Identification of compromise – prone nodes in mobile ad hoc networks using machine learning techniques

RB Sebopelo

Orcid.org/0000-0002-1291-1173

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Computer Science* at the

North West University

Supervisor: Prof N Gasela

Co-supervisor: Dr B Isong

Graduation ceremony: October 2018

Student number: 22438939

DECLARATION

I, RODNEY BUANG SEBOPELO, hereby declare that this project report titled *IDENTIFICATION OF COMPROMISE-PRONE NODES IN MOBILE AD HOC NETWORKS USING MACHINE LEARNING TECHNIQUES* is my own work carried out at North West University, Mafikeng Campus and has not been submitted in any form for the award of a degree to any other university or institution of tertiary education or published earlier. All the material used as source of information has been duly acknowledged in the text.

Signature: _____ Date: _____

Rodney Buang Sebopelo

APPROVAL

Signature: _____ Date: _____

Supervisor: **Prof. N. Gasela**

Department of Computer Science
North West University
Mafikeng Campus
South Africa

Signature: _____ Date: _____

Co-supervisor: **Dr. B. Isong**

Department of Computer Science
North West University
Mafikeng Campus
South Africa.

ACKNOWLEDGEMENTS

I would like to give my deepest expression of gratitude to the king of eternity, the Almighty God, Jehovah, for allowing me to go through this research process fully resourced and also for the people he placed around me. Jehovah, I thank you for the successful completion of this research work, and for all the wisdom and knowledge you gave me. Without your help none of this would have been possible.

I am also grateful and very thankful to my supervisor Prof Naison Gasela, co-supervisor Dr Bassey Isong for their constructive criticism, helpful support, and advice that edifies and all the new things he taught me about research. I am also grateful for the patience he exercised while supervising me and the belief he had in me to finish this degree.

I wish to express my sincere thanks to the lecturers and staff of the Department of Computer Science, North West University, Mafikeng Campus, for their help and support.

Finally, I would also like to express my gratitude to all my family and friends for their encouragement and support. Thank you all.

Abstract

Security of mobile ad hoc networks (MANETs) has become a more convoluted problem than security in other networks – due to the open nature and the lack of infrastructure of such a network. Due to the dynamic nature of the network, a device can configure and differentiate locations by itself on the move. Since the devices in MANETs are autonomous and can move in any direction, they change links with other devices very often. The networks are supposed to perform well when the nodes are trusty and act cooperatively. The characteristic of this network is that routing mechanisms are more convoluted and nodes are more vulnerable to compromise and predominantly susceptible to denial of service attacks (DoS) launched by malicious nodes or intruders. To mitigate against compromise behaviours we have developed a novel automated security mechanism using support vector machine and logistic regression algorithms to identify and defend against malicious attacks. The mechanism classified data packets as normal or abnormal. Our results indicated that the LR algorithm has the capability of detecting normal and abnormal data packets in MANETs and shows a good detection result of 100%. Thus, the LR model was used in the design and development of MANET intrusion detection framework in this research. An experimental result was evaluated using PDER, PMMR, and confusion metrics as the performance techniques to evaluate the QoS of links for better identification of attacks. We compared our two schemes and present evaluation results obtained from simulation studies. We then propose a framework for intrusion detection.

Table of Contents

List of Figures

List of Tables

List of Acronyms

MANET	mobile ad hoc network
DoS	denial of service
SVM	support vector machine
LR	logistic regression
PDER	packet delivery ratio
PMMR	packet modification and misroute rate
QoS	quality of service
ML	machine learning
BAN	body area network
PAN	personal area network,
LAN	local area network
MAN	metropolitan area net

Chapter 1

1.1 Overview

In this research, we investigate compromised nodes in a MANET network. We aim to use machine learning techniques to identify compromised nodes in MANETs in order to ensure secured packets delivery over the multi-hop wireless channel. The research also proposes the idea of identification and alleviation of compromised nodes in MANETs using SVM and LR.

1.2 Introduction

Wireless communication technology allows the user to communicate without any physical infrastructure irrespective of the geographical location. Consequently, this is why sometimes it is referred to as an infrastructure-less network [1]. A mobile ad hoc network (MANET) [2] is an independent collection of mobile nodes / devices that can communicate with each other without any central supervision as shown in Figure1.1. The mobile nodes within a radio transmission range can directly communicate, whereas others need the help of intermediate nodes to route their packets. MANETs have become one of the most researched areas in recent years due to the challenges they provide to the related protocols. With the advent of a wide range of wireless and mobile devices, provision of secure communication between different nodes in MANET's environment has become paramount. MANETs are more vulnerable to security attacks than other networks due to their unique features one of which is open network architecture, – and shared wireless medium [3]. It is also true that the requirements for secured networking require secured protocols which ensure confidentiality, availability, authenticity and integrity of network [4].

Due to the mobility of nodes in MANETs that cause frequent changes of the network topology, decentralized networking, and message delivery must be executed by the device themselves [5]. MANETs should be able be detect the presence of other devices in order to facilitate communication, – such as sharing of data and other services. The self-configuring nature of MANETs allows the devices to maintain the connections to the networks as well as to remove or add devices to the network. Due to the dynamic nature of the network, the topology may change rapidly and unpredictably overtime [6]. The dynamic nature and

mobility of mobile devices make MANET networks vulnerable to malicious attacks. Similar to other wireless networks, MANETs are susceptible to passive and active attacks [7]. Passive attacks involve only the eavesdropping of data. Active attacks involve actions such as replication, modification, and deletion of data that may be performed by an adversary. According to Karpijoki et al. [8-10] the attacks on MANETs may include actions such as dropping packets, or gaining substantiation or have access authorization by inserting forged packets into a data stream. The results – of such security attacks on MANETs may include having their effective output compromised for features like changed topology, restricted battery power and others. In contrast, in the wired network the attacker must have physical access to the wires of the network or pass through the defence lines of the firewalls and gateways. MANETs do not have a definite path; each node must be prepared to be confronted by malicious attackers directly or indirectly. MANETs are receptive to being captured, compromised, and hijacked since they can roam independently. The challenge encountered by self-configuring infrastructure nodes is that they must be willing to operate in a non-trusting mode. In general, attacks in MANETs can cause congestion and other disadvantages [11]. Since tracking down mobile nodes is hard to achieve, attacks done on them are more damaging and more difficult to detect [12].

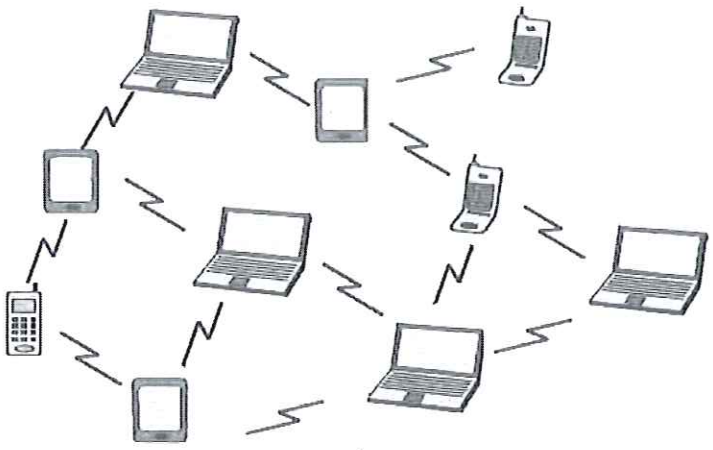

Figure 1.1: Mobile ad hoc network [13]

It is true that the dynamic nature of the network topology has been attractive to many different application areas such as military tactical networks, wireless sensor networks, and many others. These applications have in turn introduced some design issues and challenges that need to be overcome [14, 15]. Included in these issues and challenges are a limit in

bandwidth, battery power, computational power, and security [16, 17]. However, chief among these challenges is network security. Compromised nodes can disturb the correct functioning of a routing protocol by modifying or fabricating routing information. The solutions to the wired networks do not necessarily work on MANETs, which makes MANETs more vulnerable to attacks among them black hole and grayhole [18, 19].

In a blackhole attack, a malicious node claims to be having the shortest route to the node whose packets it wants to intercept, then on receiving the request the misbehaviour triggered a fake reply with the shortest route. Once a malicious node is placed between the participating nodes, it is able to control cooperative nodes in the network. In contrast, a grayhole attack which is also known as a misbehaviours attack drops the message in the active network. The attack is made up of phases where a node advertises itself as having the valid route to the destination, and the node drops the packets with a certain probability [20].

A modern security approach uses multiple layers of defence to shield a network from malicious nodes. MANETs are highly vulnerable and hence require secure communication. The main goal of MANET routing algorithms is to establish a route between a pair of nodes in the network so that a message can be delivered according to the expected quality of service parameters [21]. Route establishment should be done with a minimum overhead and bandwidth consumption. On-demand (reactive) routing is a type of protocol that establishes a route to a given destination only when a node requests it by initiating a route discovery process. Once a route has been created, the node keeps it until the final destination is no longer accessible. An example of a reactive protocol is the dynamic source routing (DSR) protocol [22] that determines the complete route to the destination node as a list of nodes of the routing path, and embeds it in the data packet. The protocol helps nodes to maintain multiple routes to a destination in a cache, which is helpful in case there is a link failure.

