (Babatunde et al., 2014; Husagic et al., 2013).

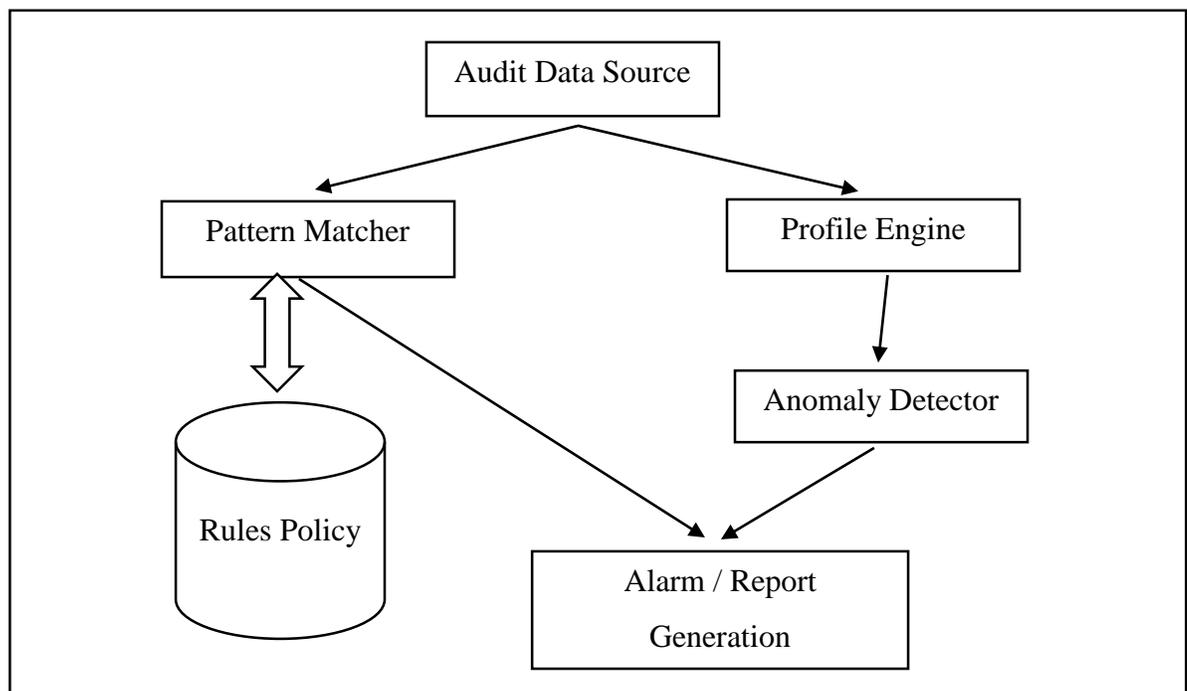


Figure 2.1: A Generic Intrusion Detection System

IDSs are dynamic security systems which can provide effective defense to the information stored in the networked system and also over the internet. It is developed for identifying and capturing unauthorized use of, access to, network system and protecting from malicious actions taking place in a computer network system (Rangadurai et al., 2012).

IDS performs three main security tasks, and they are monitoring, detecting, and responding to malicious actions respectively for both two types of attacks (internal and external). It monitors network packets and sessions, then tries to observe and detect the network users' profile as attack (abnormal) or normal through comparing the current network behavior to the known attack signatures (Jain et al., 2013; Panda et al., 2012).

The main objective of deploying the IDS is to recognize the abuse, illegitimate use, and misuse of network system (internal and external) attacks and to prevent them from carrying out their attacks (Jaisankar & Kannan, 2011). IDS emits alarms in proportionate to the abnormal or illegitimate behavior in the network and sends an alert to a security administrator if abnormality is found (Ishida et al., 2011; Mohammad et al., 2011; Panda et al., 2012).

There are many standards which can classify IDSs, depending on its various factors. Firstly, according to system operation, there are two main categories: namely, centralized and distributed. Distributed system is run as standalone concept, no agent is required. On the other hand, centralized system is run beside independent agent, and it has ability to a preemptive response and reaction measures. Secondly, IDSs are categorized into active and passive systems, depending on its behavior for the intrusion response. After detection by the usage of normal behavior information, system behavior characterizes its response for the attacks (Liao et al., 2013).

System architecture is one of these standards which have categorized IDS into two types, namely, host-based and network-based IDS. HIDSs are installed in every host by monitoring and gathering audit trail data in real time. Also, it is fully dependent on agent-based, thus; it divides network overload and CPU usage and introduces means of security administration more flexibly. On the other hand, NIDSs are distributed on network devices like hubs, routers and so on. It monitors and collects data from the packets, which are transferred over network devices to detect an attacks. Finally, IDSs can be categorized according to the method of detection process. Two broad techniques come under this classification, and they are misuse and anomaly approach (Satpute et al., 2013).

Finally, IDSs can be categorized according to method of detection process. Two broad techniques in this classification which are misuse and anomaly approach as shown in Figure 2.2 (Chapke Prajkta & Raut, 2012; Shanmugam & Idris, 2011).

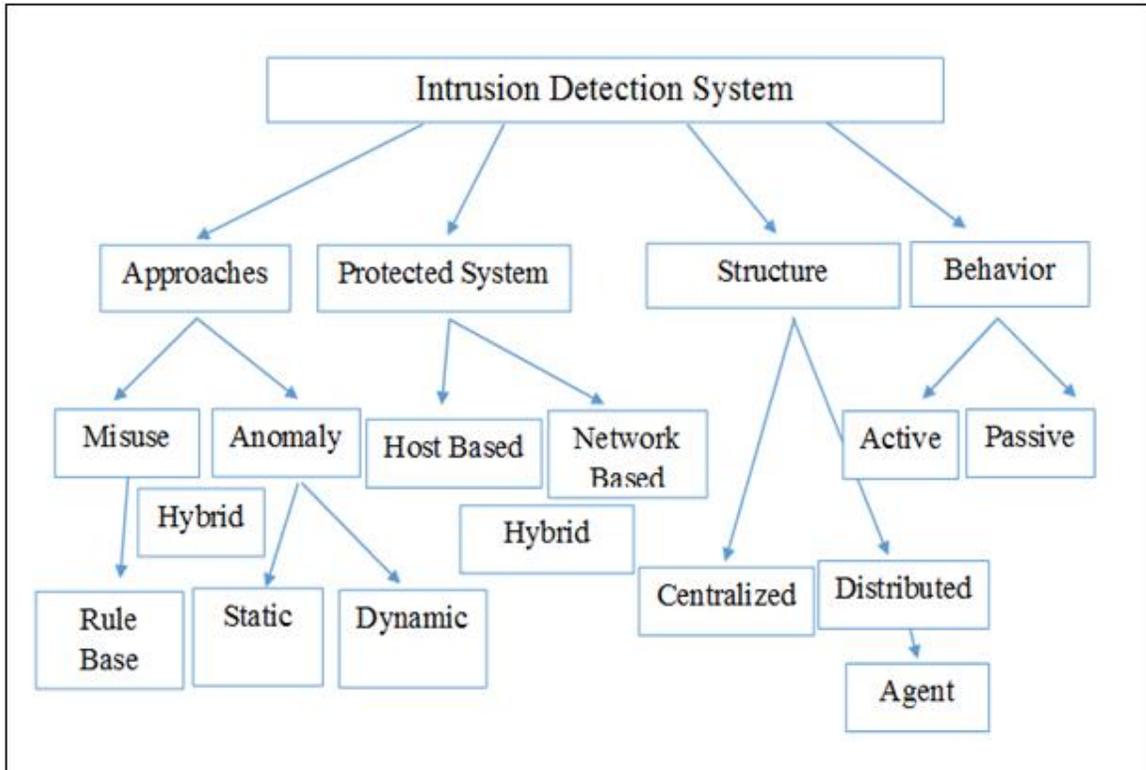


Figure 2.2: Classification of Intrusion Detection Systems

During the last couple of years, the researchers and developers have employed many methods for developing IDSs. But, there are some drawbacks which come simultaneously with the current detection systems, and they are in the following (Elbasiony et al., 2013; Mohammad et al., 2011):

- 1) **High false positives rate:** under this case, there are high number of incorrect alarms and inaccuracy attacks detection. It is used through reduction thresholds for decreasing the false alarms and increasing attacks getting undetected as false negatives. Today, one of the main problems that IDS faces is the ability of intrusion detection for detecting attacks with high accuracy.

- 2) **High false negatives rate:** several attacks are still unrecognized by current intrusion detection system. IDS still cannot detect all network attacks yet. Therefore, increasing an efficiency of IDSs for detecting attacks is another key problem for researchers.

2.5.1 Intrusion Detection System Sites

IDS is one of the powerful and essential technologies, which detects and prevents illegitimate and malicious access to network, consequently; it contributes in the protection of network and important data. Attackers (intruders) are trying to misuse resources of networks, or damage computer networks. There are two main kinds for intruders; one is internal who has access and tries to misuse it or the external one who tries to get unauthorized access for network. Other taxonomy is based on the data source which is used to make intrusion detection. The categorization can be derived, depending on the data collected from single machine (called as Host IDS (HIDS)), on the other hand; the data collected from packet is transferred over network, and this is monitored closely (Chapke Prajkta & Raut, 2012; Kumar et al., 2012).

2.5.1.1 Host Based Intrusion Detection System

HIDS is that system which is installed on each host machine locally. It is concerned with what is happening on each individual host. They are able to monitor and analyze the activities only on the host system. It receives information from operating system application subject. It monitors the status of key system files, such as the identification and authentication mechanisms, administrative action and access log files. Figure 2.3

shows the structure and location of HIDS (Elbasiony et al., 2013; Kulhare & Singh, 2013).

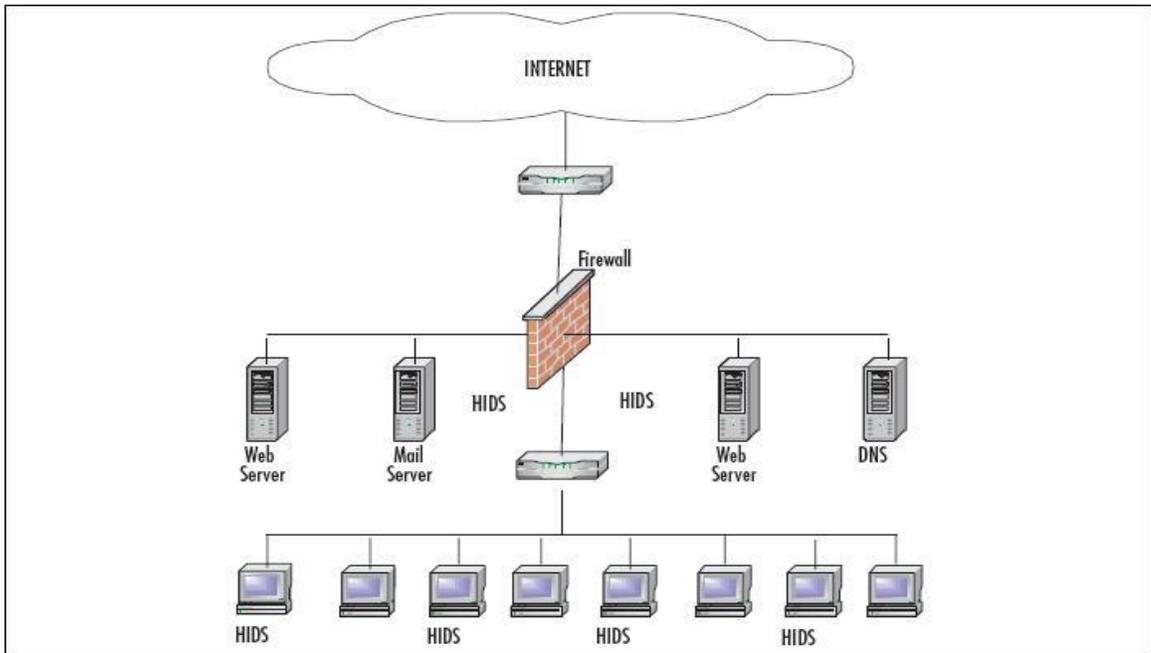


Figure 2.3: Host Based Intrusion Detection System

HIDS monitors any changes to any single system and detects the illegal changes. They typically monitor logs, system calls, and system activities in a way to detect any intrusions attempted to a system. HIDS are placed on a single host and require a lot of installation if they are implemented in a large scale. HIDS, on the other hand, monitors inbound and outbound network traffic and detects if there are any intrusions. HIDS only monitors the host that they are installed and intrusion cannot be detected on other host. HIDSs monitor and analyze the inner of a computing system rather than the outer interfaces of system. A HIDS might detect internal activity, such as program accesses and attempts of illegitimate access. HIDSs can be defined as the agents that monitor any

external and internal activities which have circumvented the policy of security that attempts are made to enforce by the help of the operating system (Chowdhary et al., 2014; Shivakumar et al., 2013).

2.5.1.2 Network Based Intrusion Detection System

NIDSs can be placed anywhere in the network and is usually attached to any network devices or being installed independently, and it monitors everything, so it can detect any intrusion at the first place. It picks and analyzes individual packets flowing through network. NIDS is run on the independent machine, such as hub, switch or router. These devices work as concentration points and allow IDS to monitor and collect audit data for the entire network (Ahmad et al., 2013).

Also it is defined as that it is an autonomous environment where it can identify attacks by analyzing network traffics and monitoring various hosts. It gains access for network traffics through linking for a hub, switching configuration to port mirroring, or tapping network. The problem with NIDS is that it cannot detect any packets that are encrypted or obfuscated. NIDS as shown in Figure 2.4 (Joshi & Varsha, 2013).

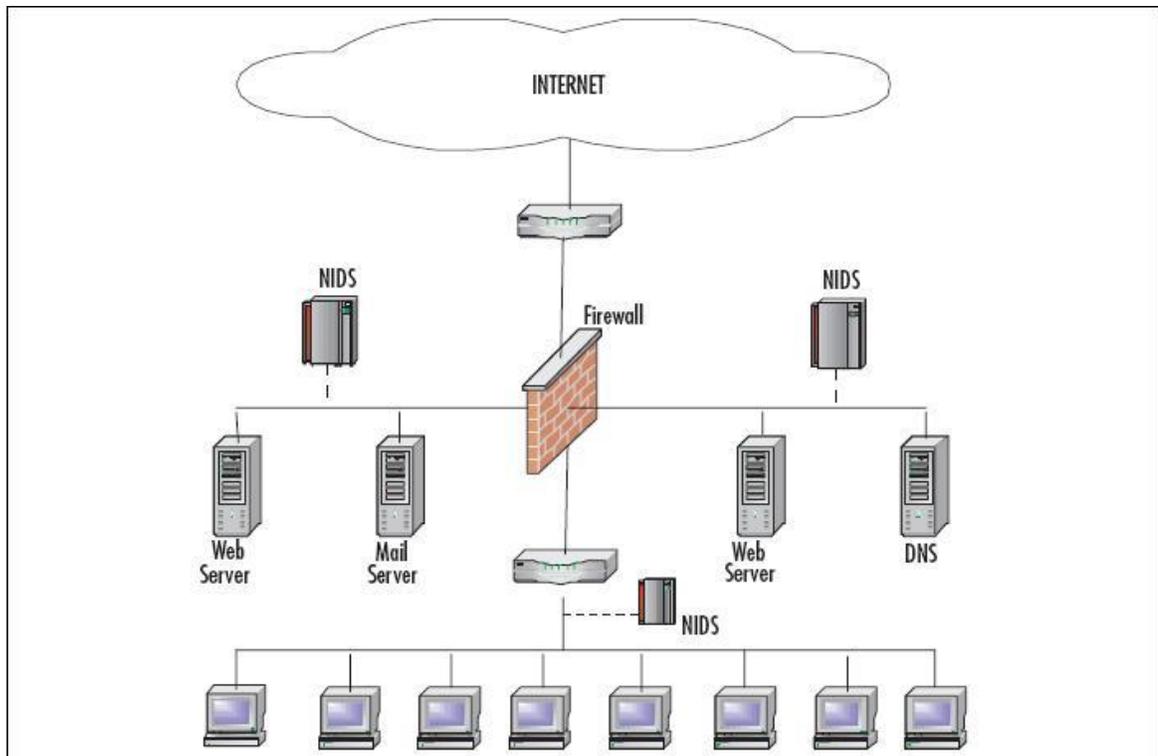


Figure 2.4: Network Based Intrusion Detection System

The NIDS reads all incoming packets or flows, looking for and finding out suspect behavior. For instance, if the administrator observes the big requests connection of TCP to most of ports in short time, that means, an attack is doing a scan for ports in one of network machines. Apart from checking the incoming traffic, a NIDS also provides valuable information about intrusion from local or out traffic. In some cases, the incoming traffic does not regard the cause because of the attacks which can be internal from monitored or segmented network. Data for intrusion detection are available in various levels of subdivisions (Shivakumar et al., 2013). Table 2.1 contains differences between HIDS and NIDS.