Machine learning is a field of artificial intelligence that gives the computer the ability to learn without explicitly being programmed. Machine learning is basically divided into three phases; data gathering, learning and classification. It is also divided into three different categories namely supervised learning, unsupervised learning and reinforcement learning [23]. A support vector machine (SVM) also called support vector network is a supervised machine learning algorithm which can be used for both classification and regression challenges [24]. It belongs to statistical learning theory and structural risk minimization: a

learning paradigm that is based on the theoretical and practical approach. It relies on the support vectors [25] to identify the decision boundaries between various classes which are located near the separation surfaces to define the largest possible margin for achievement of correct classification. The goal of SVM is to find the best separating hyper plane that separates clusters of vectors into two distinct classes. It takes the set of input training data and predicts for each input, any two feasible classes that form the input. It is a linear classifier used to detect misbehaviours [23].

A Logistic Regression (LR) model has become one of the most widely used and accepted methods for the analysis of binary output variables. The mathematical concept that underlies the LR is called the logit or the natural logarithm that is based on the odds ratio [26]. Generally, an LR describes and tests hypotheses based on the relationships between some categorical dependent variables and one or several independent variables [27]. LR, as one of the multivariate analysis models, can be used to predict the patient conditions such as heart disease and others. For this reason, these algorithms act as the best approach of data modelling and classification while providing good generalization performance balance between the complexity and the learning ability of the model. They also play an important role in nonlinear classification. The SVM, provides a quick response to compromised and malicious attacks [28]. The method provides a solution to MANET attacks and offers countermeasures using properties of metrics that classify packets as normal and abnormal. The metrics together with SVM are used to design a security system for MANETS and are easily computed [23]. The SVM checks the behaviour of each node in the behaviour of each node in the network and classifies those nodes according to their characteristics. These nodes are either classified as trusted or untrusted with the assistance of the SVM classifier.

In this study, we investigated and propose the method of identification and alleviation of compromised nodes in MANETs in order to thwart the malicious adversary attacks using SVM and LR. The essence was to identify which of the models is more efficient in the detection of compromised nodes. Although several machine learning algorithms have been applied to automate the detection of meaningful pattern for intrusion detection in MANETs, this research is geared towards applying SVM and LR for detection of malicious attacks in the network of MANETs. Thus, in this research, machine learning algorithms of SVM and LR will be utilized to categorize MANET's generated data as either normal or abnormal in

the network. It is far more resilient to the general changes in MANETs such as those due to malicious nodes changing their patterns over time or rapid changes in environmental factors.

1.3 Problem statement

Compromise-prone nodes are nodes on which an attacker has gained control after network deployment -. Generally, a compromised node exists when a malicious attacker has gained access and then directly connects devices to the network, and once attached, the attacker can take control of the nodes by extracting data, putting new data or controlling access to the data [11].

For this reason, compromised nodes can prevent correct functioning of routing protocols, this is possible by modification of routing information, fabricating of bogus routing information and prevent services from working properly [10]. These types of attacks make use of private connection to create a tunnel and bypass the actual network. Consequently, short circuit of normal flow routing messages that create a disruption in the network propagate incorrect routing information and prevent services from working properly [11].With the operation performed by compromised nodes within the network, it becomes crucial to investigate the weaknesses of such nodes and provide enhancements to mitigate against current security loopholes. It is also necessary to be able to identify whether a node has been compromised or not during the operation of a MANET. Due to their characteristics MANETS are more convoluted, vulnerable to compromise and predominantly susceptible to malicious activities launched to those compromised nodes from a variety of perspectives.

Therefore, in this study, two methods are proposed, namely Support Vector Machine (SVM) and Logistic Regression (LR) of identification and alleviation of compromised nodes in MANETs in order to thwart these malicious attacks and compare the two to determine the more efficient method.

1.4 Research Questions

Accordingly, in order to meet the aim of this study, the following research questions will be answered:

a) How can we identify if a node in a MANETs is compromised or not?

b) How can we design a predictive model to identify such nodes in the network?

c) How can we implement and evaluate the designed model?

1.5 Research Aim

The aim is to use machine learning techniques to identify compromised nodes in MANETs in order to ensure secure packets delivery over the multi-hop wireless channel.

1.6 Research Objectives

To achieve the aim of this research, the following objectives will be carried out:

a) Survey the literature on security issues related to security on MANET link when delivering packets and appropriate machine learning algorithms for prediction.
b) Design an automatic security model using appropriate machine learning and an optimization algorithm.
c) Implement the automatic security model and optimization developed.
d) Compare the results of our schemes and present evaluations.

1.7 Research Contributions

The main contribution of this research is in assisting the nodes in a MANET to reduce malicious attacks, and to tighten security by identifying compromised nodes in order to ensure secured packets delivery over the multi-hop wireless channel. This is achieved by the optimization algorithms, usage of metrics and machine learning models developed in this research as a validation of the concept and approach.

1.8 Research Methodology

Based on the goal of this research, quantitative and qualitative research methods will be used. Literature review on previous work related to MANETs will be excavated and analysed from which will be derived the activities of the investigation.

a) Qualitative Research: The framework (software and hardware tools) selection to be used and the collection of data set.

b) Quantitative Research: Simulation of the SVM and LR algorithms using experiments. Comparative data analysis of the data outputs from the two algorithms by analyzing the data produced within a time frame.

1.8.1 Comprehensive literature review

Literature review on previous work related to machine learning algorithms, MANETs and the knowledge gained will be used to design the model.

1.8.2 Model design

Based on the literature review, SVM and LR algorithm models will be optimized and practically implemented.

1.8.3 Model implementation

Based on the designs, SVM and LR models will be implemented in Python framework and evaluate suitable benchmark performances of the designed algorithms.

1.8.4 Model analysis and evaluation

In this section, the designed algorithms from the implemented model are analysed and evaluated.

1.9 Chapter summary

This chapter gave a brief introductory insight of this study, the problem statement, and the aim as well as objectives of the research. It also outlined the research methodology employed in this study. In Chapter 2, we will discuss a review of the literature on related works that have been done on security of MANETs, and identification of compromised nodes. It also gives a review of the background information regarding security related issues on identifying compromised nodes in MANETs for ensuring secured packets delivery over the multi-hop wireless channel. In Chapter 3, we investigate and propose the method of identification and alleviation of compromised nodes in MANETs in order to thwart the malicious adversary attacks using SVM and LR. The essence is to identify which of the models is more efficient in the detection of compromised nodes. In Chapter 4, we discuss and analyse the results, method of identification and alleviation of compromised nodes in MANETs in order to thwart the malicious adversary attacks using SVM and LR algorithms. The essence is to identify which of the models is more efficient in the detection of compromised nodes and

then develop a framework for automatic detection of attacks on MANETs. In Chapter 5, we present a brief summary of the research that was conducted and show how the aim and objectives were achieved. It also discusses the conclusion, recommendations and proposed direction for future work.

Chapter 2

Literature Review

2.1 Chapter outline

This chapter presents a review of the literature on related works that have been done in security of MANETs, and identification of compromised nodes. It also gives a review of the background information regarding security related issues on identifying compromised nodes in MANETs for ensuring secured packets delivery over the multi-hop wireless channel.

2.2 Security overview

Security is one of the key components for basic network functioning such as packets and routing protocols. The performance of a network might be disrupted when crucial security measures are disobeyed when designing sensitive applications. It is crucial for security countermeasures to be adopted wirelessly due to the dynamic nature of the topology of MANETs. To provide secure network communication in the network layer where attacks take place, there are security measures that need to be addressed to ensure secure packet transmission, However, Kaur et al. [4] reported that these measures concern confidentiality, availability, authentication, access control and non-repudiation of the network.

2.3 Network security concepts

The study of security in computer networks is a growing area of interest in many applications. Since networks are prone to malicious attack which might disrupt the performance of a network and propagate incorrect routing information. This is exacerbated by a lack of security in MANETs due to features like dynamic nature of the network, open medium, cooperative algorithm, and lack of a clear line of defence. These security factors have challenged MANETs against different threats. However, the idea of providing and maintaining security measures in the network is quite challenging due to the fact that the network is infrastructure-less. Papadopoulos et al. [29] reported that network security is defined as a multiplicity of systems, procedures and other approaches to supply a certain level of security to the network. Attackers or intrusion that eavesdrops on information,

illegally accessing the resources remotely, invading computers remotely, fabricating routing information, inserting erroneous information into important files and flooding, consequently degrade the network's performance. We have defined the following terms as properties (characteristics) of a good network:

a) Confidentiality

In MANETs, Zhou et al. [14] reported that the protection of information of mobile nodes in the network from being exposed to malicious entities is called confidentiality. This process is more difficult in MANETs – due to the complicated structure where an intermediate node receives the packets for other recipients. Kumar et al. [18] also reported that the malicious attack can easily eavesdrop the information being routed along the network. This leakage on such information could lead to devastating consequences for the network.

b) Availability

MANETs should provide services to available mobile devices whenever needed. However, according to Boora et al. [30] the network should be able to provide assurance of survivability without consideration of misbehaviours that can take place in MANET's layer. Furthermore, in other layers of MANETs such as physical, misbehaviour can propagate by using the techniques of jamming in order to make the communication less effective. In contrast, in the network part, it can be propagated by incorrect functions on routing protocols which may lead to disconnection of the network.

c) Authenticity

Haas et al. [12] stated that MANET is concerned with assurance that the mobile devices will communicate with each other. Authenticity provides assurance or confirmation of the origin of a communication. The process prevents the attackers from compromising normal mobile nodes so that the resources and sensitive information in the network is not accessible to unauthorized members. Furthermore, this also restricts unwanted access from other services of the network.

d) Non–repudiation

Zhou et al. [12] reported that this ensures that the source and destination nodes in MANETs should not be denied access for sending and receiving messages by other mobile nodes. It is a very useful method for identifying compromised devices. For example, consider device G that has a corrupted message from device A, a security alarm is set off by G about node A of using the message report that G is compromised.

e) Access control

Kumar et al. [31], access control prevents unauthorised use of network services and the resources of the system.

2.4 MANET overview

According to Kaushik et al. [32], the architecture of MANETs is grouped into three categories namely: enabling technologies; networking, and middleware & applications. Enabling technology contains several classes: Body (BAN), Personal (PAN), Local (LAN), Metropolitan (MAN) and Wide (WAN) area networks. BAN can cooperate with wearable computers for strong connections. This connection is made of wearable computers to distribute its components such as microphones, earphones and others. Hence, it is a very helpful connectivity in various devices. PAN on the other hand can act as a link between the users, mobile devices and other stationary devices. MAN and WAN are multi hop networks that deal with unresolved issues to be taken into considerations such as routing, security and others. However, establishing the transmission that is end-to-end requires that the sender should know the whereabouts of the receiver for transmission to take place in the network.

Bang et al. [1, 2] reported that MANETs is an independent / autonomous collection of mobile nodes (laptop, smart phones, personal digital assistant) and others that communicate with each other without any central supervision / administration. MANETs is an infrastructure-less network which consists of mobile devices that are able to communicate wirelessly without the existence of access points. Devices in MANETs are able to move in any direction independently and structuring themselves and the topology of MANET changes erratically and dynamically. The density on mobile nodes and numbers depends on how the application users perform.

2.5 Applications of MANETs

Due to an increase in devices and rapid progression in wireless communication, ad hoc networks allow devices to keep connection to the network by addition and removal of nodes. According to Goyal et al. [33, 34], some typical applications of MANETs are as follows:

a) ***Military applications:*** Frodigh et al. [34] reported that military resources nowadays contain computer equipment. Ad hoc networking provides the advantage of common place networking technology to keep information between the soldier, vehicles and others. Initially, basic ad hoc networking came from this technique.

b) ***Personal Area Network (PAN):*** Loo et al. [35] reported that MANETs apply in a short range of inter-communication between different mobile devices such as personal digital assistant, laptop and cellular phone. Wired network cables have been replaced by wireless communication. For this reason, ad hoc networking can extend access to internet or other network by mechanisms such as Wireless Local Area Network (WLAN). PAN is one of the promising application fields of MANETs in the near future.

2.6 Vulnerabilities in MANETs

Vulnerabilities of MANETs come from the open nature of such networks. Unlike wired networks that have dedicated routers, mobile devices in this case act as routers and forward data to other mobile devices. According to Lalar et al. [36] the reported features of MANETs like randomly changing topology, multi-hop topologies consist of relatively constrained bandwidth wireless link that makes the network vulnerable to malicious activities. However, MANETs have many challenges that must be studied before a wide commercial deployment considered. The wireless network is accessible to both legitimate network users and malicious attackers; since it does not possess a definite or clear line of defence from security perspectives. Some of the typical challenges that come with MANETs are as follows:

a) ***Dynamic topologies:*** According to Goyal et al. [33], the changing infrastructure of the network allows mobile nodes to move arbitrarily and independently. The nodes in a MANET without correct physical protection can result in nodes misbehaving thereby degrading the network performance and propagating incorrect routing information. The changing topology of the network–which is typically multi hop, may cause changes to the network randomly and rapidly at unpredictable times.

b) ***Bandwidth-constraint:*** Lalar et al. [36] reported that wireless links will always have lower capacity than hardwired counterparts.

c) ***Routing:*** According to Lalar et al. [36], MANETs are constantly routing packets between nodes becomes a challenging task. Multi cast routing is another challenge due to the fact that it is no longer static because of the constant movement of nodes within the network. For these reasons, routes between mobile devices may have multiple hops which are more complicated than a single hop communication.

d) ***Quality of service (QoS):*** According to Xiao et al. [21], MANETs make it difficult to offer permanent or fixed guarantees on the services that are provided by the mobile nodes.

e) ***Wireless links:*** Pal et al. [37] reported that, as nodes join and leave the network through wireless interface, they become susceptible to malicious attacks. However, the bandwidth of a wireless dynamic network of MANETs is less compared to wired network, and the competition for bandwidth may lead nodes to be compromised as attackers may prevent the normal communication operation.

f) ***Cooperativeness:*** According to Lalar et al. [36], in MANETs topology, all the routing protocols assume that nodes in the network provide secure communication. However, the self–configuring nature of MANETs allows the devices to maintain the connections to the networks as well as to remove or add the device to the network. Due to this reason, some nodes in the network may become malicious and propagate incorrect routing information and degrade the network performance.

g) ***Lack of clear line of defence:*** Chandra et al. [38] reported that, MANETs do not have a definite path; each node must be prepared to be confronted by malicious attackers directly or indirectly. Nodes in MANETs are receptive to being captured, compromised, and hijacked since they can roam independently. Each node is willing to be attacked in the network either internally or externally.

h) ***Limited resources:*** Agalawe et al. [39] reported that, MANETs are made up of various sets of mobile devices that communicate with each other such as personal digital assistant, laptop, mobile phone and others without predefined infrastructure. However, these mobile devices have different storage capacities, processing speeds, computational powers and others. These may attract malicious nodes to launch new attacks.

2.7 Types of security attacks

According to Singh et al. [40], MANETs like other network paradigms are faced with several security attacks. Due to the factors like it being an infrastructure-less, dynamic topology and absence of trust from mobile devices, the routing protocols are vulnerable to malicious kinds of attacks. Jawandhiya et al. [41] reported that the attacks of the networks can be categorizes as internal or external. Also Alani et al. [42] reported that the attack can be classified on the behaviour such as passive and or active attacks. The attacker can propagate information that is incorrect and can prevent services from working accurately either internally or externally. There can be active and passive attack against the network as well. Nguyen et al. [43] discussed the definitions of internal, external, passive and active attacks as follows:

a) External attacks

Nodes outside of the network try and make their way into the network. When access is accepted they start distributing fake information to degrade performance of the whole network. These can be countered by proposing security measures such as firewalls and other blocking mechanisms where the access to the network is denied for unauthorized nodes.

b) Internal attacks

Internal attacks consist of the malicious attack that gains a privileged position inside the network activities. This can be done by impersonation or by a compromised node. Internal attacks have normal access to the resources of the network by participating in operations such as packet transmission. These types of attacks are more damaging than the external attacks.

c) Active attacks

Active attacks consider the actions such as the replication of data that may be performed by an adversary; thus disturbing the normal operation in the network. Furthermore, active attacks can be in the form of internal and external attacks. A mobile node that is not part of the network can be classified as an external attack. In contrast, internal attacks include those mobile nodes that have been taken over by an attacker in the network. Attacks from internal nodes of the network are the most damaging and they are harder to identify as compared to external attacks.

d) Passive attacks

The passive attacks will not modify the correct functioning of the network. They eavesdrop on the data without modifying it but violate the confidentiality and security requirements. There is no clear line of defence against passive attacks. However, the network operation is not affected by this. Otherwise, data can be encrypted during transfer to prevent the attacker from snooping.

2.8 Threats in MANETs

The network layer protocols provide functions that enable the MANET nodes to use multi-hop to communicate with each other in the network without a central supervision. The autonomous nodes in MANETs can take decisions to forward data packet. In such cases, the malicious attackers can initiate routing loops with the intention to cause the network to be dysfunctional and also disrupt other services.

a) Blackhole

Sharma et al. [21, 44] reported that this type of attack occurs when a malicious node receives a packet that is requested by a certain node, and initiates route claiming to have the definite path to the destination irrespective of whether the destination is known or not.

b) Grayhole

Vishnu et al. [19, 45] reported that this type of attack involves the activity that drops data packet without forwarding it. The grayhole attack acts as a disrupter of the network process. For this reason it is a node that behaves like a trusted node during the forwarding and receiving of packets and commences dropping packets that are sent silently when arriving.

c) Denial of service (DoS)

Karpijoki et al. [8] reported that in the DoS attack, the attacker prevents the real communication facilities by stopping services from working properly and disabling the operation of the entire communication network in order to make the network performance less effective. The attacker does not interfere with the data in the network, but disables the services and replace them with its own. An attacker may put packets into the network in order to disrupt valuable network resources resulting in unintentional failure or malicious action. The result of this attack depends on the area of application of the ad hoc network.