Table 2.1:

Network Based vs. Host Based Intrusion Detection System.

NIDS	HIDS
<ul style="list-style-type: none"> - Broad in scope (watches all network activities). - Better for detecting attack from outside. - Less expensive to implement. - Examines packet headers, and detects network attacks as payload is analyzed. - Near real time response. - OS independent. 	<ul style="list-style-type: none"> - Narrow in scope (watches only specific host activities). - Better for detecting attack from inside. - More expensive to implement. - Does not see packet headers - Usually only respond after a suspicious log entry has been made. - OS specific.

2.5.1.3 Hybrid Intrusion Detection System

The main difference in both types of IDS is that HIDS monitors the traffic of network in each host. On the other hand, NIDSs monitor all traffic of network. In network security, IDSs are the systems which monitor and analyze traffics of network to recognize and sign unauthorized activities and access. It will be more useful for the IDS development for getting fully integrated into NIDS, like filter alarms in similar method to HIDS and it can be controlled from one place (Panda & Patra, 2008).

Hence, full secure network must have both kinds of IDSs: NIDS and HIDS are used to implement detection systems to provide a full defense shield against various threats, doing efficient and effective monitoring. In addition to it, they are used to tap detection and response against illegitimate access and activities. Figure 2.5 below indicated the hybrid intrusion detection system.

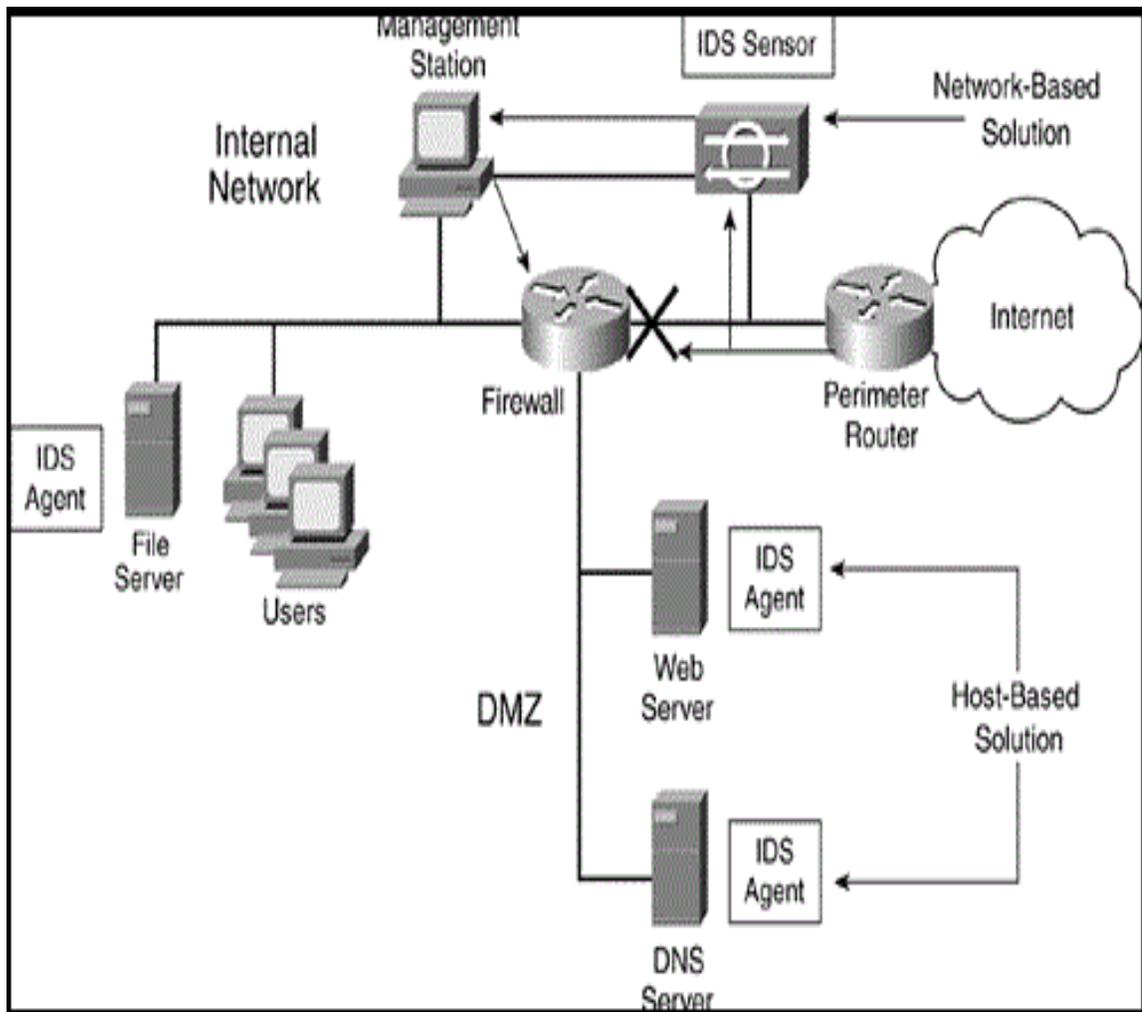


Figure 2.5: Hybrid Based Intrusion Detection System

2.5.2 Intrusion Detection System Behaviors

Behavior on detecting of IDS is one of the standards which can classify the IDS on this basis. It describes response of the system after detecting attacks. There are two types of IDS based on behavior, namely: active and passive, and they depend on the response for attacks.

2.5.2.1 Passive Behavior

It does not disrupt the process of the connection among the nodes and routing protocols. But it tries to discover and retrieve the important information by listening to network traffic. Passive threat is often difficult to detect, therefore; it is complex to have defense against passive attack. The passive system can detect anomalous behavior of packages only without emitting any kind of alarm for the administrators. Examples for this type of behavior are analysis of traffic, eavesdropping, and monitoring (Upadhyaya & Jain, 2013).

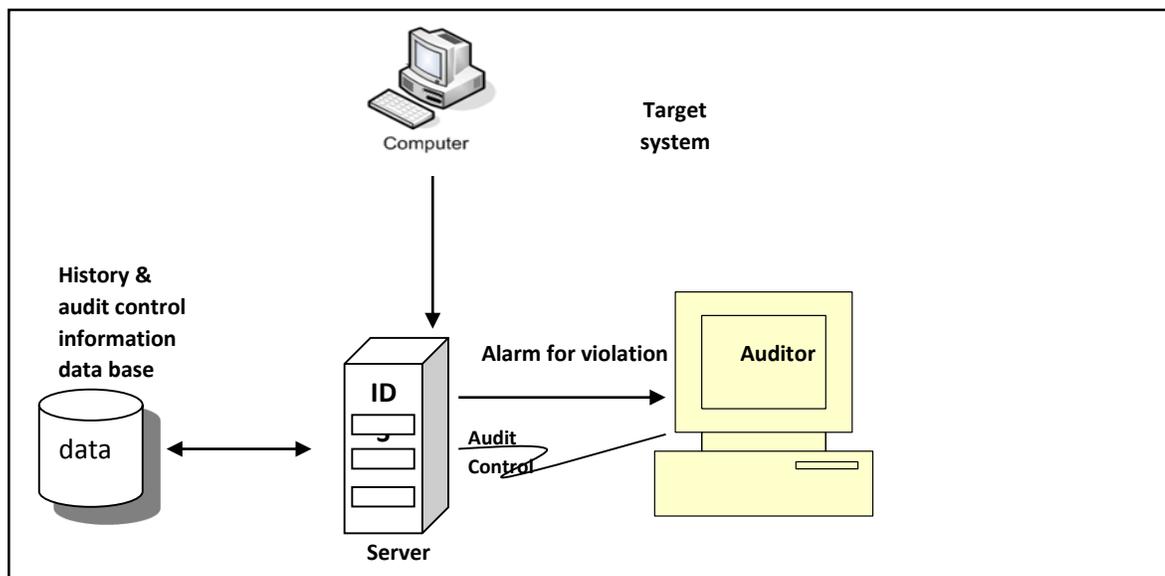


Figure 2.6: Passive Intrusion Detection System

2.5.2.2 Active Behavior

An active attack inserts its own data into the data stream, and this process attempts to disrupt and impede the normal function. Such attacks are attempted to modify or destroy the data being exchanged in the network so as to ensure availability limitation, get authentication or attacks packages prepared for another nodes. The examples of this are impersonating, jamming, denial of services, modification, and message reply. The active IDS can respond the anomalous activities easily by signing out an attacker or blocking the traffic of network from the suspicious anomaly users (Upadhyaya & Jain, 2013).

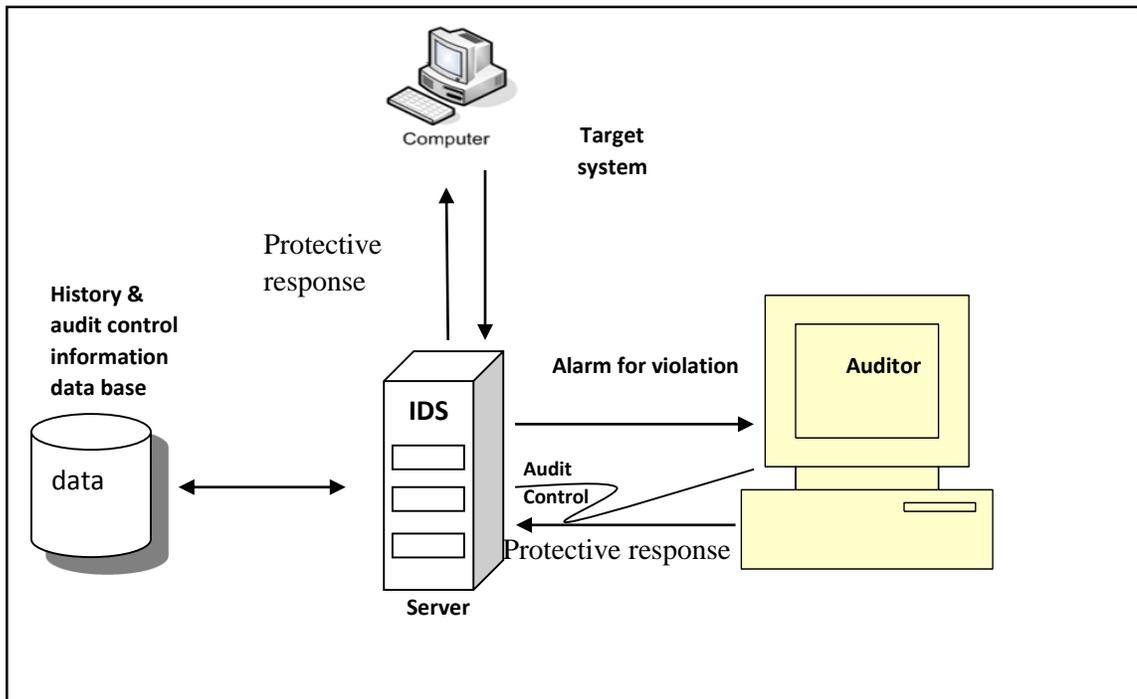


Figure 2.7: Active Intrusion Detection System

2.5.3 Intrusion Detection System Approaches

There exist two wide methods to solve network intrusion detection problem and they are applied for both two types of network intrusion detection system, namely, anomaly detection and misuse (signature) detection approaches.

2.5.3.1 Misuse Detection Approach

Misuse approach is an ability for detecting attacks, depending on predefined signatures of malicious activities. Systems store patterns (signatures) of known attacks and use them to compare with the actual activities or captured data. For signature detection method, the IDSs analyze the data which are collected and compared to attacks pattern, which are saved in the big database of known attacks (Powers & He, 2012).

Basically, IDSs search about certain attacks which have been stored already, and they come under detection systems. IDS software are solely good as much as the attacks have pattern database, which is utilized for comparing packages against attacks. When it detects an activity matching, a signature is stored in an attack signature table. The system alarms to inform the administrators (Bahrololum & Khaleghi, 2008; Upadhyaya & Jain, 2013).

The main advantages of this technique are that it can accurately detect known attacks and it has very low false alarm ratio too. In addition to it, patterns are understood and developed easily if network behavior is able to know what it is attempting to recognize. On the other hand, the disadvantages of it are that it cannot detect novel attacks whose

signatures are unknown and the attacks database need to be updated all the time. Also, it depends on string matching and regular expression only, therefore; it can deceive easily. This technique deals with strings inside packages which is transferring data over network (Govindarajan, 2014; Jaisankar & Kannan, 2011).

The misuse of systems' efficacy is reduced highly as they have to make new patterns to every attack type. The performance of systems engine reduces because of the number of attack signature is still increasing, consequently; a lot of IDS engines are used with many processors and network cards. In order to avoid the new attack on the network, IDSs designers start making a new pattern before the intruders do anything. The precedence of making new patterns among intruders and designers assigns and showcases the system efficiency (Jyothisna & Prasad, 2011).

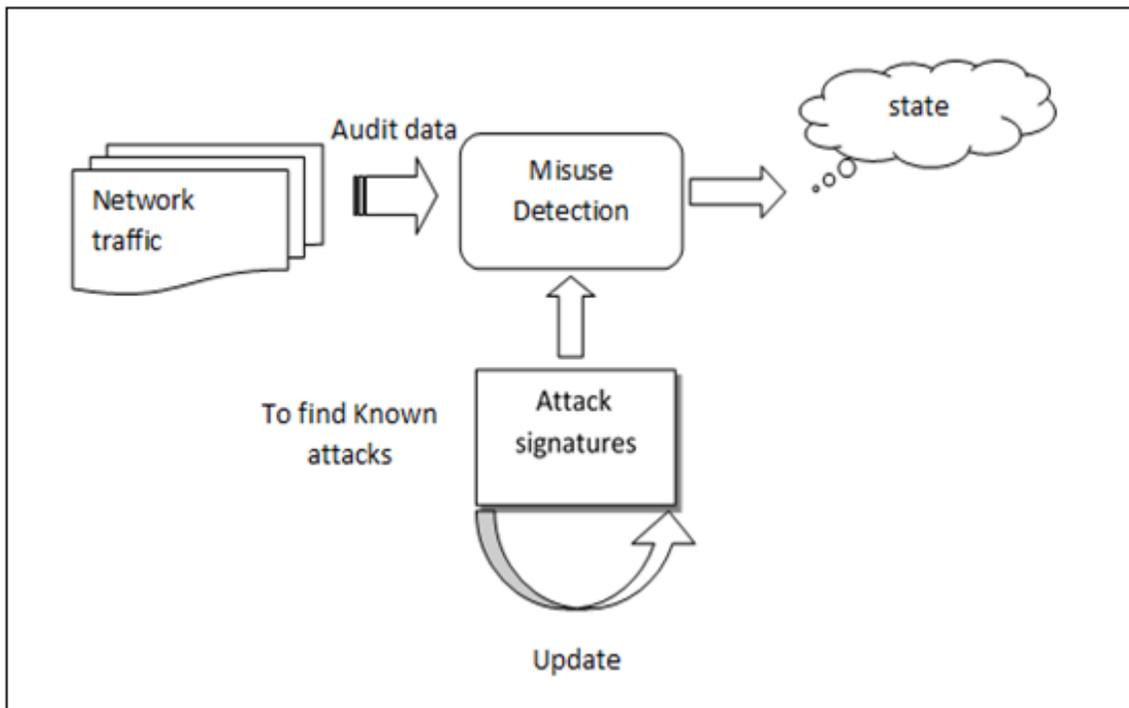


Figure 2.8: Misuse Intrusion Detection System

2.5.3.2 Anomaly Detection Approach

Anomaly detection approach depends on defining a network behavior (profile) and trying to detect traffic on deviation created by normal network behavior. This method compares a captured data (current activity) with the stored normal behavior. For anomaly method, the administrators of system determine the normal profile (baseline) for the traffic of network, protocols, typical package size, and breakdown. The detectors of anomaly approach monitor the packets by comparing their behaviors to the normal profile of network and finally they recognize the attacks (Bahrololum & Khaleghi, 2008; Panda et al., 2012; Upadhyaya & Jain, 2013).

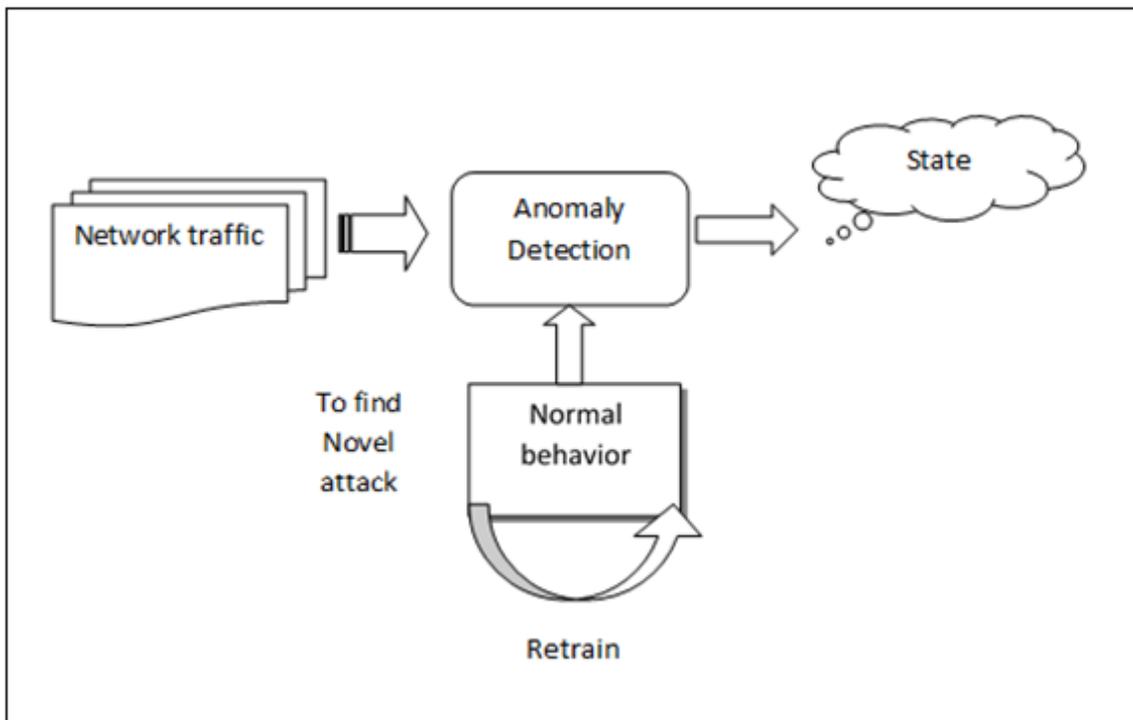


Figure 2.9: Anomaly Intrusion Detection System

Normal profile is defined by the earlier observed user behavior which is usually created during a training phase. The essential stage in determining normal network profile is that IDS engine has ability for cutting over the different protocols in all levels. The engine should be able to understand the aims of protocols and process them. Many benefits are created by protocol analysis, (which is computationally expensive) such as growing a rule set helps for decreasing false positive alerts. If the system detects the difference from the normal profile, the alarm is generated to inform the administrators or to respond properly (Elbasiony et al., 2013; Powers & He, 2012; Shanmugam & Idris, 2011).

The key advantage of anomaly detection method in comparison with signature method is that the anomaly technique can detect a novel intrusion. When a pattern is unknown, its behavior drops out of normal network profile. But it also has many drawbacks; the main disadvantage of it is how to determine the rules. System efficiency is based on how well it is applied and tested on all protocols. For the accuracy of detection, the detailed information about normal behavior is needed for getting development by system administrator. Sometime malicious activities of the intruders fall under the normal profile, and after that it goes undetected, therefore; it has low accuracy, and it generates a large number of false positive alarm. Also, it is difficult to determine among normal profile and malicious profile in networks because of the high overlapping in monitoring data. Several techniques are used in anomaly method. Several techniques are used in anomaly method as shows in Figure 2.5 (Bhuyan et al., 2013; Jyothsna & Prasad, 2011).

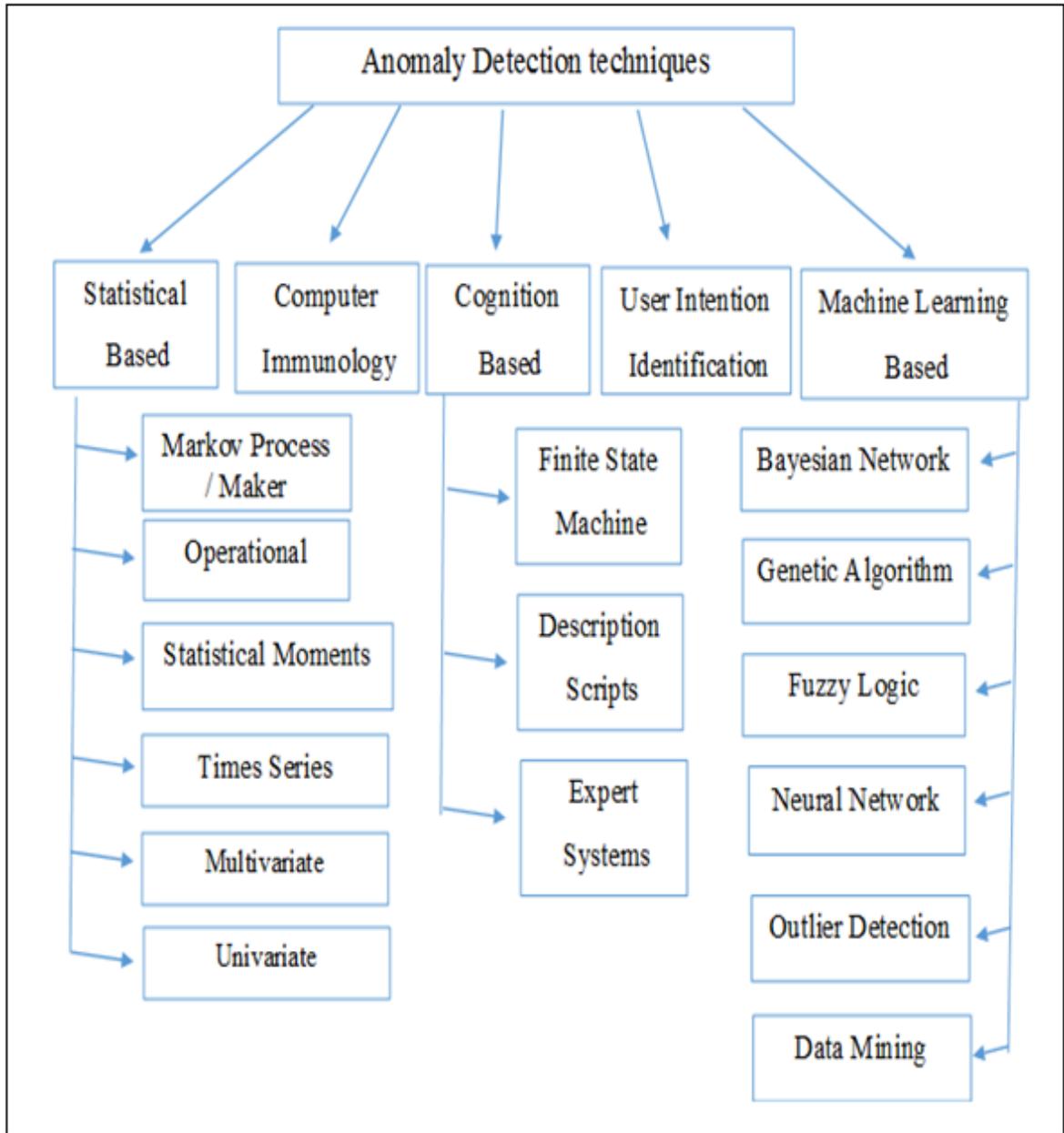


Figure 2.10: Classification of Anomaly Based Intrusion Detection Techniques

Here are the advantages and disadvantages of misuses and anomaly intrusion detection in the following table:

Table 2.2:

Misuse vs. Anomaly Intrusion Detection System

	Advantages	Disadvantages
Misuse IDS	<ul style="list-style-type: none"> - Can name attacks. - System administrator can write their own signatures. - Easy to implement. - Generate less number of false alarms. <p>This is because the system only gives an alert when already known signatures are found in the audit trail.</p>	<ul style="list-style-type: none"> - Cannot detect novel or new attacks. - Need to update with new signatures database. - The signature database tends to get big and after while this can slow down the system.
Anomaly IDS	<ul style="list-style-type: none"> - Can detect new or novel attacks. - Can use more sophisticated rules. 	<ul style="list-style-type: none"> - Complex to implement, it is very hard to define what exactly normal behavior is. - High rate of false alarm.

2.5.3.3 Hybrid Intrusion Detection Approach

Both of the intrusion detections are misuse-based and anomaly-based approaches, and they are merged into to cope up with their drawbacks. Anomaly method can detect a novel attack to raise the detection rate. On the other hand, misuse-method recognizes known attacks. If the known attacks are removed, the number of attacks can be decreased greatly in datasets for the sake of unsupervised anomaly detection. Misuse approach has high detection rate with low false positive rate. Hence, integrating anomaly and misuse approaches will be able to enhance the performance of the IDS (Pillai et al., 2011).

2.6 Artificial intelligence for Intrusion Detection

AI techniques perform a key role in intrusion detection process by decreasing the data which is utilized to detect and also to classify the data according to the needs and it is used in both approaches (anomaly and misuse detection). AI mechanisms have been utilized to automate an intrusion detection operation, for examples, fuzzy inference systems, ANN, evolutionary computation, SVM and AIS (Kumar et al., 2012; Shanmugam & Idris, 2011).

2.6.1 Artificial Immune Systems (AIS)

AISs use ideas of the human immune system operation and they perform to detect computational problems. Particularly, for intrusion detection problems, the immune systems can be considered as anomaly detection because it can recognize between normal and abnormal self. This technique is applied by a specific kind of lymphocyte named as T-cell (Powers & He, 2012).

In the last few years, AISs have been researched widely, especially for anomaly detection. Many researches have been applied to use negative selection as this model is appropriate to anomaly detection. The ease of explaining the detectors in term of domain knowledge is the key to getting advantage in this technique. However, the scalability applied to real-world IDS is still the main problem for this technique (Jain et al., 2011; Powers & He, 2012).

2.6.2 Artificial Neural Networks (ANN)

ANN contains a set of processing elements which are strongly interconnected and can convert a collection of inputs for collection of required outputs. The conversion results are assigned by the characteristics of the elements and the weights related with the interconnections among them. The networks are able to adapt the required outputs by modifying the connections among the nodes (Wang et al., 2010).

There are many advantages in this technique, such as flexibility, ability to process from many resources and the inherent speed of neural networks. It has also two main drawbacks as well; the first one is the requirements of training for the neural network. The training data and methods utilized are critical because the ANN ability for identifying of intrusion is fully dependent on the accurate training phase of system. The training procedure needs a very big amount of data to ensure accuracy of results statistically. Secondly, the most significant disadvantage of applying neural networks to intrusion detection is the "black box" nature of the neural network (Ibrahim, 2010).

2.6.3 Fuzzy Logic (FL)

Fuzzy logic techniques are used to refer uncertainty management widely. Few existing (IDS) use this powerful reasoning technique. Two main categories of IDS that are based on fuzzy logic are: fuzzy clustering and rule based system. Data mining techniques are used by rule-based IDS for detecting hidden information by scrutinizing a big collection of data. The system challenges discover the more significant rules and set them for classifying the hidden information. The other use of fuzzy clustering is related to the performance meant for fuzzy logic for IDS (Ghadiri & Ghadiri, 2011; Shanmugam & Idris, 2011).

The goal of clustering is to divide a collection of data for clusters. Clusters must have the following properties, and they are homogeneity within the clusters, concerning data in same cluster, and heterogeneity between clusters. On the other hand, data belonging to different clusters should be as different as much as possible. The training phase is clustering data into many subsets through fuzzy clustering module (Wang et al., 2010).

2.6.4 Genetic Algorithm (GA)

GA works with the concepts of biological development, genetic recombination and natural selection. GA works on an individual basis called as chromosome and evolves the group of chromosomes to a population of quality individuals. A fitness function will be there for each rule which is a measurement of each rules implementation. Three genetic operators are applied for every single, through the generation process which are: selection, crossover, and mutation. It has been applied heavily to solve security problems with particular reference to network intrusion detection. It offers the benefit of evolving their behavior to have comparison with the behavior of final users. Various GAs can be utilized to perform the required purpose; each one is characterized by peculiar features. The drawbacks of these are that it cannot determine the attacks in audit trails and cannot detect new attack as it needs more domain specific knowledge to handle the new attack. Genetic based intrusion detection has no ability to detect multiple, simultaneous responses, and this detection approach meets computation complexity problems (Akbar et al., 2011; Liu Li , Wan Pengyuan Wang , & Songtao, 2014).

2.6.5 Support Vector Machine (SVM)

SVM is one of machine learning techniques, and it performs the training vectors in high dimensional feature space and also categorizes every vector to its class. It classifies the data by identifying the groups of support vectors. It can be used for solving the highly nonlinear classification and regression problems with the linear learning machine method in the sample space (Chitrakar & Chuanhe, 2012; Govindarajan, 2014; Kong & Xiao, 2009).

SVM has become the main technique to be used in anomaly intrusion detection because of its good generalization nature and the ability to solve the curse of dimensionality problem. The aim of SVMs is to determine a classifier or regression function which can minimize the empirical risks and the confidence interval (which corresponds to the generalization or test set error) (Chitrakar & Chuanhe, 2012; Govindarajan, 2014).

2.6.6 Hidden Markov Models

HMM is a double embedded random processing, which is consisted of an underlying random processing that is hidden. But it can be only noted by other set of stochastic processing that produces the sequence of observations. HMM processing, such as transition and state are invisible. Every transition emits a separated set of symbols rather than an output symbol. These sets of symbols are the observed variable utilized for training HMM.

HMM is such machine learning technique that the only technique can explicitly learn state based classification (sequential modelling). Thus, it is not limited to performing stateless attacks detection. This advantage enables HMM for doing more complete intrusion detection and detecting many-stage intrusion. In spite of having advantages, HMM has disadvantage too; HMM has not been applied to intrusion detection widely like other machine learning techniques (Jain et al., 2011; Rangadurai et al., 2012).

2.6.7 Naïve Bayes

NB is a simple version of Bayesian network, which offers machine learning abilities. There are two main particular disadvantages of Bayesian networks. The requirements of a priori information about the problem for defining probabilities and the approach are computationally expensive. In the first case, it is possible for extracting probabilities from training data, if available data is applied to NB. However, NB assumes that all data's features are independent to each other, and this is the reason to apply Bayesian networks instead of NB to database intrusion detection (Chitrakar & Huang, 2012; Jain et al., 2011).

2.6.8 Data Mining

Data mining or (knowledge discovery) is the operation of identifying and defining the patterns from big amount of data. Also, it is defined as a process of finding and discovering useful and meaningful patterns and their relationships in huge volumes of data. This scientific area also contains tools, techniques, and algorithms from artificial intelligent, such as machine learning algorithms and neural network and statistical

environment to analyze huge digital collections, which are called as data sets. Knowledge discovery is broadly used in numerous purposes which is the best technology for discovering the knowledgeable patterns.

Recently, data mining algorithms and techniques have played a key role in network intrusion detection. Data mining can be applied to get insightful knowledge of intrusion prevention mechanisms. They can help to recognize new types of attacks and vulnerabilities as well as intrusions; they discover previous unknown patterns of attackers' behaviors and provide decision support for intrusion management. Data mining techniques, such as clustering and classification are proving efficient solution for network intrusion problem by analyzing and dealing with large amount of network traffic (Dhawan, 2013; Wankhade et al., 2013).

2.6.9 Hybrid Artificial Intelligence Approach

The best possible high detection rate, low false alarm rate and overcoming the disadvantages of intrusion detection techniques can be gotten by applying hybrid intelligence techniques. However, the work to make false alarm rate as low as possible is an ongoing affair. Different techniques, such as combination of neural network, machine learning, and Genetic algorithm technique can be used to hybrid intelligence approaches (Chitrakar & Huang, 2012; Wankhade et al., 2013).

2.7 Performance Evaluation

This section has discussed the intrusion datasets, the matrices of performance evaluation and evaluation formulas.

2.7.1 Intrusion Detection Dataset

The experiments for training and testing the proposed hybrid intelligent approach for network intrusion detection are applied by using a real dataset stream named intrusion detection dataset. These datasets contain a standard set of data to be audited, and they also include a wide variety of intrusion types simulated in a network environment. They are used as benchmark for security researches (Engen et al., 2011; Ibrahim et al., 2013; Raghuveer, 2012).