2.9 Routing protocols in MANETs

Karpijoki et al. [8] reported that routing is the most important component in MANETs and consists of different kinds of routing protocols that can be affected by malicious attacks. The aim of MANETs routing algorithms is to establish the correct functioning of nodes in the network so that a message can be delivered according to the expected quality of service parameters. A key issue in MANETs is that routing protocol should be aware of the changes within the network. MANETs routing protocols are able to establish paths between devices and to forward packets to the destination. However, MANETs do not have secure communication due to their high vulnerability. If routing of a message can be misdirected the entire operation of a network is also compromised. Jhaveri et al. [46] proposed protocols for MANETs and are classified into three categories, namely: reactive, proactive and hybrid.

2.9.1 Reactive protocol

Abolhasan et al. [47] reported that the protocol maintains information or other activity of the network without communication. In other words, if any node participating in the network wants to forward information to another mobile node the routing protocol searches for the routes and creates the path for sending and receiving information packets. Ali et al. [48] reported the example of reactive protocols.

2.9.2 Proactive protocol

Deng et al. [49] also referred to a proactive protocol as table driven. It contains protocols that allow one of the nodes to provide tables one or many for storage data packets. The protocol maintains a stable updated topology of the network. The mobile devices know in advance the information related to routing which is kept in different tables. As the topology of MANETs changes, these tables are also updated according to the changes in the network. Rahman et al. [50] reported an example of a proactive protocol as Destination Sequenced Distance Vector (DSDV) routing. Proactive protocols provide consistency and updated information for a complete network. Ali et al. [48] stated that they also reduce delay in communication and quickly provide access reachable nodes in the network.

2.9.3 Hybrid protocol

In contrast, a hybrid routing protocol combines good properties of both the reactive and proactive routing protocols and overcomes their weaknesses when sending information over the network. This type of protocol allocates nodes into network topological zones. In addition, Pandey et al. [51] reported that as the network separates into zones, routing information is maintained within each zone using the proactive protocol approach, while the reactive protocol is used for routing packets between different zones. Kaur et al. [52] discussed an example of hybrid protocol namely Zone Routing Protocol (ZRP).

2.10 Machine Learning (ML) overview

Ayodele et al. [53] reported that ML is a subfield of artificial intelligence that gives the computer the ability to learn and adapt on the logical, binary and other operations that gather information from a set of examples. Mannila et al. [54] stated that it is considered as a powerful collection of techniques used for data mining and knowledge discovery. Alsheikh et al. [55] stated that ML is concerned with the computational methods that give experience in order to bring and improve performance or to provide with good predictions, where the experience here is referred to as the past information available to the learner in order to collect and analyse electronic data such as digitized human labelled training sets or any other types of information interacting with the environment. It aims to classify expressions that are simple enough to be understood by humans. Kotsiantis et al. [56] also reported that, it requires the notion of sample complexity to measure and evaluate the sample size for the algorithm and learn different families of concepts. Taiwo et al. [53] also discussed that, learning includes the following:

a) To gain knowledge, comprehension or to acquire skills of training.
b) To provide the experience for training samples.
c) To learn something by experience using examples or practice.

ML can also be defined as building the systems for computers that simultaneously provide experience that makes use of learning as a process. According to Muller et al. [57] ML is also defined as automatic learning theory of data by model fitting or by training samples.

Michalski et al. [58] reported that the categories of ML can be prepared or organized according to the following areas:

a) Task oriented studies: analysis of learning and also development of the model for solving a defined task.
b) Cognitive simulation/cognitive modelling: learning which includes investigations and simulations of human learning processes.
c) Theoretical analysis: the possible learning methods and also algorithms which include theoretical exploration for independent application.

Since learning with machines depends on the data used, more generally learning techniques are data driven methods by combining fundamental concepts. Machine learning has been widely used for a number of tasks such as classification and regression, it is also used for a variety of applications which include bioinformatics, speech recognition, face detection, computer vision and others. The algorithms and techniques of machine learning come from different fields including statistics, mathematics, and computer science. Alsheikh et al. [55] proposed that the techniques of ML that can be used to solve problems are supervised and unsupervised learning as well as reinforcement learning.

2.10.1 Supervised learning technique

Patel et al. [59] reported that, supervised learning techniques involve the practice of training observations that consist of predefined input and known output generated by system generic parameters. The learning technique creates a knowledge structure in order to support the work of classifying new examples into predefined classes. Kotsiantis et al. [56] reported that the output classes are predefined by the training samples. The output of the learning technique is a classification model that is made up of or constructed from the instances provided. The learning is concerned with modelling the input/output relationships. It is also considered as the mapping from the input feature space to output class.

Nguyen et al. [60] reported that there also are various kinds of supervised learning technique algorithms, each differing in the form of classifying the model and how the optimization algorithm is constructed and used to search for the accurate model. Janakiram et al. [61] proposed that supervised learning algorithms be used in networking to solve a number of different challenges such as security and intrusion detection.

a) Linear Regression

Linear Regression is one of the basic and most common types of predictive analysis. |It measures the relationship between dependent variable and one or more independent variables.

However, the situation involving one independent variable is referred as simple linear regression and for more than one independent variables, the process is called multiple linear regression. To model the relationship, the technique called linear predictor functions is used [62]. However, a disadvantage of linear regression is that, a technique is limited to linear relationship between dependent and independent variables, sensitive to outliers, and the dataset must be independent [63].

b) Random forest

Also called random decision forests, is one of the learning methods for classification and regression analysis and other tasks. The technique operates by constructing a multitude of decision trees in the training time, and output the class classifications. The technique is a very popular method for different machine learning tasks [64]. Random forests have been observed to overfitting for most datasets with noisy classification or regression analysis, and algorithms that have large numbers of trees can make the algorithm slow for real-time prediction. The technique requires almost no input preparation, also it performs implicit feature selection, is very quick to train, versatile and simple with lots of excellent applications and implementations [65].

2.10.2 Unsupervised learning techniques

While supervised learning techniques use predefined training instances of the classes, unsupervised learning techniques are not provided with guidance or labels. This means there is no output vector, and for this reason they discover the patterns of clusters/groups in the data using internalised heuristics. The basic goal of unsupervised learning is to assign training observation sets into various groups by gathering the similarity between them. Alsheikh et al. [55] reported that the scheme is widely used in situations like node clustering and data aggregation scenarios. The method does this by finding patterns in the input training observation. It does this by clustering instances with common properties into groups. Huang et al. [66] also reported that the identified groups may be exclusive so that any example may belong to one of the following groups: overlapping, probabilistic or hierarchical.

a) K means clustering

The K-means clustering is a popular clustering method in data mining used for vector quantization. The aim of k-means is to partition the number of n observations into k clusters in which each independent variable belongs to the cluster. The results of k–means are to

partition data space into voronoi cell. The disadvantage of the k–means is computationally difficult; however there are other algorithms that are employed for the technique in order to make convergence very quickly to a local optimum. The algorithm uses an iterative refinement technique, and is easy to implement. With the large number of variables, the algorithm can be faster than hierarchical clustering; however it can also produce tighter clusters than hierarchical clustering. In addition, it is difficult to predict the number of clusters (k–value), and the initial value has a strong effect on the outcome. The order of the data also has an effect on the outcomes. The technique is also sensitive to scale, for example rescaling the datasets also affects the results completely. The technique is also sensitive to outliers, time complexity is not suitable for large data sets, and is not possible to undo the previous step [67].

b) Apriori

Apriori is an algorithm that has frequent item sets for mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Hence the association rules are determined by the frequent item sets together with an Apriori algorithm in the database. Apriori also has disadvantages such as assumption of the transaction database in memory resident; also the algorithm requires many database scans, can also be very slow and the bottleneck is candidate generation [68].

For this reason, several machine learning algorithms have been applied to automate the detection of meaningful pattern; this study is geared towards applying SVM and LR. Thus, our research machine learning algorithms of SVM and LR will be utilized to categorize MANET generated data as either normal or abnormal in the network. It is far more resilient to the general changes in MANETs such as those due to malicious nodes changing their patterns over time or rapid changes in environmental factors.

2.10.3 Support vector machines

Gunn et al. [24] reported that SVM also called support vector network is a supervised learning algorithm that can be used for both classification and regression challenges. Cortes et al. [25] also reported that since it follows the structural risk minimization principle, it reduces the risk of error occurrence during the training process and relies on the support

vectors to identify the decision boundary between various classes which are located near the separation surfaces to define the largest possible margin for achievement of correct classification. Amari et al. [69] proposed that the support vector machine technique is classified as one of the good linear classifier that improves the robustness of the classifier and is not sensitive in the sparsity and correlation of characteristics of data. Elish et al. [70] also reported that, it also remains one of the existing classification techniques of data mining application fields such as software defect prediction, face recognition, voice recognition and others.

In addition, Gunn et al. [71] reported that, a line that separates the training observations into two classes can be represented as equation 1 :

$$< w.x > + b = 0 \tag{1}$$

where w is an orthogonal dimensional vector or weight vector, x is the real valued input pattern of metrics, b is a bias term used to identify the perpendicular of a distance from the origin to the hyperplane, +1 represents the normal data packet, -1 represent the abnormal data packet. However, Smola et al. [72] also reported that, data points of the two parallel hyper planes can be represented by equations 2 and 3 as follows:

$$< w.x > + b \geq +1 \tag{2}$$

and

$$< w.x > + b \leq -1 \tag{3}$$

The process that involves support vector machine is represented in Figure 2.1

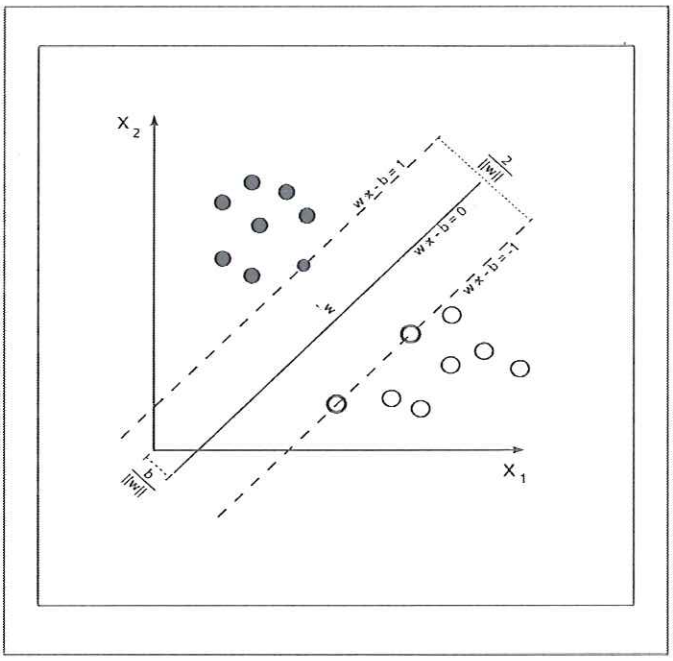

Figure 2.1 Support vector machine model [73]

2.10.4 Logistic Regression

Dreiseitl et al. [26] reported that LR is considered as a regression method that is used to make predictions based on a dependant variable. The LR is widely used in case of predictions such as the presence or absence of characteristic or the results that are based on the values of a set of predictor variables. However, Kurt et al. [27] also reported that this method is equivalent to a linear regression model but is very useful in cases where the dependent variable is dichotomous. Peng et al. [74] also reported that this method is used to provide with the estimation of probability that considers response of binary variables that consist of many features.

Furthermore, Harrell et al. [75] reported that LR measures the relationship between a dependent outcome and other more independent instances by making the probabilities which consist of logit function. It is also a special process of a generalised model that is linear, which is made up of different reasons based on the relationship of outcomes and independent instances. However, Mevlut et al. [27, 76] indicated that, the logistic function of the logistic regression f (x) is defined as the following:

$$f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \quad (4)$$

and Lida et al. [74] stated that the graph of the LR function is often referred to as the sigmoid or S-shaped curve on the interval.

The process that involves LR is represented in Figure 2.2

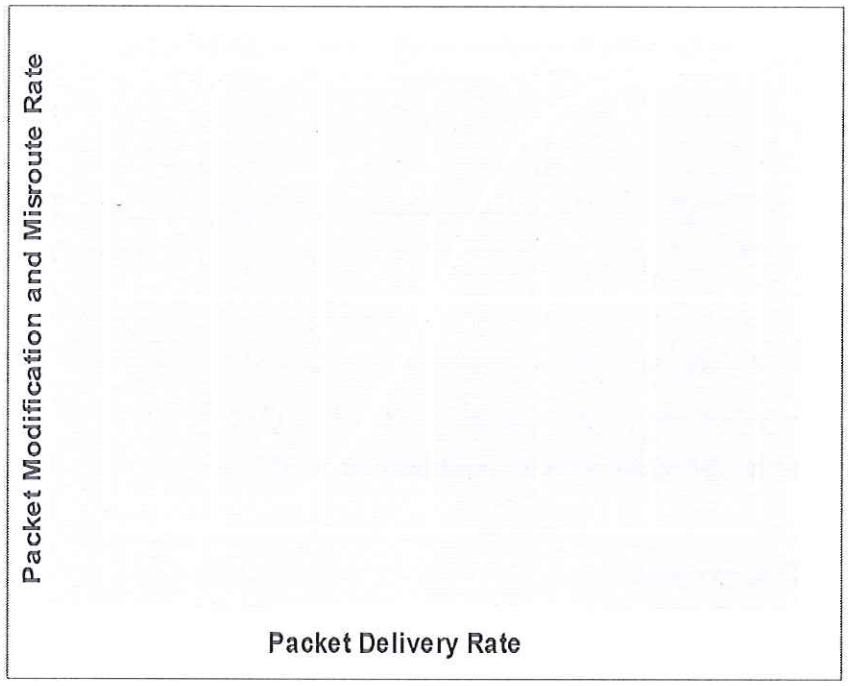

Figure 2.2 Logistic Regression model [77]

2.11 MANET metrics

Patel et al. [20] mentioned that to analyse the behaviour and performance of each node in the network, the following metrics packet delivery ratio, packet modification ratio, and packet misroute rate are used. But in this study, we focused on providing a more effective and efficient model to detect and enable mitigation plans to be provided in the event of attacks on nodes. In our case, we intended to utilize these metrics on both SVM and LR in order to determine the more effective model based on classification accuracies. Also Patel et al. [20] also proposed that packet modification ratio and packet misroute rate were considered as separate metrics. But this study combines them into a single metric that accounts for compromise in the packets delivered.

Patel et al. [20] discussed the metrics PDER, PMOR and PMISR as follows:

1. $PDER = \frac{\text{number of packets tranmsitted by node}}{\text{total number of incoming packets}}$

PDER is the delivered number of packet data with the total number of packets that are to be delivered by mobile node. Hence the larger number of PDER the better the performance of the mobile node.

2. $PMOR = (\frac{\text{number of packet's modified } by\ node}{\text{total number of incoming packets}})$

PMOR are the resultant changes of the packet content by the attackers during a transmission process. This is done accidentally or intentionally to some of the nodes present in the network. This also can be the process of communication when a node sends the packet to the wrong destination.

3. $\text{PMISR} = (\frac{\text{number of packets misroute by node}}{\text{total number of incoming packets}})$

PMISR is the process of communication when a node sends the packet to the wrong destination.

2.12 Related studies on MANET compromised node identification

Nikos et al. [12], projected two phase detection technique for identification of unauthorized nodes and compromised nodes for a specific service in MANETs. This approach is enabled with the operations of the network that can be found in the link and network layer.

Sukla et al. [78], proposed the technique of addressing the problem of packet forwarding misbehaviour based on the mechanism of detecting and removing the blackhole and grayhole attacks, exhibiting packet forwarding misbehaviour. The technique finds the chain of nodes with cooperating misbehaviours of dropping an important fraction of data packets. The approach involves the use of algorithms for detection of these attacks

Shabana et al. [79], proposed secure power–aware ant routing algorithm (SPA-ARA) for MANETS from ant colony optimization (ACO) algorithms that are from the swarm

intelligence technique. The introduced algorithms were used with a combination of two metrics: new metric and next-hop availability. It provides path availability and can minimize travel time of packets and this was very helpful in selecting accurate paths and improving the use of resources. The protocol used for the model helped to identify unauthorized and other nodes in MANETs.

Jhaveri et al. [80], projected the approach occurring on ad hoc on demand distance vector (AODV) protocol to detect nodes (blackhole and grayhole attacks), for this reason the sender nodes check through the routes available and check out whether the received destination for its message is not damaged. To avoid the malicious node that disrupt with the transmission, the sender will have to broadcast a "check" and provide a destination message reply would pursue the same route requested.

Shila et al. [81], introduced an approach to investigate selective forwarding attack (grayhole attack) in the wireless network. The algorithms were presented to defend against forwarding attack based on routing AODV. The algorithm used consisted of phase's counter-threshold, the technique that uses detection threshold and the data packet counter and detect the attacks, also query-based for acknowledgement the intermediate mobile nodes to localize adversaries.

Xiaopeng et al. [82], proposed a method to thwart denial of service (DoS) attacks. Grayhole attacks were used as a type of denial of service attacks. The Dynamic Source Routing (DSR) protocol and aggregate signature algorithm were used. Aggregate signature algorithms were used to trace packed dropping nodes. The approach consisted of three related algorithms – namely the proof-creating proof algorithm used to create proof, the check–up algorithm to check up source route nodes and the diagnosis algorithm to locate the malicious nodes.

Sengar et al. [83], proposed an approach of blackhole attack on the AODV routing protocol and its effects were elaborated by confirming how the attack disrupted the performance of MANETs.

Tamilselvan et al. [84], proposed a system for time-based threshold detection augmenting the original AODV. The technique sets a timer to measure time in the table for gathering requests from other nodes after having access to the first request. This will provide a packet's sequence consisting of a number that received the time in a Collect Route Reply Table (CRRT). For this reason the counting timeout value is based on incoming time and the initial route that has been requested.

Sardana et al. [85], proposed analysis of black hole performance occurring in AODV by always changing the mobile nodes and black hole. Metrics used for performance analyses which included end average delay.