Generally, there are many types of benchmark datasets that are used for testing of any experiment of network intrusion detection, such as KDD'99, Kyoto 2006+, NSL-KDD, ISCX 2012 (Shiravi et al., 2012; Tavallaei et al., 2009; Yassin et al., 2013).

2.7.2 Evaluation Metric

The following factors which make a confusion matrix are often used to evaluate the intrusion detection accuracy and false alarm rate of ID whether true positives, true negatives, false positives, and false negatives. A true positive indicates that the intrusion detection approach detects precisely a particular attack having occurred. A true negative indicates that the intrusion detection approach has not made a mistake in detecting a normal condition. A false positive indicates that a particular attack has been detected by

the intrusion detection approach, but in reality, that type of attack did not actually occur. The following table shows a confusion matrix (Jawhar & Mehrotra, 2010).

Table 2.3:

Confusion Matrix

Actual	Predicated	
	Attack	Normal
Attack	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

It represents the accuracy of the detection system. If it is consistently high, that makes the approach to remain in a dangerous status. A false negative indicates that the intrusion detection approach is unable to detect the intrusion after a particular attack has occurred. The evaluation is performed by applying standard mathematic equations in terms of accuracy (*A*), detection rate (*DR*) and false alarm rate (*FAR*). As the equation (1), (2) and (3) (Chitrakar & Huang, 2012; Hameed & Sulaiman, 2012).

$$A = (TP+TN) / (TP+TN+FP+FN) \dots\dots\dots (1)$$

$$DR = (TP) / (TP+FP) \dots\dots\dots (2)$$

$$FAR = (FP) / (FP+TN) \dots\dots\dots (3)$$

Where,

TP = True Positive (attack detected as attack).

A = Accuracy.

TN = True Negative (normal detected as normal).

DR = Detection Rate

FP = False Positive (normal detected as attack).

FAR = False Alarm Rate

FN = False Negative (attack detected as normal).

2.8 Existing hybrid intelligent approaches

The best possible detection rate and accuracy can be done by using hybrid intelligent approaches. Many artificial intelligence techniques have been used to propose hybrid approaches (Chitrakar & Huang, 2012). Recently, many researchers have proposed hybrid different intrusion detection models and methods which are the combination of both of misuse and anomaly detection. Table 2.4 shows the related works and the proposed hybrid approaches to solve intrusion detection problem.

Chitrakar and Chuanhe (2012) propose a new combination to solve intrusion detection problems. In this approach, the requirement of big samples by the earlier methods is decreased by deploying SVM technique while maintaining the high quality clustering of k-Medoids. Their result shows that the proposed approach has increased the detection rate as well as has reduced the false positive rate. But the main drawback of this approach is that it has high time complexity because of k-Medoids clustering.

Upadhyaya and Jain (2013) suggest a novel method which combines the K-Medoids clustering and Naïve-Bayes classification. The proposed method has used clustering on data into a set, and after that the said method performs a classifier for classification purpose. Their result shows that the new method performs better in terms of CPU utilization, but it still has high false alarm rate.

Hameed et al. (2013) have proposed a hybrid method integrated both into Modified fuzzy Possiblistic C-Means (MFPCM) and symbolic fuzzy clustering in one algorithm called Extended Modified Fuzzy Possiblistic C-means (EMFPCM). Their result indicates that the proposed algorithm can recognize among natural and attack behaviors with reasonable detection rate.

Elbasiony et al. (2013) propose a hybrid detection framework, which depends on data mining classification and clustering techniques. Random forests classification algorithm is used to build intrusion patterns automatically from a training dataset. The k-means clustering is used to detect novel intrusions by clustering the network data to collect the most of intrusions together in one or more clusters. The disadvantages of that method are that, it works with specific network environment and the process is run offline.

Chapke Prajkta and Raut (2012) have suggested a novel method using hybrid model based on enhanced fuzzy and data mining techniques, which can detect both misuse and anomaly attacks. Then they have mentioned to use improved Kuok fuzzy data mining algorithm, which is in turn a modified version of APRIORI algorithm. A few

contributions have done in this study, such as improved priority algorithm for faster rule generation and reduced querying frequency. But it has many drawbacks, such as bottle neck in packet processing, searching for known rules and very high false positive rates.

Table 2.4:

Existing Hybrid Intelligent Approaches For Network Intrusion Detection

No.	Authors	Methods	Finding
1.	Danziger and de Lima Neto (2010).	Multi agent system, Danger theory and Bayesian classifier.	The findings of proposed approach are low number of negatives false alarms, ability to detect events not known to the system, and low cost of extra traffic upon the network. But it have high positive false alarms.
2.	Ghadiri and Ghadiri (2011).	Fuzzy C-Means, GK fuzzy clustering and RBF neural networks.	The combination provides more robustness to noise is expected due to both fuzziness and modularity of our approach.
3.	Ishida et al. (2011).	A modification of OptiGrid clustering and a cluster labelling algorithm using grids algorithm.	The finding of this study is improve the detection rate. The extraction process of training dataset and the reduction of parameters are the disadvantages.

No.	Authors	Methods	Finding
4.	Rangadurai et al. (2012).	Hidden Markov Model (HMM) based model with Naive Bayesian (NB).	The proposed method performed well in detecting intrusions. But the drawback of this model is that the difficulties of implementing HMM model in real time environment were described.
5.	Chitrakar and Chuanhe (2012).	Support Vector Machine Classification with k-Medoids Clustering.	Results show that it accomplished high detection rate as well as in decreasing the false positive rate. It consumes long time to perform detection.
6.	Husagic et al. (2013).	Principal Component Analysis, Fuzzy C Means and Nearest Neighborhood method.	The obtained results showed the system classification for five clusters was 35.7 %, and the best overall performance was accomplished with U2R attack by 58.8 % detection rate.
7.	Upadhyaya and Jain (2013).	K-Medoids clustering and Naïve-Bayes classification.	The proposed approach performs better in term of CPU utilization, but it still has high false alarm rate.

No.	Authors	Methods	Finding
8.	Elbasiony et al. (2013).	Framework based on random forests and weighted k-means.	It uses to build intrusion signatures using a training dataset, and classify the network behavior to the main kinds of intrusions. The disadvantages of this method are that, it works with specific network environment and the process run offline.
9.	Ahmad et al. (2013).	Software as a Service Intrusion Detection Services (SaaSIDS) and using hybrid analysis engine Artificial Immune System (AIS).	SaaSIDS is able to identify malicious activity and would generate appropriate alerts and notification accordingly.
10.	Chapke Prajkta and Raut (2012)	Fuzzy and data mining techniques.	A few contributions have done in this study such as improved priority algorithm for reducing query frequency and faster rule generation. But it has many drawbacks like very high false positive rates, bottle neck in packet processing, and search for known rules.

No.	Authors	Methods	Finding
11.	Liu Li et al. (2014).	Clustering algorithm and Hybrid Genetic Algorithm (HGA).	Enhanced Intrusion Detection Algorithm (EIDA) for intrusion detection in Ad hoc networks. It has the capacity to deal with different types of data and detection rate and false positive rate has been improved effectively.

The above table summarized the latest existing hybrid intelligent approaches which combine different AI techniques. It showed the finding, advantage and disadvantages of applying these methods. In a like manner of the proposed approach, the mentioned approaches have focused on the approaches which use clustering or classification techniques. However, the discussed approaches have indicated that the better results have come from combination of clustering and classification techniques together. Even though that, the results have shown some problems in the applied clustering and classification techniques.

On the clustering side, the utilized techniques such as OptiGrid, k-Medoids, fuzzy clustering, and others are sensitive to amount of input data and highly execution time. On the other sides, the classification techniques like NB, random forests, Bayesian, HMM and so on have shown that they have caused high false alarm rate. In addition, some of classification techniques work offline and environment dependency. Therefore,

the proposed hybrid intelligent approach combines the k-means clustering and SVM classification. The reasons of choosing that techniques are those, both of k-means and SVM are insensitive to amount of data, low time execution. As well, k-means is efficient data mining algorithm. While, SVM technique performs very well in terms of labeling data classification and applies in various computing environments. Hence, all of previous advantages for the proposed techniques lead to improve the detection accuracy and decrease the false alarm rate of classification.

2.9 Summary

This chapter has summarized key concepts for this study. It has dealt with the definition of network security and the type of attacks on network briefly. The importance of security today and the types of security that should be avoided is also mentioned. It has focused on the intrusion and explained its concepts. Also this chapter talks about the concepts of intrusion detection, intrusion detection system, the approaches of solving intrusion detection problem, the role of AI in intrusion detection problem and the main techniques which are used in this field. This chapter has shown and discussed the existing hybrid intelligent approaches for network intrusion detection and their limitations.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter covers research methodology, which proposes to solve the research problem. Research methodology is adapted from Chitrakar and Huang (2012) and Jawhar and Mehrotra (2010). It has four phases which are: Selection of Experiment Data, Data Pre-Processing, Classification, and Performance Evaluation. Each phase of the methodology highlights the operation to undertake as shown in Figure 3.1 below:

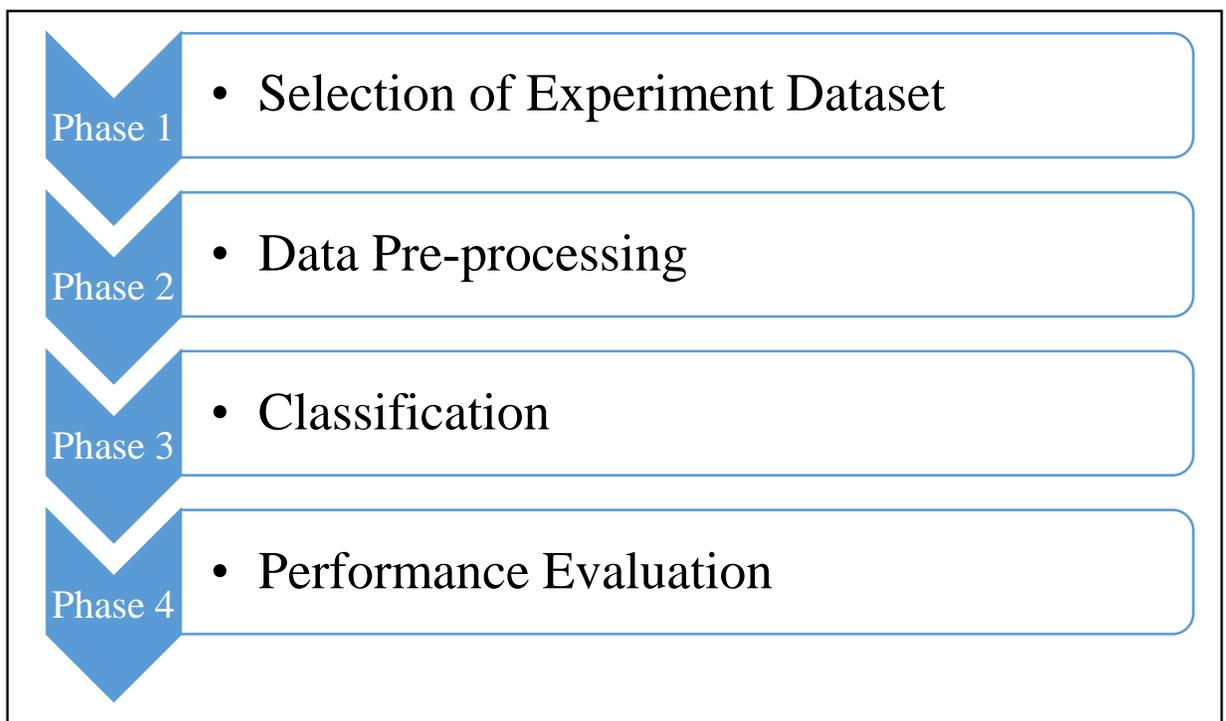


Figure 3.1: Research Methodology Phases

In the first phase, the appropriate and better dataset which has the biggest variety of attack types and it is mostly recommended from other studies and researchers will select it for testing a suggested hybrid intelligent approach. This phase's achievement depends on literature review and the previous studies. In the second phase, a network packets and experiment dataset, which is selected from pervious phase, will prepare for classification phase. The data preprocessing phase will be applied by four steps, and they are data transformation, feature selection, data normalization and data clustering.

In the third phase, a classification phase will classify all the network traffic into natural or intrusion behavior and assign each attack to its specific category. The network attacks fall in four categories, and they are DoS, U2R, R2L and Probing. Lastly, in the fourth phase, a result and performance of the proposed hybrid method will be evaluated. The evaluation has two steps; the first step is applied by using mathematical equations and the other step will be done by comparing the result of the proposed hybrid intelligent approach with the results of existing hybrid intelligent approaches.

3.2 Phase I: Selection of Experiment Dataset

The experiments for training and testing the proposed hybrid intelligent approach for network intrusion detection is applied by using a real dataset stream named as intrusion detection dataset. These datasets contain a standard set of data to be audited, and that datasets include a wide variety of intrusion types simulated in a network environment (Revathi & Malathi, 2014).

To validate the efficiency and accuracy of the proposed hybrid intelligent approach, NSL-KDD intrusion dataset has chosen. It is a new version of KDD'99 dataset. NSL-KDD dataset has some advantages over KDD'99 dataset and other intrusion dataset. It has solved several inherent problems of the KDD'99, and it is considered as the standard benchmark for intrusion detection evaluation. The intrusion dataset of NSL-KDD consist of approximately 150,000 single connection vector; each of which contains 41 features and is labeled as either normal or abnormal type with exactly one of the specific attack type (Bhavsar & Waghmare, 2013; Panwar et al., 2014).

NSL-KDD has many advantages comparing with other datasets. It has been chosen for testing the proposed hybrid intelligent approach for intrusion detection due to following reasons.

- i. The similar connection vectors from training dataset were removed.
- ii. The duplicate connection vectors in the testing dataset were deleted for improving detection performance.
- iii. The classification by using NSL-KDD dataset gives an accurate evaluation of various machine learning algorithms.
- iv. The usage of NSL-KDD dataset is affordable for experiment purposes. Also, it has a reasonable numbers instances and good variety of attack types both in the training and testing dataset.

Tables 3.1 shows the NSL-KDD dataset features as below :

Table 3.1:

List of Attributes in NSL-KDD Dataset.

No.	Feature Name	No.	Feature Name	No.	Feature Name
1.	Duration	2.	Protocol-type	3.	Service
4.	Flag	5.	Src-bytes	6.	Det-bytes
7.	Land	8.	Wrong fragment	9.	Urgent
10.	Hot	11.	Num-failed-login	12.	Logged-in
13.	Num-compromised	14.	Root-Shell	15.	Su-attempted
16.	Num-Root	17.	Num-File-Creation	18.	Num-shell
19.	Num-access-file	20.	Num-outboundcmds	21.	Is-hot-login
22.	Is-guest-login	23.	Count	24.	Srv-count
25.	Serror-rate	26.	Srv-serror-rate	27.	Rerror-rate
28.	Srv-error-rate	29.	Same-srv-rate	30.	Diff-srv-rate
31.	Srv-diff-host-rate	32.	Det-host-count	33.	Dst-host-srv-co
34.	Dst-host-same-srvrate	35.	Dst-host-diff-srvrate	36.	Dst-host-same
37.	Dst-host-diff-srvhost-rate	38.	Dst-host-serror-rate	39.	Dst-host-srv-rate
40.	Dst-host-error-rate	41.	Dst-host-srv-rer-rate		

Features 1-9 stands for the basic features of a packet, 10-22 features for content features, 23-31 features for traffic features and 32-41 features for the host features. There are 38 different types of attacks in training and testing dataset combined. Simulated attacks are grouped into four categories, namely, DoS, R2L, U2R, and Probing. DoS and Probing attacks come with greater frequency and can be easily separated from normal activities. In contrast, U2R and R2L attacks are embedded in the data portions of the packet, thus; it is difficult to achieve detection accuracy for those two types of attacks. Intruders attempt to get access from a remote computer by exploiting the holes in local machine and embed their attack with transferred packages in the case of R2L attacks. Whilst in the case of U2R attacks, intruder has legal access to the victim machine and attempts to use vulnerabilities of the system so as to gain an access to a root of the system and super user privileges. Attacks are grouped into following four categories as shown in Table 3.2.

Denial of Service (DoS): such as back, land, neptune etc.

Remote-to-Local (R2L): such as imap, sendmail, phf etc.

User to Root (U2R): such as buffer_overflow, sqlattack

Probing: such as mscan, saint, satan etc.

Table 3.2:
Attack Categories.