Akbani et al. [86], proposed a reputation system for detection of malicious nodes in the network and for thwarting attacks such as the spread of viruses or worms. The proposed method collects information about past transactions and uses it to predict the future behaviour of the network. A Machine learning based Reputation System (RS) was used to devise the model and defend against many patterns of attacks. SVMs were used as the basis of the RS. Support Vector Machine (SVM) based RSs were evaluated and compared with RS trust guard.

Fatemeh et al. [87], proposed one class of SVM for anomaly detection for statistical learning in MANET occurring in AODV routing protocol named the ManetSVM. The approach suggested is in the form of a predefined model and was used to detect malicious behaviours such as flooding, black hole, and rushing among others. This technique was compared with other dynamic anomaly detection techniques for detection rate and false alarm rate.

Patel et al. [23], proposed support vector machine for detection and identification of packet dropping nodes in MANETs. An SVM was used to categorise them as either normal or malicious activity. The technique was implemented with AODV routing protocol. The method was evaluated using Packet Delivery Ratio (PDER), End-To-End delay, Average throughput, Normalized Routing Overhead, and Average Energy Consumption.

2.13 Chapter Summary

This chapter is a review of the background information regarding security related issues in MANETs for ensuring secured packets delivery over the multi-hop wireless channel. Furthermore, it provides information about machine learning whether supervised learning or unsupervised learning. In addition, it also explains key concepts about two types of ML techniques; SVM and LR and an overview of related work has also been explained, to buttress the direction of this research. It finally gives a survey of existing related works done on the security of MANETs regarding malicious attacks and identification of compromised nodes.

Chapter 3

Research Methodology

3.1 Chapter outline

This chapter presents the methodology employed to achieve the main goal of this research. It begins with the explanation of how the methods will be employed in the design of models to detect compromised nodes in MANETs. Given the models, it provides in each case, a detailed explanation about the model parameters and inputs. Moreover, it also provides details about the experimental settings, the collection of data sets, model design, and model evaluations. It further gives the analysis of how the designed models and algorithms work together to achieve the main goal of this research.

3.2 Overview

In this study, we investigate and propose a method that will identify and alleviate compromised nodes in MANETs in order to thwart the malicious adversary attacks with the help of the algorithms SVM and LR. Although several machine learning algorithms have been applied to automate the detection of meaningful pattern for intrusion detection in MANETs, in this chapter, machine learning algorithms of SVM and LR will be utilized to categorize generated data as either normal or abnormal in the network. Patel et al. [20] reported that the machine learning algorithms are far more resilient to the general changes in MANETs such as those due to malicious nodes changing their patterns over time or rapid changes in environmental factors.

3.3 Methodology

The methodology employed in this study is characterized by the goal of this research. The methods that were used to achieve the objectives include both qualitative and quantitative research methods. We discuss each of these in the following sections. In addition, we also discuss the various tools used for the research.

3.3.1 Qualitative research

Qualitative research is an exploratory research method used to gain and understand underlying reasons, opinions and motivations [88].

The qualitative method in this research is achieved using the following approaches:

1. ***Comprehensive literature review*** which has already been achieved in Chapter 2 and will be continued throughout the study. The objective here was to gain some understanding about MANETs and the challenges they face such as malicious attacks, as well as tools for mitigating those challenges; more specifically machine learning algorithms and so on.
2. ***Model design*** which is the core objective in this chapter. Literature studies in chapter 2 regarding identification of malicious attacks in MANETs were utilized as the premise to design the models for the study. Producing a model from the study has been utilized as part of the comparative examination. We propose the method of identification and mitigation of malicious activities in MANETs in order to thwart these malicious attacks using SVM and LR algorithms using data packets. Details about the formulation of the models and their parameters such as dependent and independent variables will be discussed in subsequent sections.

The classification results from the training data was used to answer the research questions stated in chapter 1, in particular, *How do we identify if a node in MANETs is compromised or not?* Data packets collected will be categorised into a number of class lists so as to fit and predict the results using identified algorithms. From the designed model we built a system that recognized actual input from data packets. In addition, we trained the system using data packets to make estimations and predictions about the awareness of malicious attacks.

3.3.2 Quantitative research

A Quantitative research method provides a clear picture of what to expect in the research compared to the qualitative research [89]. In particular, it deals with numerical computations. Thus in this research, numerical data about data packets was collected and utilized as inputs to the models designed based on the qualitative research method above. In this case, the data packets were trained for classification, detection and analysis of the compromised nodes.
Data packet was fitted onto the formulated mathematical models of the selected machine learning algorithms utilizing metrics of Packet Delivery (PDER), Packet modification (PMOR) and Packet Misroute (PMISR) [20]. Based on this, we will be able to categorise data packets as either normal or abnormal. If a data packet is normal it means it is not compromised, otherwise it is abnormal.

3.3.3Research tools

In this research, in order to achieve our aim as stated in Chapter 1, several tools were employed. In particular, to fit in the data packet collected into the models and to generate the results for analysis; we employed the Python programming language and other related packages such as sklearn, pandas, matplotlib and scipy. The choice of these tools is informed by that they have been tested and proven to be effective in the implementation of machine learning algorithms[90]. Simulation with python provides precise measurements and analysis of target concepts. The models were used to classify the data packet as either normal or abnormal by integrating it with simulated MANETs scenarios to automate the process.

Furthermore, the collected data was used to gather qualitative information about the nature of MANETs intrusion detection. However, by comparing the different response times from the classification and regression analyses, conclusions were drawn about the percentage accuracy of the models for awareness of cyber threats in MANETs. Chen et al. [91] reported that the confusion matrix can be used to evaluate the accuracy of models to gauge the performance.

3.4 Model design

This section presents a discussion of SVM and LR that have been selected as the algorithms to be used to perform the determinations in this study. The SVM model was used to maximize the margin between the two parallel hyperplanes for classification analysis, and the LR model for maximizing the probability of data packets.

3.4.1 Support vector machine (SVM)

The SVM technique is known as one of the good linear classifiers that improves the robustness of the classification and is not sensitive to the scarcity or correlation of the characteristics of data. Elish et al. [70, 92, 93] reported that, it remains one of the existing classification techniques in the field of data mining application such as software defect prediction, face recognition, and voice recognition. An SVM constructs a specific model that is complex enough to accommodate real world applications, and can also be simple enough to be analysed mathematically. Moreover, it provides linear and non-linear classifiers that offer time efficient training, prevention of over fitting in a high dimensional space and application of symbolic data.

SVM is used in a modular purpose and depends on the separability of feature space such as maximum margin classification for linear separable data. It allows some noise in the training data. A linear programming support vector machine can be used for classification purposes as well as regression problems. It also provides good practice for separating hyperplanes in higher dimensional space. However, Smola [72] proposed that the construction of such a hyperplane uses a maximum margin classifier as a linear separating machine.

As shown in chapter 2, equation 1 of SVM has the following format:

a) Optimal separating hyperplane

$w.\ x + b = 0$

and as shown in chapter 2, equation 2 and equation 3 have the following formats:

b) Two parallel hyperplanes

$w.\ x + b \geq +1 \ \ if\ y_i = +1$
$w.\ x + b \leq -1 \ \ if\ y_i = -1$

where x denotes the real value input pattern of metrics PDER & PMMR in our case, w is an orthogonal dimensional vector or weight vector that is to be computed by solving convex optimization [94] and b is a bias term used to identify the perpendicular of a distance from the origin to the hyper plane, $y_i = +1$ represents the normal data packet, and $y_i = -1$ represent the abnormal data packet.

The above functions will be utilized to formulate our model to detect nodes that have been compromised or attacked in MANETs.

However, the time taken and accuracy of the SVM is usually represented by the proportion of the accurate classifications. A small instance of data packets classifies well by finding the separating hyper plane In addition, Zanaty et al. [95] reported that in order to improve the accuracy of the SVM, a suitable linear kernel function is used and is appropriate for the training data and testing of variables, and is suitable for achieving better performance results. It maximizes the margin between the two parallel hyperplanes. The method gives good

results on pattern recognition and therefore knowledge is directly built into the algorithm, and can be used to generate training examples. Some support vectors characterize the solution to the problem such that if all the other training data are removed, and the system is retrained, the solution would be unchanged. Those support vectors which are not error prone are close to the decision boundary.

3.4.2 Logistic Regression

Dreiseitl et al. [26] reported that LR is a regression method used to make predictions based on a dependant variable. Pregibon et al. [96] stated that LR has been widely used for diagnostic measures to aid the analyst in detecting the observations or by quantifying the effect on various aspects. Ingersol et al. [74] also proposed that LR is widely used for predictions such as the presence or absence of characteristics in a powerful analytical technique when the outcome variable is dichotomous or have results that are based on the values of a set of predictor variables. It is equivalent to a linear regression model but is very useful in cases where the dependent variable is dichotomous [27]. LR is made up of one or more predictor variables that can either be continuous or categorical. It consists of the binomial type of possible dependent variables that can accept only two values "0" and "1" which are the likely outcomes of representations such as pass/fail, win/lose and others.