Attack types	Category	Attack types	Category
Normal	Normal	ftp_write	R2L
apache2	DoS	guess_passwd	
back		imap	
land		multihop	
mailbomb		named	
neptune		phf	
pod		sendmail	
processtable		snmpgetattack	
smurf		snmpguess	
teardrop		spy	
udpstorm		warezclient	
buffer_overflow		warezmaster	
httptunnel	worm		
loadmodule	xsnoop		
perl	U2R	ipsweep	Probe
ps		mscan	
rootkit		nmap	
sqlattack		portsweep	
xterm		saint	
		satan	

A sample of original NSL-KDD data set connection record is shown in Figure 3.2

```
0, tcp, netbios_ns, S0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 96, 16,  
1.00, 1.00, 0.00, 0.00, 0.17, 0.05, 0.00, 255, 2, 0.01, 0.06, 0.00, 0.00, 1.00,  
1.00, 0.00, 0.00, neptune  
-----  
0, tcp, http, SF, 300, 13788, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 8, 9,  
0.00, 0.11, 0.00, 0.00, 1.00, 0.00, 0.22, 91, 255, 1.00, 0.00, 0.01, 0.02, 0.00,  
0.00, 0.00, 0.00, normal
```

Figure 3.2: The Original NSL-KDD Dataset Connection

3.3 Phase II: Data Pre-Processing

Pre-processing of original NSL-KDD intrusion data set is an important phase to make it as an appropriate input for classification phase. Based on the result of vulnerability assessment and network topology, hosts exists, and services running, intrusion detection process needs to configure preprocessing phase. The main objective of preprocessing phase is to reduce ambiguity and provide accurate information to detection engine. So, here we have presented preprocessing phase which cleans network data, grouping, labeling, and the handles missing or incomplete dataset (Davis & Clark, 2011; Li & Yuan, 2010; Teng et al., 2010).

IDSs usually have high-dimensional datasets, furthermore; intrusion datasets should be suitable and compatible by transforming all the dataset for uniform type, selecting the effective and relevant features for intrusion detection purpose, and normalization. Additionally, grouping and labeling the dataset before classification phase. Preprocessing phase helps to reduce the false alarm rate and improve detection accuracy. Dataset pre-processing has achieved the following by applying the following stages sequentially:

- i. Dataset transformation.
- ii. Feature Selection
- iii. Dataset normalization.
- iv. Dataset clustering.

i. Dataset Transformation

The training and testing dataset of NSL-KDD consist of approximately 150,000 single connection instances. Each connection instance contains 41 features, including attacks or normal. There are many symbolic attributes like flag, services types, and protocol types. These attributes have nominal values, such as RSTOS0, private, icmp in the dataset. Under this step, some useless data will be filtered and modified. For example, some text items need to be converted into numeric values. Hence, one needs to transform these nominal values to numeric values beforehand to make it suitable input for classification phase using SVM.

For this transformation stage, Table 3.3 has used to transform all the nominal values of dataset features into the numeric values. For instance, the flag type of “OTH” is transformed to 1, “REJ” is transformed to 2, “RSTO” is transformed to 3 and so on. Also, the numeric value has to assign last feature in the connection instance, which is the target class. For doing this, we have assigned a target class “0” for normal connection and class “1” for any deviation from that (i.e. if that is an attack) as per transformation Table 3.3.

Table 3.3:

Transformations Table

Type	Feature Name	Numeric Value
Attack or Normal	Normal	0
	Attack	1
Protocol Type	TCP	1
	UDP	2
	ICMP	3
Flag	OTH	1
	REJ	2
	RSTO	3
	RSTOS0	4
	RSTR	5
	S0	6
	S1	7

	S2	8
	S3	9
	SF	10
	SH	11
Services	All Services	1 to 70

ii. Features Selection

Features selection is the most critical stage in building intrusion detection models and is equally important to improve the efficiency of data mining algorithms. In general, the input data to classifiers is in high dimension feature space, but not all of the features are relevant to the classes to be classified. Most of the data includes irrelevant, redundant, or noisy features. In these cases, irrelevant and redundant features can introduce noisy data that distract the learning algorithm, even degrade the accuracy of the detector and causing slow training and testing processes (Maldonado et al., 2011; Stein et al., 2005).

Features selection is the operation for choosing a subset of original attributes, which depend on specific criteria, and it is a necessary and common used technique in data mining for dimension reduction. It decreases the number of attributes, eliminates irrelevant, noisy, or redundant features, and brings about palpable effects on applications, such as speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility. During this step, the set of attributes or features deemed to be the most effective attributes, which are extracted

in order to construct suitable detection system (detectors). The goal of features selection increases the detection rate and decreases the false alarm rate in network intrusion detection (Chae et al., 2013; Maldonado et al., 2011).

In this study, WEKA 3.7, which is a machine learning tool has been used to compute the features selection subsets for SVM classifier to test the classification performance on each of these feature sets. ClassifierSubsetEval and BestFirst algorithms have been applied, and this study has used all training dataset and 10-fold cross validation for this purpose. In 10-fold cross-validation, the training dataset is randomly separated into 10 individual subsets, which have almost same size. Then, a single of the subsets is applied as the test dataset and the other nine subsets are utilized for building the classifier. The test set is used to estimate the accuracy, and the accuracy estimate is the mean of the estimates for the classifier.

iii. Dataset Normalization

Dataset normalization is a preprocessing type, which plays an important role in classification. It is a good way to reduce the difference of the data and improve the speed. So, normalizing the input data will help speed up the learning phase. Normalization is an important stage to enhance the performance of intrusion detection system when datasets are too large. Some kind of data normalization also may be necessary to avoid numerical problems, such as precision loss from arithmetic overflows. Attributes with initially large ranges will outweigh attributes

with initially smaller ranges, and then dominate the distance measure (Z. Liu, 2011; Wang et al., 2009).

Min-Max Normalization applies in a linear transformation on the original dataset x into the specified interval. This method scales the data from (X_{\min}, X_{\max}) to (New_{\min}, New_{\max}) in proportion. The advantage of this method is that it preserves all relationships of the data values exactly. It does not introduce any potential bias into the data (Z. Liu, 2011).

$$X_{new} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (4)$$

iv. Data Clustering (K-Means)

Clustering technique groups a set of data that exhibit similar characteristics into meaningful subclasses (known as clusters) according to some pre-defined metrics, and there should be high intra cluster similarity and low inter cluster similarity, i.e. items within same cluster are more “similar” to each other than they are to items in the other clusters. A clustering method which results in such type of clusters is considered as good clustering algorithm (Al-Jarrah & Arafat, 2014; Panwar et al., 2014).

K-means clustering is a well-known data mining algorithm, and it has been used in an attempt to separate the user behavior into anomalous and normal and make them as labeled groups, and group unusual behavior in network traffic as well. K-Means is an unsupervised approach, which overcomes the problem of the lack of training datasets with known intrusions. After an initial random assignment of an example to K clusters, the centers of clusters are computed and the examples are assigned to the clusters with the closest centers. The process is repeated until the cluster centers do not significantly change. when the cluster assignment is fixed, the mean distance of an example to cluster centers is used as the score. A set of n vectors X_j , $j = 1, \dots, n$, are to be partitioned into C groups G_i , $i=1, \dots, C$. The cost function, based on the Euclidean distance between a vector x_k in group j and the corresponding cluster centre C_i , can be defined as: (Gao & Wang, 2014; Yassin et al., 2013).

$$J = \sum_{i=0}^c J_i \sum_{i=1}^c \left[\sum_{k, x \in G_i} \|X_k - C_i\|^2 \right] \quad (5)$$

By applying the K-means clustering algorithm, two clusters were specified and created for each output class. As the algorithm iterates through the training data, each cluster's architecture is updated. In updating clusters, elements are deleted from one cluster and transferred to another. The updating of clusters causes the values of the centroids to modify. This change is a reflection of the current cluster elements. When, there are no changes to any cluster, the clustering of the K-Means algorithm becomes complete.

At the end of the K-Means clustering, the K cluster centroids are created, and the algorithm is ready for classifying traffic. For each element to be clustered, the cluster centroids with the minimal Euclidean distance from the element will be the cluster for which the element will be a member.

The algorithm consists of the following steps:

Input: the intrusion dataset for intrusion detection and K value which presents number of clusters.

Output: A different set of K-clusters that minimizes the squared error criterion.

Algorithm:

- 1.** Initialize K clusters (randomly select k elements from the data).
- 2.** Initialize the k cluster centroids. This can be done by arbitrarily dividing all objects into k clusters, computing their centroids, and verifying that all centroids are different from each other. Alternatively, the centroids can be initialized to k arbitrarily chosen different object.
- 3.** Iterate over all data points in the data set and compute the distances to the centroids of all clusters. Assign each data point to the cluster with the nearest centroid.
- 4.** Re calculate “k” new centroids as per centers of the clusters resulting from the previous step.
- 5.** Repeat step 3 until the centroids do not change any more.

3.4 Phase III: Classification

A classification based IDS will classify all the network traffic into natural or intrusion behavior and assign each attack to its specific category. The network attacks fall into four categories, and they are DoS, U2R, R2L and Probing. Classification process consists of two steps. The first step is learning by training phase, and the second step is classification. In the learning step a classifier is formed and in the classification step, that model is used to predict the class labels for a given data. In this phase, processed data from previous phase are classified as normal or attack and assign every attack to its type. The classification purpose has applied and used supervised machine learning classification technique, which is support vector machine.

Support vector machine (SVM) is a supervised learning model; and it is one of the best techniques, and it is widely used in machine learning tasks. It plots the training vectors in high dimensional feature space, assign each vector by its class. It classifies data by determining a set of support vectors, which are the members of the set of training inputs that outline a hyper plane in the feature space. SVMs provide a generic mechanism to fit the surface of the hyper plane to the data via the use of a kernel function. In the standard supervised learning, we are given n training samples (x_i, y_i) $i=1, 2, \dots, n$, where $x_i \in X$ denotes the input vector, and $y_i \in Y$, $y_i \in \{+1, -1\}$ denotes the corresponding output value. The typical classification function of SVM usually is (RavinderReddy et al., 2014; Song et al., 2011).

$$f(x) = w \cdot \phi(x) + b \quad (6)$$

ω denotes the weight vector and b denotes the bias term. In order to construct a hyper plane that has the smallest number of errors, we introduce the non-negative slack variables $\xi_i \geq 0, i = 1, 2, \dots, n$, and the penalty parameter C , which denotes a cost function for measuring the empirical risk and is determined by the user. The coefficients ω and b are estimated by minimizing the following regularized risk function:

$$\text{Max} \sum_{i=1}^n \xi_i - \frac{1}{2} \sum_{i,j=1}^n \xi_i \xi_j y_i y_j K(x_i, x_j) \quad (7)$$

The main advantage of SVM is the low expected probability of generalization errors. There are several other reasons for selecting SVM for network intrusion detection. The first is performance speed: as real-time environment performance is of primary importance to intrusion detection systems. Any classifier that can potentially run “fast” is worth considering. The second reason is scalability; SVMs are relatively insensitive to the number of data points; and the classification complexity does not depend on the dimensionality of the feature space, so they can potentially learn a larger set of patterns, thus; it will be able to scale better than neural networks (Chang & Lin, 2011; Maldonado et al., 2011; Sung & Mukkamala, 2003).

3.5 Phase VI: Performance Evaluation

Network intrusion detection methods are usually applied to distinguish between anomalous and normal traffic. So, here we are interested in performance measures which are applied in classification. The evaluation of the proposed hybrid intelligent approach has two steps; the first step is applied by using mathematical equations, and the other step is done by comparing the result of the proposed hybrid intelligent approach with the result of the existing hybrid intelligent approaches.

The first stage is performed by applying standard mathematic equations in terms of accuracy (A) which is the metric, metric indicates the total number of connections that are correctly classified, including normal and intrusive connections (Equation 8), detection rate (DR) that is the amount of attack detected when it is actually attack / the amount of attack sample and false alarm rate (Equation 9), (FAR) which is the amount of attack detected when it is actually normal / the amount of normal sample (Equation 10) (Chitrakar & Huang, 2012; Hameed & Sulaiman, 2012).

$$A = (TP+TN) / (TP+TN+FP+FN) \dots\dots\dots (8)$$

$$DR = (TP) / (TP+FP)\dots\dots\dots (9)$$

$$FAR = (FP) / (FP+TN)\dots\dots\dots (10)$$

Where,

TP = True Positive (attack detected as attack).

A = Accuracy

TN = True Negative (normal detected as normal).

DR = Detection Rate

FP = False Positive (normal detected as attack).

FAR = False Alarm Rate

FN = False Negative (attack detected as normal).

This study uses confusion matrix method to present the classification results. The confusion matrix is the best way of presenting the binary classification results. The following factors are often use to evaluate the detection accuracy and false alarm rate of ID in confusion matrix which are: true positives, true negatives, false positives, and false negatives. While, in second step, the detection rate and false alarm rate of proposed hybrid intelligent approach will compare to existing hybrid intelligent approaches through the literature reviews.

3.5.1 Confusion Matrix

Confusion matrix is the best way of presenting the binary classification result. It is a visualization tool for tabulating the overall performance of the classier. Each row of the matrix represents the instances in a predicted class, on the other hand, each column represents the instances in an actual class. The following factors are often used to evaluate the detection accuracy and false alarm rate of ID: true positives, true negatives, false positives, and false negatives. Table 3.5 bellow shows the factors of confusion matrix and the relationship between these factors.

Table 3.5:

Confusion Matrix.

Actual	Predicated	
	Attack	Normal
Attack	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

A true positive indicates that the intrusion detection approach detects precisely a particular attack have occurred. A true negative indicates that the intrusion detection approach has not made a mistake in detecting a normal condition. A false positive indicates that a particular attack has been detected by the intrusion detection approach, but such an attack did not actually occur. It represents the accuracy of the detection system. If it is consistently high, that makes the approach remain in a dangerous status. A false negative indicates that the intrusion detection approach is unable to detect the intrusion after a particular attack has occurred.

For the sake of comparison, some performance measures extract part of the information from the confusion matrix and produce some numeric values that are easily comparable. In addition, there are some other metrics derived directly from the confusion matrix values (Tavallae, 2011).

3.6 Summary

This chapter has presented the research methodology processes for solving the problem of this study. It contains four phases, which are selection of experiment dataset, data preprocessing, classification, and performance evaluation. NSL-KDD dataset has been selected for testing the proposed hybrid intelligent approach. The dataset was described in details. Then, data pre-processing phase which consist of four stages that have been applied on NSL-KDD dataset to prepare it for the classification phase. These stages are, firstly, dataset transformation which has uniformed the types of dataset to one type (numeric) to improve the performance of machine learning algorithms. Secondly, features selection process has performed to select relevant feature for intrusion detection purpose. It has chosen 21 from 42 features which are most effect on intrusion detection.

Thirdly, data normalization which has reduced the difference of the data and improves the speed by ranging it between (0, 1) values using Min-Max formula. Lastly, dataset clustering for grouping the dataset is for two main types; that are normal and abnormal. K-Means algorithm was used for clustering purpose. In addition to that, Classification phase is performed SVM technique to classify network behavior to normal or attacks and identify the types of attacks. Finally, Performance evaluation factors and methods for the proposed hybrid intelligent approach was used mathematic equations, confusion matrix.

CHAPTER FOUR

HYBRID INTELLIGENT APPROACH DESIGN

4.1 Introduction

This chapter presents the design of the proposed hybrid intelligent approach for network intrusion detection. It shows the workflow of the proposed hybrid intelligent approach and it explains the machine learning algorithms which are used in the proposed approach in details. In addition, the advantages of these algorithms are also discussed. Various artificial intelligence techniques have been utilized in network intrusion detection process; different classifiers, such as combination of artificial intelligence techniques have been used to hybrid learning approaches. In this study, the proposed hybrid intelligent approach combines k-means clustering and SVM classification.

4.2 Approach Design

A clustering and classification ensemble algorithm for intrusion detection are proposed to achieve the high speed, high detection rate, and low false alarm rate. A clustering algorithm, which has used k-means divides and labels the data for the corresponding groups before applying a classifier technique to classification purpose. on the other hand, a classification algorithm that uses SVM technique for intrusion detection is proposed in which when the K-means finishes its process. The learning ensemble based on evidence accumulation algorithm, overcomes the disadvantages of K-means and SVM algorithms. Figure 4.1 below the workflow of proposed approach in details.

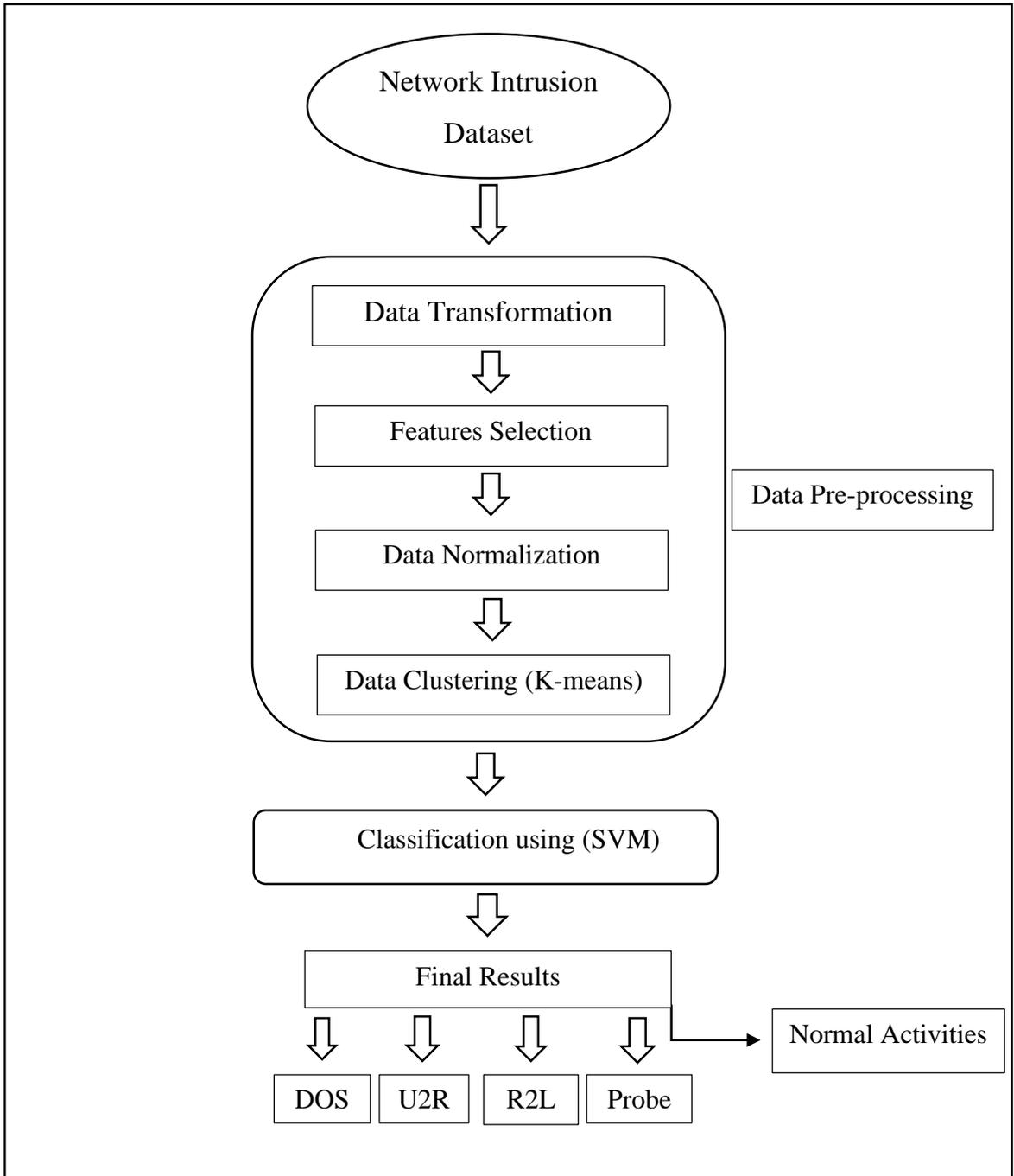


Figure 4.1: Workflow of Proposed Hybrid Intelligent Approach

4.3 Clustering

Data clustering is a popular technique for intrusion detection. Nowadays, the manual labeling is not useful, expensive and time consuming because of the huge amount of network data available. Clustering is the process of grouping, labeling the data and determining it into sets of similar objects.

Every set is named as cluster. It contains a number of the similar clusters that are same and numbers from the different clusters are different from another clusters. Hence, clustering techniques are useful for detecting intrusions from network data. Clustering techniques can discover complicated intrusions from various time period. Clustering is a machine learning technique to detect attacks from unlabeled data with many dimensions. They are techniques that handle the unlabeled data (Bahrololum & Khaleghi, 2008; Wankhade et al., 2013).

Labelling and grouping of dataset are important, and natural patterns in the data are extracted. Most of clustering algorithms are unsupervised and this key advantage of clustering and are commonly applied to anomaly detection. This generally obtains better results on benchmark data.

There are two type of clustering namely: hierarchical and partitional clustering. A hierarchical clustering performs a nested sequence of partitions by either an agglomerative (bottom-up) or divisive (splitting or top-down) approach. The agglomerative approach starts by placing each object in its own cluster and then merges them into larger and larger cluster until all objects are in one cluster. The divisive approach reverses the process, it starts with one cluster contains all of the objects, proceeds by splitting the single cluster up into smaller sized clusters. While, Partitional clustering essentially deals with the task of partitioning a set of entities into a number of homogeneous clusters, with respect to a suitable similarity measure. Partitional clustering tries to optimize certain criteria. They start from a given group definition and proceed by exchanging elements between groups until a certain criteria is optimized. Typically, the criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters (Bhuyan et al., 2013; Jain et al., 2011; Panda & Patra, 2008). Some advantages of using clustering are given below

- i. For a partitioning approach, if k can be provided accurately then the task is easy.
- ii. Incremental clustering (in supervised mode) techniques are effective for fast response generation.
- iii. It is advantageous in case of large datasets to group into similar number of classes for detecting network anomalies, because it reduces the computational complexity during intrusion detection.

In our proposed hybrid intelligent approach, we choose K-means as a clustering technique.

K-Means Clustering

K-means is one of well-known data mining clustering algorithms. K-means has been performed in an attempt to detect abnormal network user behavior, and novel behavior in network traffic as well. After an initial random assignment of the examples to K clusters, the centers of clusters are calculated and the samples dataset are allocated to the clusters with the closest centers. This process is repeated until the cluster center does not significantly change. Once the cluster assignment is fixed, the mean distance of an example to cluster centers is used as the score. Using the K-means clustering algorithm, different clusters are specified and generated for each output class (Panda & Patra, 2008; Yang & Ning, 2010).

The difficulty of recognizing among normal and intrusion behavior in network systems is a major problem with clustering techniques because of the big overlapping in data monitoring. The detection processing causes false alarms resulting in the intrusion detection based on the anomaly intrusion detection system. The main aim of applying the K-Means clustering algorithm is to separate the set of normal and abnormal data that behave similarly for different partitions which are known as K -th cluster centroids (Jawhar & Mehrotra, 2010; Yassin et al., 2013).

4.4 Classification

Classification is a machine learning technique that deals with every instance of data set and classifies it to a specific class. It extracts the models for defining important data classes. Such types of classes are called as classifiers. A classification based on IDS will classify all the network traffic into natural or intrusion behavior. Classification process consists of two steps. The first step is learning by training phase, and the second step is classification. In the learning step a classifier is formed and in the classification step that model is used to predict the class labels for a given data.

The analysis of classification needs that the analyst knows prior of time how classes are defined. Every record available in the data set already had values to the attributes applied for defining the classes. Classification is a supervised machine learning mechanism. It can handle only labeled data. So, the major disadvantage of classification technique is that, it is less efficient in the field of intrusion detection as compared to clustering because classification cannot handle unlabeled data, which degrades the performance of intrusion detection system (Wankhade et al., 2013).

This problem will be solved in the proposed hybrid approach by apply clustering data before making classification to increase the efficiency of classification technique. In the proposed hybrid intelligent approach, support vector machine have been chosen as a classification technique.

Support Vector Machine

Nowadays, SVM technique is mature enough to apply for different classification problems. The SVM transforms data into a feature space F which usually have a big dimension. It interests for noting that SVM generalization bases on the geometrical characteristics of the training data, not on the dimensions of the input space. Training a SVM leads to a quadratic optimization problem with bound constraints and one linear equality constraint.

The main benefits of using SVM to an intrusion detection problem lie in that the system takes an accurate detection model from a mass of audit data automatically for decreasing artificial intervention and it can be used to construct an intrusion detection system in various computing environments because of universality of mining process itself processing techniques. The first reason for using SVM is its performance in terms of execution speed. The other cause is scalability: SVMs are comparatively insensitive to the amount of data points and the complexity of classification does not depend on the dimensionality of the feature space (Chitrakar & Chuanhe, 2012; Mohammad et al., 2011).

4.5 Summary

This chapter has discussed the design of the proposed hybrid intelligent approach for network intrusion detection. It has shown the workflow of the proposed approach in details. The proposed approach has integrated two machine learning techniques, which are k-means clustering and SVM classification techniques. This chapter has explained in detail the clustering and classification work and their advantages. In addition to that, it has presented the current problems in the existing hybrid intelligent approaches which combine clustering or classification techniques. Also, it has shown the k-means algorithm and SVM in details and their advantages. As well as, the ability of these two techniques in overcoming the current drawbacks in the existing hybrid intelligent approaches.

CHAPTER FIVE

EXPERIMENTAL RESULTS AND EVALUATION

5.1 Introduction

This chapter shows the experimental results for classification of the proposed hybrid intelligent approach for network intrusion detection. Confusion matrix uses to present the results. In addition, the process of evaluation has done and shown and made the comparison with an existing intelligent approach for network intrusion detection. The NSL- KDD dataset are taken to test the proposed hybrid intelligent approach for network intrusion detection. All experiments have been performed using Intel Core i5-2410M, 2.3 GHz processor with 8 GB of RAM and Waikato Environment for Knowledge Analysis (WEKA) machine learning software.

5.2 Preprocessing Results

In the final analysis, the testing of the proposed approach was done as follows: the full NSL-KDD dataset which has 150,000 approximately connection records has undergone the transformation, feature selection, normalization process, and clustering. The experimental results are presented based on every step of preprocessing phase. Accordingly, the transformation step which converts network data to uniform type firstly. The NSL - KDD dataset has some symbolic attributes like flag, service types, and protocol types.

Hence, these nominal values have been transformed to numeric values beforehand to make the data suitable input for classification phase. After transformation process, the feature's value of original NSL- KDD dataset has become as shown in Figure 5.1.

<p>0, 1, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 96, 16, 1.00, 1.00, 0.00, 0.00, 0.17, 0.05, 0.00, 255, 2, 0.01, 0.06, 0.00, 0.00, 1.00, 1.00, 0.00, 0.00, 1</p>
<p>0, 1, 2, 10, 300, 13788, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 9, 0.00, 0.11, 0.00, 0.00, 1.00, 0.00, 0.22, 91, 255, 1.00, 0.00, 0.01, 0.02, 0.00, 0.00, 0.00, 0.00, 0</p>

Figure 5.1: The NSL-KDD Dataset Connection After Transformation

Afterwards, for the purpose of decreasing the number of attributes, eliminates irrelevant, noisy, or redundant features, the process of feature selection using ClassifierSubsetEval and BestFirst algorithms have been applied. As a result of features selection step, the features of the input data were reduced from 41 to 21. This step is speeding up a classification phase and improving classification accuracy. Table 5.1 shows the result of features selection process which selected 21 features from NSL-KDD dataset for network intrusion detection purpose.

Table 5.1:

The Result of Features Selection Process

No.	Features	No.	Features
2.	Protocol_type	3.	Services
4.	Flag	5.	Src_byte
6.	Dst_byte	8.	Wrong_fragment
12.	Logged_in	23.	Count
24.	Src_count	25.	Serror_rate
26.	Srv_serror_rate	29.	Same_srv_rate
30.	Diff_srv_rate	32.	Dst_host_count
33.	Dst_host_srv_count	34.	Dst_host_same_srvrate
35.	Dst_host_diff_srvrate	36.	Dst_host_same_srvhost_rate
37.	Dst_host_diff_srvhost_rate	38.	Dst_host_serror_rate
41.	Dst_host_srv_rer_rate		

Subsequently, the normalization step has been applied for the selected features from previous step using Min-Max formula. It has used to reduce the difference of the data, improve the speed and enhance the performance of an intrusion detection process. Transforms data to have a mean value of zero, so that outliers can be easily detected, the results of normalization step has shown in Figure 5.2.

0.0039, 0.0039, 0.0235, 0.0000, 0.0000, 0.0000, 0.000, 0.3764, 0.0627, 0.0039, 0.0039, 0.0006, 0.0001, 1.0000, 0.0078, 0.0000, 0.0002, 0.0000, 0.0000, 0.0039, 0.0000
0.00007, 0.0001, 0.0007, 0.0217, 1.0000, 0.0000, 0.00007, 0.0005, 0.0006, 0.0000, 0.0000, 0.00007, 0.0000, 0.0065, 0.0184, 0.00007, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000

Figure 5.2: The NSL-KDD Dataset Connection After Normalization

Eventually, the clustering step using K-Means algorithm has performed for grouping the network dataset for two clusters (groups) as (normal and anomaly). The results of clustering have grouped 49% of network dataset as a normal behavior and the other 51% as an anomaly (attack) behavior. Figure 5.3 below show the clustering results of testing of the proposed hybrid intelligent approach using K-Means algorithm.

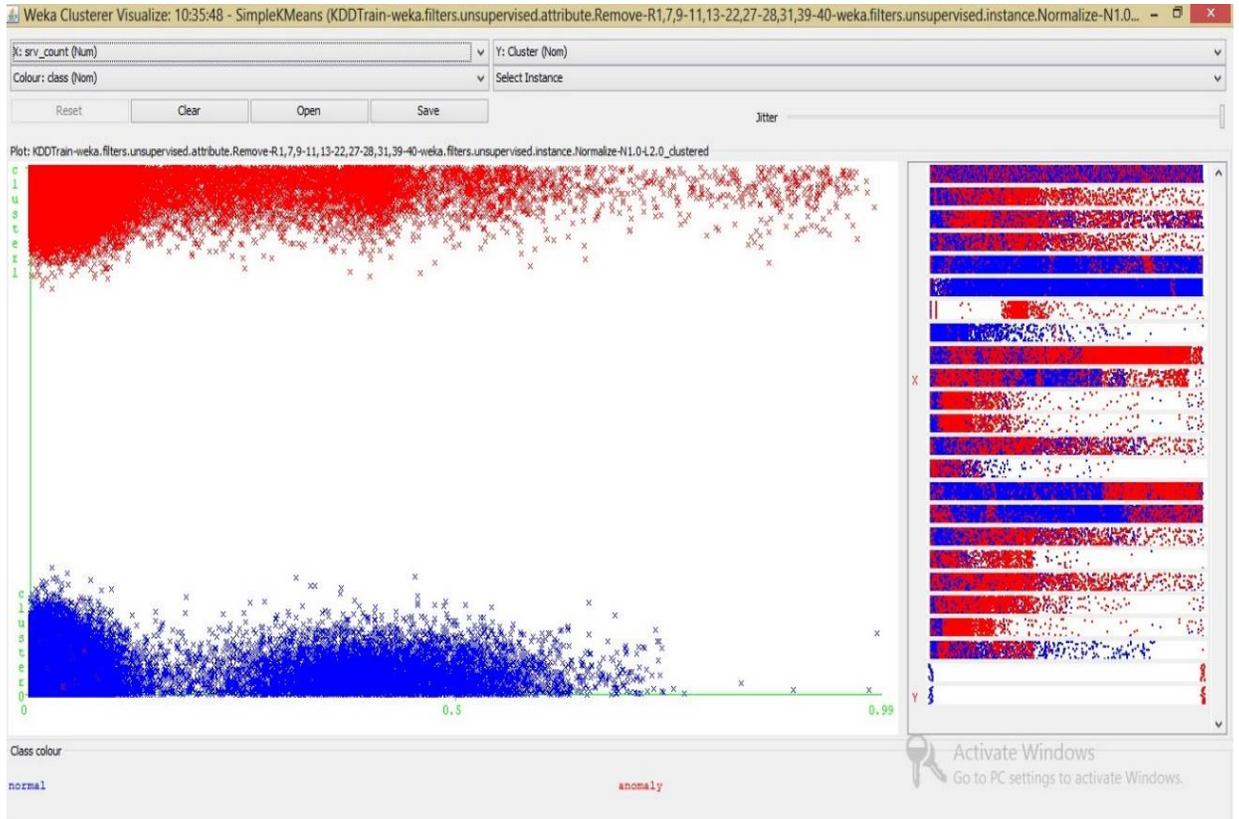


Figure 5.3: Clustering Results.

5.3 Classification Results

After applying all the preprocessing steps, the classification phase has been performed using SVM technique. SVM classification divides the network behavior to normal and abnormal and assign the attack behavior to its category. The confusion matrix has obtained from the classification of the proposed hybrid intelligent approach using full NSL-KDD intrusion dataset. From 148,517 connection instances of full NSL-KDD dataset, Table 5.2 shows the obtained confusion matrix by connection records of testing of the hybrid intelligent approach for network intrusion detection.