Furthermore, LR measures the relationship between the dependent variable and some independent variables by producing probabilities consisting of logit function. Harrell et al. [75] also proposed that it is a special case of a generalised linear model, which is made up of different assumptions based on the relationship between the dependent and independent variables.
LR has the following format:

As shown in chapter 2, equation 4 is the Logistic function that is expressed as

$$f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

where t can be represented as

$$t = \beta_0 + \beta_1 x$$

However, *f* denotes the probability function where t is time taken for the packet delivery and packet modification and x is independent packet represented as the metrics PDER and PMMR, *f(x)* denotes the probability of events that will be occurring or not as the results of independent packets being transmitted in the network of MANETs, β_0 is the y intercept from the linear regression equation (a criterion when the predictor $\beta_1 x$ is equal to zero), $\beta_1 x$ denotes the coefficient to the regression by predictor value of packet delivery and misroute rate, e denotes the value 2.7182, the exponential function of the probability of the LR model.

The accuracy of the LR model comes with a powerful tool that predicts a dependent variable from a set of predictors. However, LR performance analysis of the obtained model results in a good classification and regression analysis or fitting results that are an accurate prediction of the outcomes based on the subjects of interest. LR deals with the quality of fit and how to evaluate its performance in order to avoid poorly fitted models. Coefficients are estimated in order to maximize the overall likelihood of the data packets, however the maximum likelihood method used to provide estimation comes with a mathematical maximization procedure, providing the opportunity for capitalization on chances. LR takes little time to perform classification because data is very small to fit and validate. Giancristofaro et al. [97] proposed that, the ability of LR to distinguish correctly between classes is based on the separation of subjects with different responses.

3.5 Model parameters

This section discusses the types of variables used in this research. The parameters were used as inputs to data packets. It covers the way the independent variables and dependent variables were processed. It also discusses how the variables work together to achieve the main goal of this research.

3.5.1 Independent variable

An independent variable is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable. The objective of an independent variable is to influence a dependent variable by causing a change. In the context of this research, MANETs data packets (x_1, x_2) in the form of metrics PDER and PMMR which are

inputs to the models. These metrics are recorded in a list of arrays and fit the normal and abnormal packets. These variables are also unknown variables or unseen data sets and can be manipulated. They are represented as binary arrays of values of data packets in the set of vectors $x = (x_0, x_1, x_p)$ and determines the prediction of the dependent variable.

3.5.2 Dependent variable

A dependent variable is the: the outcome based on the independent variable. The objective of a dependent variable is to determine values that are to be predicted. These values include the targets (y_1,y_2) for data packets. The class's value depends on the feature reading of metrics (x_1,x_2) entered in the form of arrays. Once the experimental value of metric changes, a dependent variable is observed and recorded. The dependent variable outcome can account for dichotomous variables solutions that allow the results to be classified as "Normal" "1" or "Abnormal" "Intrusion" "0" [96].

3.6 Model formulation

In this section we describe how to formulate our model using the two chosen techniques, namely SVM and LR.

3.6.1 Model formulation using SVM

Consider the problem of separating data packets that belong to the following two separate classes: normal and abnormal packets. As shown in chapter 2 by equation 1,

$$w.\ x + b = 0$$

is the line that separate two classes; normal and abnormal packets, x represents the subsets of data packets of metrics PDER & PMMR, w is an orthogonal dimensional vector or weight vector that is to be computed by the convex optimization [94] and b is a bias term used to identify the perpendicular of a distance from the origin to the optimal separating hyper plane as shown in chapter 2 by equation 2 and equation 3.

$$w.\ x + b \geq +1 \quad if\ y_i = +1$$

$$w.\ x + b \leq -1 \quad if\ y_i = -1$$

The two equations represent the two parallel hyperplanes that separate the two classes of data packets into normal or abnormal packets. Equation 2 denotes a linear separating hyper plane of data packet with support vectors of the normal packet. The label $y_i = +1$ represents the dependent variable of class of normal packets. Equation 3 denotes the separating hyperplane of abnormal packets. The label $y_i = -1$ represents the dependent variable of abnormal packets, where x indicates the input data of binary values; PDER and PMMR.

3.6.2 Model formulation using LR

LR is given by

$$f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

where for independent variables, f(t) is the function of the generated data packet, t denotes metrics PDER and PMMR of data packet. For dependent variables, f(t) denotes the "normal packet" or a binary "1" or "abnormal packet" or binary "0", hence f(PDER & PMMR) = "0" or "1"equalling "Abnormal" or "normal" respectively.

Accordingly,

$t = \beta_0 + \beta_1 x$

where t denotes the combination of the metrics PDER and PMMR and coefficients β_0 represents the y intercept (the value of the criterion when the metric $\beta_1 x$ is equal to zero). $\beta_1 x$ is the regression coefficient computed multiplied by some value of the metrics of data packets.

3.7 Model evaluation

This section discusses the evaluation of the models that will be used for the experiments, this will be achieved by finding the best model that represents data packets as normal or abnormal packets, and how well the chosen model will work in the future. Hence for evaluation, a confusion matrix and cross validation will be used.

3.7.1 Confusion matrix

Fawcett et al. [98] reported that the objective of the confusion matrix or error matrix is to come up with a summary of prediction results on a problem of the classification. For evaluation of both SVM and LR models, the confusion matrix will be used as follows:

For operational processing of data packet, two classes of data packets using PDER & PMMR are divided into the train, test and split methods. These methods verify the number of correct and incorrect predictions for data packets and summarize the count values for each class. The methods confirm confusion for predictions of the models, and computed using the Python programming language as shown by Appendix A and B.

A confusion matrix checks the accuracy of classification and regression analysis by dividing data packet into y test and y predict. The y test denotes the test class label; y predict is the prediction of the y test of class label of data packet. However, each row of a matrix represents the actual conditions of data packets while each column denotes the predicted conditions of data packets, and hence compares the performance of SVM and LR algorithms. The confusion matrix process can be represented as in Figure 3.1.

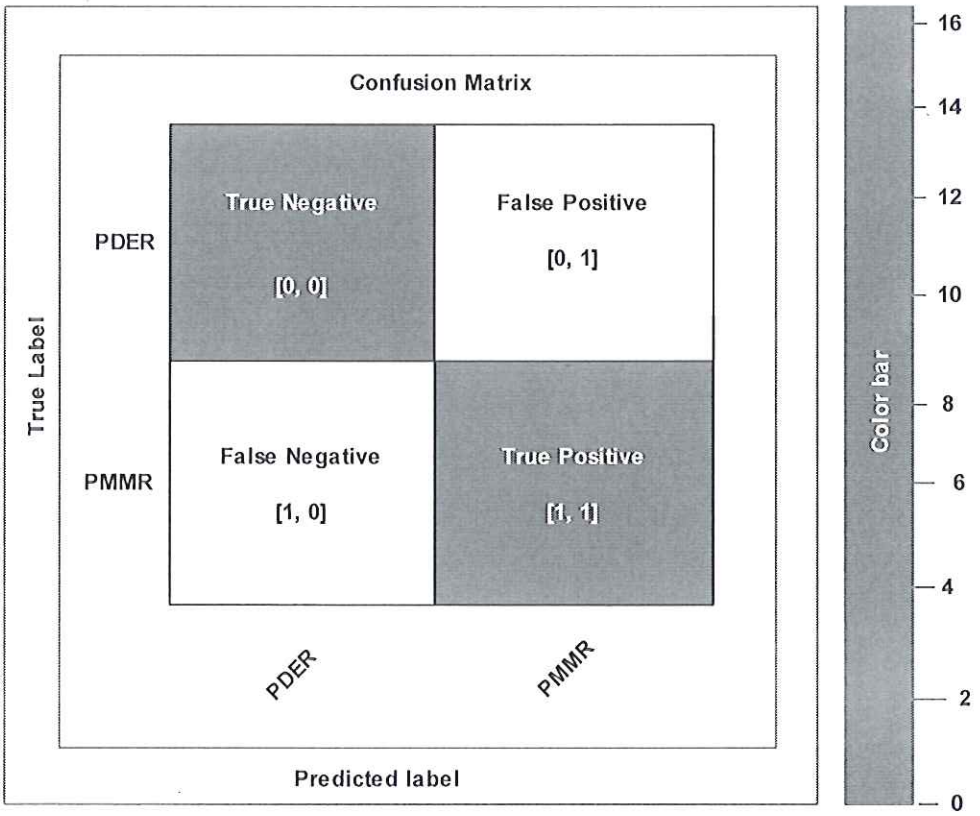

Figure 3.1 Confusion matrix process

Data packets will be classified as true positive, true negative, false negative and false positive as shown in Figure 3.1. Fawcett et al. [98] also reported that the objective of a confusion matrix or error matrix is to provide a summary of prediction results based on the classification problems. However, the classifier together with instances performs five possible outcomes as follow:

1. True positive: Involves the situation where a true data packet of predictions is a positive data packet [1, 1].
2. True negative: Considers the situation that involves incorrect data packet of predictions that is negative data packet [0, 0].
3. False positive: Involves a case that involve the correct predictions of data packet where the data packet is abnormal [0, 1].
4. False negative: Involves incorrect predictions of data packet that involve the positive data packet [1, 0].
5. Colour bar: The units on the colour bar represent the number of data packets. Vertical bar represents true data packet; while the horizontal bar represents the data packet that is predicted.