Table 5.2:

Confusion Matrix for Classification (number of connection records).

Actual	Predicated	
	Attack	Normal
Attack	68,436	3,027
Normal	1,070	75,984

Depend on the obtained confusion matrix for classification of the proposed approach, the percentages of the true positive rate (TP), false positive rate (FP), false negative rate (FN), and true negative rate (TN) of hybrid intelligent approach detection have been calculated as shows in Table 5.3. Obviously, the following table of result has shown the high rate of detection. The approach has detected 95.76 percentage as attack from 71,463 real attack connection records. While, the other 4.23 percentage as normal. Nevertheless, the full number of normal connections of records in the NSL-KDD dataset which is 75,984 have been classified as 96.28 percentage as normal and 3.71 percentage of the connection as an attack. Table 5.3 presents the results of classification as percentages of the true positive rate, false positive rate, false negative rate, and true negative rate of hybrid intelligent approach classification.

Table 5.3:

Confusion matrix for classification of proposed approach.

Actual	Predicated	
	Attack	Normal
Attack	95.766% (TP)	4.237% (FP)
Normal	3.715% (FN)	96.284% (TN)

The details of detection for the proposed hybrid intelligent approach have shown that, the higher the detection for DOS attacks, as well, the other category of attacks which is R2L and Probe. In spite of the highly assigning rate for last three categories, the proposed hybrid intelligent approach shows a lack in U2R attack detection. Figures 5.4 bellow shows the detection rate for every category of attack, as well as, the normal network behavior.

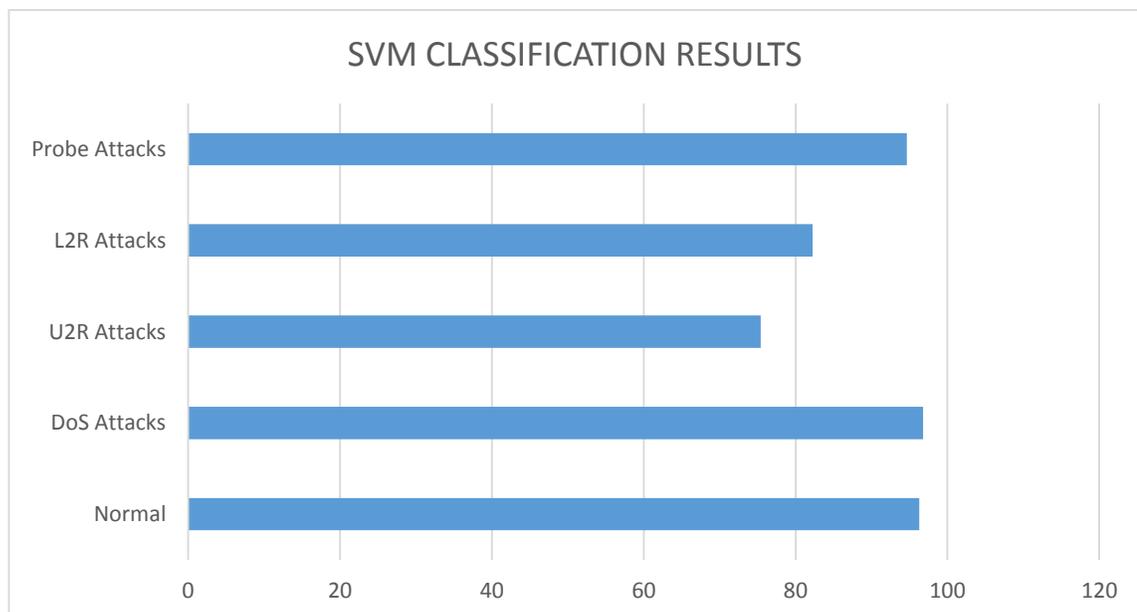


Figure 5.4: Detection Rate for Attack Categories.

5.4 Performance Evaluation

The performance evaluation of the proposed hybrid intelligent approach consists of two stages. Firstly, it has applied using mathematical equations and secondly, it has done by making a comparison between the result of the proposed approach and an existing hybrid intelligent approaches. Depend on the true positive rate (TP), false positive rate (FP), false negative rate (FN), and true negative rate (TN) values, which obtained from the classification of the proposed hybrid intelligent approach for network intrusion detection, which combine the K-means clustering and SVM classification algorithms, the terms of accuracy, detection rate, and false alarm rate will calculate using the mathematical equations (which explained in section 3.4 from the previous chapter).

The accuracy (A), which is the metric indicates the total number of connections that are correctly classified including normal and intrusive connections. Secondly, the detection rate (DR), is the amount of attack detected when it is actually attacked. Lastly, the false alarm rate (FAR), which is the amount of attack detected when it is actually normal. Table 5.4 presents the results of accuracy, detection rate, and false alarm rate as follows.

Table 5.4

Result of Performance Evaluation

Metric	Formula	Values
Accuracy	$(TP+TN) / (TP+TN+FP+FN)$	96.025%
Detection Rate	$(TP) / (TP+FP)$	96.265%
False Alarm Rate	$(FP) / (FP+TN)$	3.715%

The second step in the evaluation process is making a comparison with existing intelligent approaches for network intrusion detection to prove that, the proposed hybrid intelligent approach for network intrusion detection has improved the detection rate and decrease the false alarm rate. For this purpose, the proposed hybrid intelligent approach has compared to four of latest hybrid intelligent approaches for network intrusion detection. Table 5.5 presents the comparison and figure 5.10 shows the difference between these approaches in detection rate.

Table 5.5:

Comparison of Existing Approaches With The Proposed Hybrid Intelligent Approach

ATHUR / YEAR	TECHNIQUES	DATASET	DETECTION RATE
Kshirsagar and Patil (2010)	AdaBoost Algorithm	NSL-KDD	90.31%
Panda, Abraham, and Patra (2010)	NB + PCA	NSL-KDD	94.84%
Govindarajan and Chandrasekaran (2012)	RBF + SVM	NSL-KDD	85.19%
Patra and Panigrahi (2013)	SVM + PCA	NSL-KDD	94.06%
The proposed hybrid intelligent approach (2014)	K-Means + SVM	NSL-KDD	96.26%

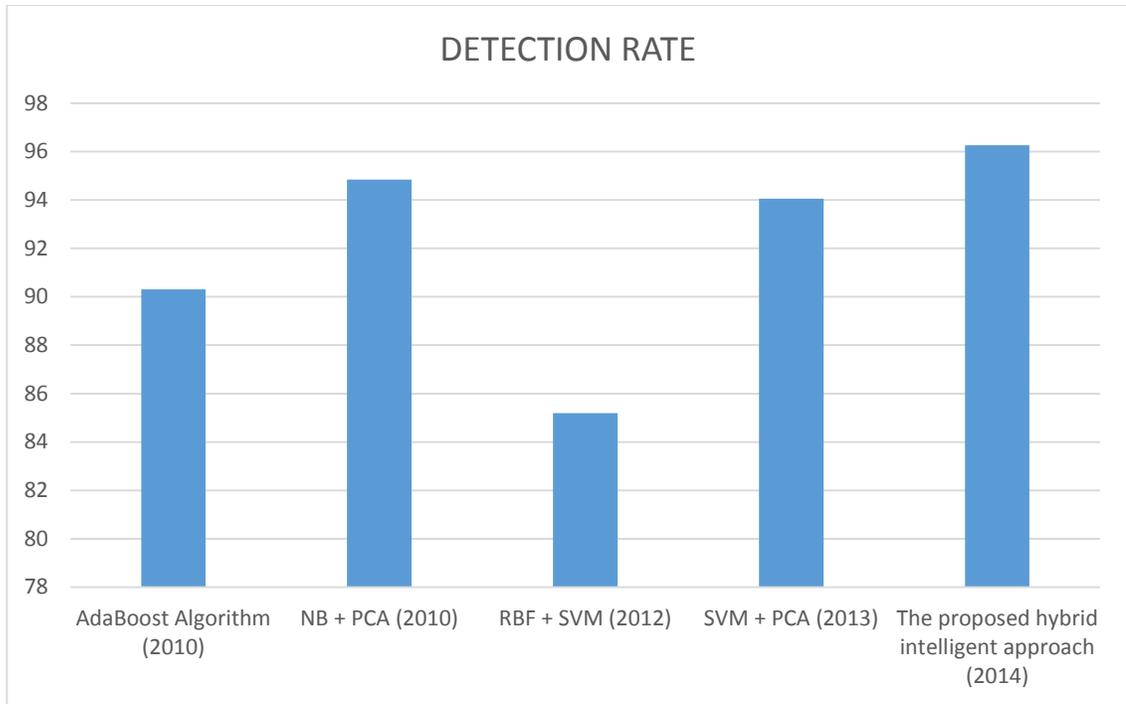


Figure 5.5: Comparison of Proposed Approach's Detection Rate With Others

5.5 Summary

In this chapter, the results of training and testing of the proposed hybrid intelligent approach for network intrusion detection have been presented. Confusion matrix has been used to present the result. Confusion matrix used to present the percentages of the true positive rate (TP), false positive rate (FP), false negative rate (FN), and true negative rate (TN). In addition, the process of performance evaluation has been done by using mathematical equations. Finally, a comparison the results of the proposed approach have made with the existing network intrusion detection approaches.

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6.1 Conclusion

Recently, the impact of a successful attack on an institution can have disastrous consequences such as a privacy breach, data loss, or service interruption. Therefore, many network intrusion detection approaches have developed and improved that are meant to combat such threats. Due to limitation of finding novel attacks in previous intrusion detection approaches and weakness in detection rate, this thesis has proposed a hybrid intelligent approach for network intrusion detection to overcome the main shortcomings of the existing intrusion detection approaches such as accuracy and false detection rate. For this purpose, the proposed approach integrated two machine learning techniques by merging K-Means clustering and SVM classifier for network intrusion detection.

Hence, a three objectives have been achieved by this thesis, which are design, develop, and evaluate the proposed approach. The methodology which uses to achieve the research objectives consist of four phases. Firstly, selection of experimental data which chooses an intrusion dataset depend on previous studies for testing the proposed approach. Secondly, data pre-processing which prepares the input data for classification. The main objective of preprocessing phase is to reduce ambiguity and provide accurate information to detection engine. Thirdly, classification phase has been classified all the network traffic into natural or intrusion behavior and assign each attack to its specific

category. Eventually, performance evaluation, the evaluation of the proposed hybrid intelligent approach has two steps, the first step has been applied by using mathematic equations to calculate accuracy, detection rate, and false alarm rate and other step have been done by making a comparison the result of the proposed approach with existing hybrid intelligent approaches for network intrusion detection.

According to previous studies, this research has concluded that NSL-KDD dataset is very ideal for training and testing different intrusion detection approaches. Hence, NSL-KDD intrusion dataset has been used for testing the proposed approach. The second phase is data preprocessing which is important steps to make it as an appropriate input to classification phase and clean network data, grouping, labeling, and the handles missing or incomplete dataset. It has been done by applying the four steps which are, data transformation, feature selection, data normalization, and data clustering.

In general, there are some symbolic attributes in NSL-KDD like flag, service types, and protocol types. Therefore, in transformation step, the symbolic values have been converted into numeric values beforehand to make it suitable input for classification phase. On the other hand, the usage of all the features in the network data to detect the intrusion patterns leads to time consuming detection and also the performance degradation of the system.

On the other hand, some of the features in this are redundant, make the noisy and irrelevant for the intrusion detection process. In view of the mentioned reasons, ClassifierSubsetEval and BestFirst algorithms have been applied for reducing the dimensionality of the data. They have chosen 21 features from all 41 features which are pertinent to intrusion detection purpose.

Later, the normalization stage has been performed Min-Max formula. Data normalization reduces the difference of the data by range it between (0, 1) values. Hence, it improves detection accuracy and speed up the execution time of classification. In the final step of data preprocessing, data clustering has been done. The essential aim of data clustering is that, it splits and labels the network's behaviors into two clusters (i.e. normal and anomaly) and also identified the corresponding cluster percentage. The labeled clustered data are later re-classified into specific classes which are attack classes and normal class. The clustering step has done by using k-means algorithm. After data preprocessing phase, the first contribution of this thesis has been achieved, which is that, reducing the manual work through categorizing and labeling the network data by performing the clustering before classification.

In the classification phase, the processed data from previous phase have been classified into two types of behavior namely, attack behavior and normal behavior. As well as, assigning the attack behavior to specific attack category. At the end of classification phase, the first couple objectives of this thesis have been accomplished, which are; design the hybrid intelligent approach for network intrusion detection and developing

the proposed hybrid intelligent approach using WEKA machine learning environment. The final results of hybrid intelligent approach detection have been presented using confusion matrix method depend on false positive rate, true positive rate, false negative rate and true negative rate terms.

At the end of this study, the final detection results of the proposed hybrid intelligent approach have been evaluated in terms of accuracy, detection rate, and false alarm rate. The evaluation phase has been done by using mathematical equations and held a comparison in term of detection rate among the proposed approach and the result of latest hybrid intelligent approaches for network intrusion detection. Lastly, after evaluation phase, the third objective of this thesis has been achieved, which is, evaluating the performance of the proposed hybrid intelligent approach for network intrusion detection.

The final results of the proposed hybrid intelligent approach have shown that, it has achieved highly detection rate of classification network behaviors between normal and attacks reached to 96.26 percentages and low false alarm rate reached to 3.71 percentages. Also, the results have accomplished good percentages of alarms in terms of: false positive rate, true positive rate, false negative rate and true negative rate. In general, the proposed hybrid intelligent approach has reached a high detection rate for DoS, L2R, and Probe attacks categories and normal behavior. While, the results have shown the some lack in detection rate of U2R attack category.

U2R attacks are embedded in the data portions of the packet and hence it is difficult to achieve detection accuracy for this type of attacks, as well, it is an internal attack. The intruders have legal access to the victim's system. Finally, the second contribution of this thesis has been achieved, which is that, the best combination for two machine learning techniques with higher detection rate and lower false alarm rate comparison with existing hybrid intelligent approaches for network intrusion detection.

6.2 Recommendation and Future work

As a final point, the work that was done in this thesis provides a basis for future research of intrusion detection systems in several areas. Firstly, implementing the proposed hybrid intelligent approach to intrusion detection system with one of programming languages such as Java, C#, VB.net, or others into real environment for both types of intrusion detection system which are (NIDS) and (HIDS). Secondly, testing the proposed hybrid intelligent approach by using another intrusion dataset with other types of attacks. Thirdly, as a mentioned in the conclusion, the proposed hybrid intelligent approach has weakness in the detection of U2R attack category. So, one of the future direction is that studying and focusing on the features of U2R attack to improve the accuracy's detection for this attack. Finally, the objectives of this thesis focus on improving the accuracy and false detection rate network intrusion. With this in mind, it can work with other impacts of intrusion detection systems such resource and time consuming.

References:

- Ahmad, A., Bharanidharan Shanmugam, Norbik Bashah Idris, Ganthan Nayarana Samy, & AlBakri, S. H. (2013). Danger Theory Based Hybrid Intrusion Detection Systems for Cloud Computing. *International Journal of Computer and Communication Engineering*, Vol. 2(No. 6), (pp. 650-654).
- Akbar, S., Rao, K. N., & Chandulal, J. (2011). Implementing rule based genetic algorithm as a solution for intrusion detection system. *Int. J. Comput. Sci. Netw. Secur*, 11(8), 138.
- Al-Jarrah, O., & Arafat, A. (2014). *Network Intrusion Detection System using attack behavior classification*. Paper presented at the Information and Communication Systems (ICICS), 2014 5th International Conference on (pp. 1-6). IEEE.
- Babatunde, R., Adewole, K., Abdulsalam, S., & Isiaka, R. (2014). Development of an intrusion detection system in a computer network. *International Journal of Computers & Technology*, 12(5), 3479-3485.
- Bahrololum, M., & Khaleghi, M. (2008). Anomaly Intrusion Detection System Using Hierarchical Gaussian Mixture Model. *International journal of computer science and network security*, 8(8), 264-271.
- Baili, N. (2013). *Unsupervised and semi-supervised fuzzy clustering with multiple kernels*. (Doctor of Philosophy), University of Louisville.
- Bansal, D. R., Gupta, V., & Malhotra, R. (2010). Performance analysis of wired and wireless LAN using soft computing techniques-A review. *Global Journal of Computer Science and Technology*, 10(8), 67-71.

- Bhavsar, Y. B., & Waghmare, K. C. (2013). Intrusion Detection System Using Data Mining Technique: Support Vector Machine. *International Journal of Emerging Technology and Advanced Engineering*, 3(3), 581-586.
- Bhuyan, M., Bhattacharyya, D., & Kalita, J. (2013). Network Anomaly Detection: Methods, Systems and Tools. *Communications Surveys & Tutorials, IEEE*, 16(1), 303 - 336.
- Chae, H.-s., Jo, B.-o., Choi, S.-H., & Park, T.-k. (2013). Feature Selection for Intrusion Detection using NSL-KDD. *Recent Advances in Computer Science*, 184-187.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.
- Chapke Prajkta, P., & Raut, A. (2012). Hybrid Model For Intrusion Detection System. *International Journal Of Engineering And Computer Science*, 1(3), 151-155.
- Chen, J., Wang, X., & He, L. (2008). *An architecture for differentiated security service*. Paper presented at the Electronic Commerce and Security, 2008 International Symposium on (pp. 301-304). IEEE.
- Chimphlee, W. H., Abdul; Sap, Mohd Noor Md; Chimphlee, Siriporn; Srinoy, Surat. (2007). A Rough-Fuzzy Hybrid Algorithm for computer intrusion detection. *The International Arab Journal of Information technology*, 4(3), 247-254.
- Chitrakar, R., & Chuanhe, H. (2012). *Anomaly detection using Support Vector Machine classification with k-Medoids clustering*. Paper presented at the Internet (AH-ICI), 2012 Third Asian Himalayas International Conference on (pp. 1-5). IEEE.

- Chitrakar, R., & Huang, C. (2012). *Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification*. Paper presented at the Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on (pp. 1-5). IEEE.
- Chowdhary, M., Suri, S., & Bhutani, M. (2014). Comparative Study of Intrusion Detection System. *International Journal of Computer Sciences and Engineering*, 2(4), 197-200.
- Daniel, J. V., Joshna, S., & Manjula, P. (2013). A Survey of Various Intrusion Detection Techniques in Wireless Sensor Networks.
- Danziger, M., & de Lima Neto, F. B. (2010). *A hybrid approach for IEEE 802.11 intrusion detection based on AIS, MAS and naïve Bayes*. Paper presented at the Hybrid Intelligent Systems (HIS), 2010 10th International Conference on (pp. 201-204). IEEE.
- Davis, J. J., & Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*, 30(6), 353-375.
- Dhawan, A. (2013). Data mining with Improved and efficient mechanism to detect the Vulnerabilities using intrusion detection system. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2(2), pp: 787-791.
- Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., & Fahmy, M. M. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Engineering Journal*, 4(4), 753-762.

- Engen, V., Vincent, J., & Phalp, K. (2011). Exploring discrepancies in findings obtained with the KDD Cup'99 data set. *Intelligent Data Analysis*, 15(2), 251-276.
- Eric, K. (2009). *Simulation on Network Security Design Architecture for Server Room in Rwanda Information Technology Agency*. Universiti Utara Malaysia.
- Esh Narayan, P. S. a. G. K. T. (2012). Intrusion Detection System Using Fuzzy C_ Means Clustering with Unsupervised Learning via EM Algorithms. *VSRD-IJCSIT*, Vol. 2(6), 502-510.
- Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20): Siam.
- Gao, M., & Wang, N. (2014). A Network Intrusion Detection Method Based on Improved K-means Algorithm. *Advanced Science and Technology Letters*, 53, 429-433.
- Garzia, F., Tirocchi, N., Scarpiniti, M., & Cusani, R. (2012). Optimization of Security Communication Wired Network by Means of Genetic Algorithms. *Communications & Network*, 4(3), 196-204.
- Ghadiri, A., & Ghadiri, N. (2011). *An adaptive hybrid architecture for intrusion detection based on fuzzy clustering and RBF neural networks*. Paper presented at the Communication Networks and Services Research Conference (CNSR), 2011 Ninth Annual. (pp. 123-129). IEEE.
- Govindarajan, M. (2014). A Hybrid RBF-SVM Ensemble Approach for Data Mining Applications. *International Journal of Intelligent Systems and Applications (IJISA)*, 6(3), 84 - 95.

- Govindarajan, M., & Chandrasekaran, R. (2012). *Intrusion Detection using an Ensemble of Classification Methods*. Paper presented at the Proceedings of the world congress on engineering and computer science (Vol. 1).
- Hameed, S. M., Saad, S., & AlAni, M. F. (2013). An Extended Modified Fuzzy Possibilistic C-Means Clustering Algorithm for Intrusion Detection. *Lecture Notes on Software Engineering*, 1(3), 273-278.
- Hameed, S. M., & Sulaiman, S. S. (2012). Intrusion Detection Using a Mixed Features Fuzzy Clustering Algorithm. *Iraq Journal of Science (IJS)*, 53(2), 427-434.
- Husagic, A., Koker, R., & Selman, S. (2013). *Intrusion detection using neural network committee machine*. Paper presented at the Information, Communication and Automation Technologies (ICAT), 2013 XXIV International Symposium on (pp. 1-6). IEEE.
- Ibrahim. (2010). Artificial Neural Network for Misuse Detection. *Journal of Communication and Computer*, 7(6), 38-48.
- Ibrahim, Basheer, D. T., & Mahmud, M. S. (2013). A Comparison Study For Intrusion Database (Kdd99, Nsl-Kdd) Based On Self Organization Map (SOM) Artificial Neural Network. *Journal of Engineering Science and Technology*, 8(1), 107-119.
- Idika, N. C., Marshall, B. H., & Bhargava, B. K. (2009). *Maximizing network security given a limited budget*. Paper presented at the the Fifth Richard Tapia Celebration of Diversity in Computing Conference: intellect, initiatives, insight, and innovations (pp. 12-17). ACM.

- Ishida, M., Takakura, H., & Okabe, Y. (2011). *High-performance intrusion detection using optigrd clustering and grid-based labelling*. Paper presented at the Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on (pp. 11-19). IEEE.
- Jain, Sharma, S., & Sisodia, M. S. (2011). Network Intrusion Detection by using Supervised and Unsupervised Machine Learning Technique-A Survey. *International Journal of Computer Technology and Electronics Engineering*, 1(3), 14 - 20.
- Jain, Singh, T., & Sinhal, A. (2013). A Survey on Network Attacks, Classification and Models for Anomaly-based network intrusion detection systems. *International Journal of Engineering Research and Science & Technology*, 4(2), 64 - 73.
- Jaisankar, N., & Kannan, A. (2011). A Hybrid Intelligent Agent Based Intrusion Detection System. *Journal of Computational Information Systems*, 7(8), 2608-2615.
- Jawhar, M. M. T., & Mehrotra, M. (2010). Design network intrusion detection system using hybrid fuzzy-neural network. *International Journal of Computer Science and Security*, 4(3), 285.
- Jiang, S. (2012). Internet Development Versus Networking Modes *Future Wireless and Optical Networks* (pp. 17-35): Springer.
- Joshi, S., & Varsha, S. P. (2013). Network Intrusion Detection System (NIDS) based on Data Mining. *International Journal of Engineering Science and Innovative Technology (IJESIT)*, 2(1), 95 - 98.
- Jyothsna, V., & Prasad, K. M. (2011). A Review of Anomaly based Intrusion Detection Systems. *International Journal of Computer Applications*, 28(7), 26 - 35.

- Kong, Y.-H., & Xiao, H.-M. (2009). *A new approach for intrusion detection based on Local Linear Embedding algorithm*. Paper presented at the Wavelet Analysis and Pattern Recognition, 2009. ICWAPR 2009. International Conference on (pp. 107-111). IEEE.
- Kshirsagar, V., & Patil, D. R. (2010). Application of Variant of AdaBoost based Machine Learning Algorithm in Network Intrusion Detection. *International Journal of Computer Science and Security (IJCSS)*, 4(2), 1-6.
- Kulhare, R., & Singh, D. (2013). Survey paper on intrusion detection techniques. *International Journal of Computers & Technology*, 6(2), 329-335.
- Kumar, Gulshan, K., & Krishan. (2012). The use of artificial-intelligence-based ensembles for intrusion detection: a review. *Applied Computational Intelligence and Soft Computing*, 2012, 1 - 20.
- Li, L., & Yuan, Y. (2010). Data Preprocessing for Network Intrusion Detection. *Applied Mechanics and Materials*, 20, 867-871.
- Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1), 16-24.
- Liu, Wan, P., Wang, Y., & Liu, S. (2014). Clustering and Hybrid Genetic Algorithm based Intrusion Detection Strategy. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(1), 762-770.
- Liu Li , Wan Pengyuan Wang , Y., & Songtao, L. (2014). Clustering and Hybrid Genetic Algorithm based Intrusion Detection Strategy. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(1), 762-770.

- Liu, Z. (2011). A method of svm with normalization in intrusion detection. *Procedia Environmental Sciences*, 11, 256-262.
- Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1), 115-128.
- Mohammad, M. N., Sulaiman, N., & Khalaf, E. T. (2011). A Novel Local Network Intrusion Detection System Based on Support Vector Machine. *Journal of Computer Science*, 7(10), 1560-1564.
- Muniyandi, A. P., Rajeswari, R., & Rajaram, R. (2012). Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. *Procedia Engineering*, 30, 174-182.
- Neethu, B. (2013). Adaptive Intrusion Detection Using Machine Learning. *IJCSNS*, 13(3), 118.
- NSL-KDD intrusion dataset* , (2014, June 1). Retrieved from <http://www.nsl.cs.unb.ca/NSL-KDD/>.
- Omit, I.-V. (2008). *A fuzzy feature evaluation framework for network intrusion detection*. (Doctor of Philosophy), The university of New Brunswick.
- Panda, M., Abraham, A., & Patra, M. R. (2010). *Discriminative multinomial naive bayes for network intrusion detection*. Paper presented at the Information Assurance and Security (IAS), 2010 Sixth International Conference on (pp. 5-10). IEEE.
- Panda, M., Abraham, A., & Patra, M. R. (2012). A hybrid intelligent approach for network intrusion detection. *Procedia Engineering*, 30, 1-9.

- Panda, M., & Patra, M. R. (2008). Some clustering algorithms to enhance the performance of the network intrusion detection system. *Journal of Theoretical and Applied Information Technology*, 710-716.
- Panwar, S. S., Sharma, R., Kumar, V., & Maheshwari, V. (2014). A Comprehensive Study of Clustering Techniques to Analyze NSL-KDD Dataset and Research Challenges. *International Journal of Enhanced Research in Science Technology & Engineering*, 3(1), 557-564.
- Patra, M. R., & Panigrahi, A. (2013). *Enhancing Performance of Intrusion Detection through Soft Computing Techniques*. Paper presented at the Computational and Business Intelligence (ISCBI), 2013 International Symposium on (pp. 44-48).
- Pei, L., Li, C., Hou, R., Zhang, Y., & Ou, H. (2013). *Computer Simulation of Denial of Service attack in Military Information Network using OPNET*. Paper presented at the 3rd International Conference on Multimedia Technology (ICMT-13).
- Pillai, M. B., Singh, M. U. P., & Lncet, A. P. C. (2011). NIDS For Unsupervised Authentication Records of KDD Dataset in MATLAB. *IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Wireless & Mobile Networks*, 57-61.
- Powers, S. T., & He, J. (2012). A hybrid artificial immune system and Self Organising Map for network intrusion detection. *arXiv preprint arXiv:1208.0541*.
- Prabha, K., & Sukumaran, S. (2013). Single-Keyword Pattern Matching Algorithms for Network Intrusion Detection System. *International Journal of Computer and Internet Security*, 5(1), 11-18.

- Raghuveer, K. (2012). Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set. *International Journal of Information and Network Security (IJINS)*, 1(4), 294-305.
- Rangadurai, K., R, Hattiwale, V. P., & Ravindran, B. (2012). *Adaptive network intrusion detection system using a hybrid approach*. Paper presented at the Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on (pp. 1-7). IEEE.
- RavinderReddy, R., Kavya, B., & Ramadevi, Y. (2014). A Survey on SVM Classifiers for Intrusion Detection. *International Journal of Computer Applications*, 98(19), 34-44.
- Revathi, S., & Malathi, A. (2014). Network Intrusion Detection Using Hybrid Simplified Swarm Optimization and Random Forest Algorithm on Nsl-Kdd Dataset. *International Journal Of Engineering And Computer Science*, 3(2), 3873-3876.
- Sanyal, S., & Thakur, M. R. (2012). A Hybrid Approach towards Intrusion Detection Based on Artificial Immune System and Soft Computing. *arXiv preprint arXiv:1205.4457*.
- Satpute, K., Agrawal, S., Agrawal, J., & Sharma, S. (2013). *A survey on anomaly detection in network intrusion detection system using particle swarm optimization based machine learning techniques*. Paper presented at the Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA).
- Selman, A. H. (2013). Intrusion Detection System using Fuzzy Logic. *SouthEast Europe Journal of Soft Computing*, 2(1), 14 - 20.

- Shanmugam, B., & Idris, N. B. (2011). Hybrid Intrusion Detection Systems (HIDS) using Fuzzy Logic. *Intrusion Detection Systems, Dr. Pawel Skrobanek, Ed. Croatia, Europe: InTech*, 135-155.
- Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security, 31*(3), 357-374.
- Shivakumar, M., Subalakshmi, R., Shanthakumari, S., & Joseph, S. J. (2013). Architecture for Network-Intrusion Detection and Response in open Networks using Analyzer Mobile Agents. *International Journal of Scientific Research in Network Security and Communication, 1*, 1-7.
- Siddiqui, M. (2004). *High performance data mining techniques for intrusion detection*. University of Central Florida Orlando, Florida.
- Song, G., Guo, J., & Nie, Y. (2011). *An Intrusion Detection Method based on Multiple Kernel Support Vector Machine*. Paper presented at the Network Computing and Information Security (NCIS), 2011 International Conference on (Vol. 2, pp. 119-123). IEEE.
- Stein, G., Chen, B., Wu, A. S., & Hua, K. A. (2005). *Decision tree classifier for network intrusion detection with GA-based feature selection*. Paper presented at the Proceedings of the 43rd annual Southeast regional conference-Volume 2.
- Sung, A. H., & Mukkamala, S. (2003). *Identifying important features for intrusion detection using support vector machines and neural networks*. Paper presented at the Applications and the Internet, 2003. Proceedings. 2003 Symposium on on (pp. 209-216). IEEE.

- Tavallae, M. (2011). *An Adaptive Hybrid Intrusion Detection System*. University of New Brunswick.
- Tavallae, M., Bagheri, E., Lu, W., & Ghorbani, A.-A. (2009). *A detailed analysis of the KDD CUP 99 data set*. Paper presented at the Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.
- Teng, S., Du, H., Wu, N., Zhang, W., & Su, J. (2010). A cooperative network intrusion detection based on fuzzy SVMs. *Journal of Networks*, 5(4), 475-483.
- Upadhyaya, D., & Jain, S. (2013). Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification. *International Journal of Computer Science Issues (IJCSI)*, 10(3), 231 - 236.
- Wang, Hao, J., Ma, J., & Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. *Expert Systems with Applications*, 37(9), 6225-6232.
- Wang, Zhang, X., Gombault, S., & Knapskog, S. J. (2009). *Attribute normalization in network intrusion detection*. Paper presented at the Pervasive Systems, Algorithms, and Networks (ISPAN), 2009 10th International Symposium on (pp. 448-453). IEEE.
- Wang, Y. (2004). *A comparative study of classification algorithms for network intrusion detection*. (Degree of Master), Florida Atlantic University.
- Wankhade, K., Patka, S., & Thool, R. (2013). *An Overview of Intrusion Detection Based on Data Mining Techniques*. Paper presented at the Communication Systems and Network Technologies (CSNT), 2013 International Conference on (pp. 626-629). IEEE.

- Xiang, C., Xiao, Y., Qu, P., & Qu, X. (2014). Network Intrusion Detection Based on PSO-SVM. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(2), 1502-1508.
- Yang, J., & Ning, Y. (2010). *Research on feature weights of fuzzy c-means algorithm and its application to intrusion detection*. Paper presented at the Environmental Science and Information Application Technology (ESIAT), 2010 International Conference on (Vol. 3, pp. 164-166). IEEE.
- Yassin, W., Udzir, N. I., Muda, Z., & Sulaiman, M. N. (2013). *Anomaly-based intrusion detection through K-Mean clustering and Naives bayes classification*. Paper presented at the 4th International Conference on Computing and Informatics, ICOCI, Sarawak, Malaysia.
- Zhou, M. (2005). *Network Intrusion Detection: Monitoring, Simulation and Visualization*. University of Central Florida Orlando, Florida.