3.7.2 Cross validation

Kumar et al. [99] reported that cross validation is a statistical operation of comparing and evaluating the learning algorithms based on the training observations by dividing data into two segments. For evaluation of both SVM and LR models, cross validation will be used as follows:

For the evaluation process, data packets are partitioned equally into two segments or folds: learning and validation of the models. During the evaluation, these two segments cross over in successive rounds. During the experimental process of data packet, the training and validation are performed in such a way that the data packets are used for validation and the other remaining data packets are for learning the system.

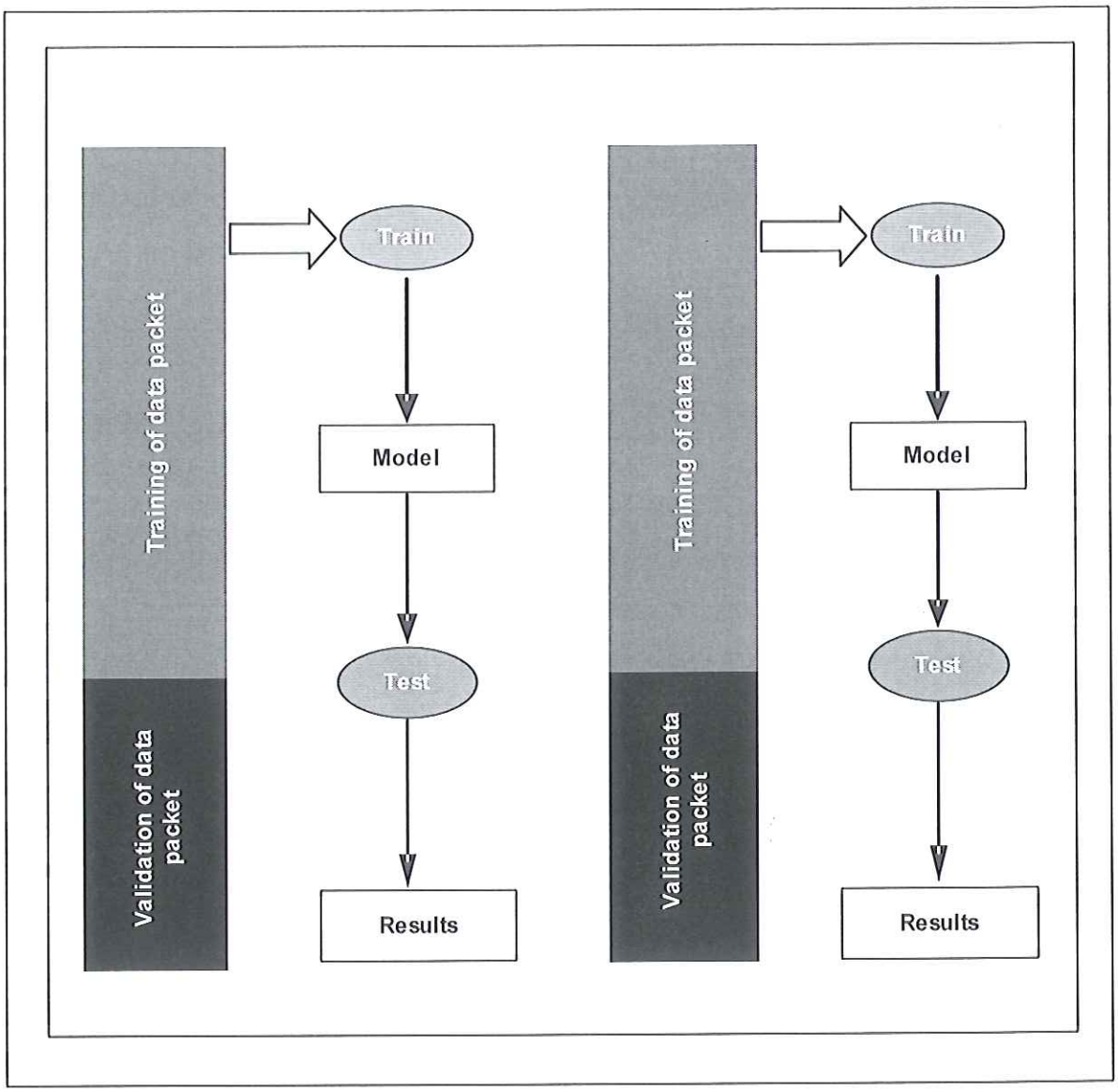

Figure 3.2 Procedure of two folds cross validation

Data packets will be partitioned as in Figure 3.2, the blue colour represents the learning of data packets, and the black colour represents the validation of the data packet. During the research the algorithms use data packets in order to learn from the models and make predictions for the validation. Hence the performance of each model is tracked using a predetermined performance metric accuracy.

3.8 Experimental design process

This section provides detailed information about the experimental design process to achieve the objectives of this research. It also discusses the metrics used for the tasks and provides information about how the setup of the experiment is conducted and how the algorithms are utilized. Moreover, the operation of the experiment is explained.

3.8.1 MANET data and compromised node detection

This section provides detailed information about the experimental design process to achieve the objectives of this research. It also discusses the data packet obtained using metrics for the experiments and provides information about how the setup of the experiment is conducted and how the algorithms are utilized. Moreover, the operation and accuracy issues of the experiment are explained.

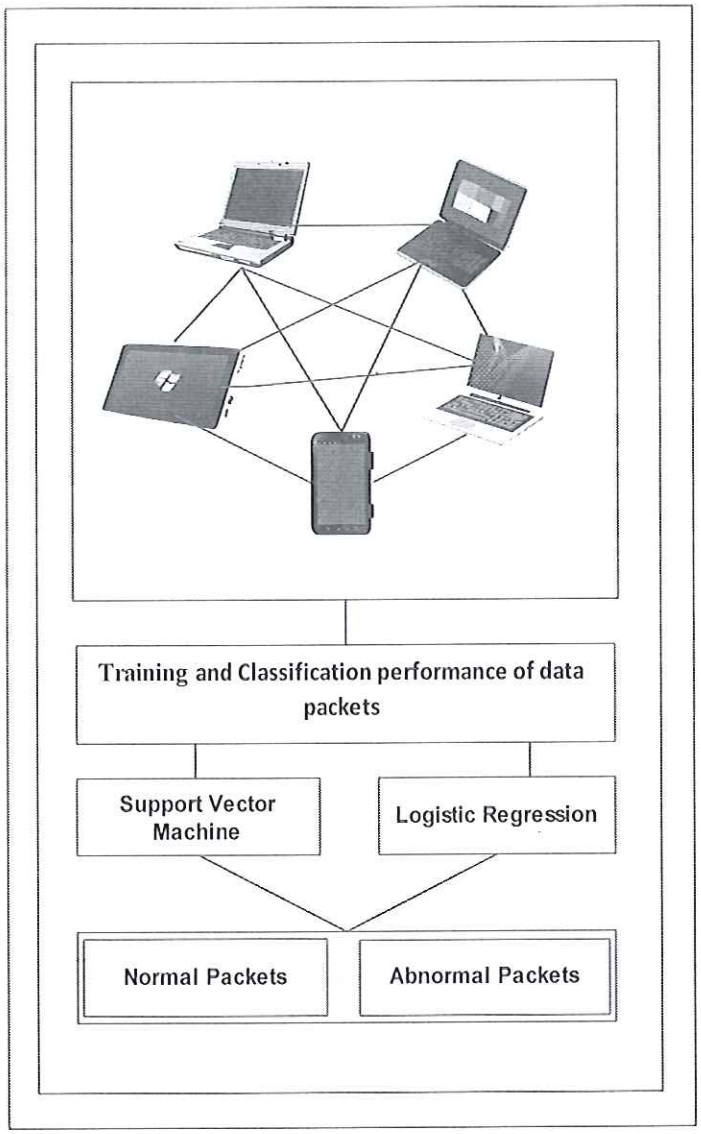

Figure 3.3 SVM and LR classification model

Based on the network setup in Figure 3.3, data packets generated were be collected and utilized. However, in order to build the model that automatically identifies nodes that have been attacked or not attacked; this study used existing historical data. In this we used the data set that has previously been collected on the metrics of PDER and PMMR where nodes were classified as either attacked or not attacked. This data was used in the model as training data and testing. As such, subsequently generated data will be used to identify nodes that are prone to attacks so that early mitigation plans can be initiated.

For the purpose of the experiments in this research, we designed a binary classification for SVM and LR classifier for data packet where the dependent variables were either ("0" or "1") and the independent variables were PDER and PMMR. The data packet was organized as records each of which consisted of a vector of attributes X = $(x_1, x_2, , x_M)$ followed by the target Y ϵ (y_1, y_2, y_M) where M is the number of the n^{th} attribute of the data packet.

3.8.2 System requirements and tools

To determine which model is more effective in terms of the accuracy of intrusion detection, the machine learning algorithms of SVM and LR were employed utilizing the following different software and related packages: pycharm-professional-5.0.3 python, scipy, pandas, numpy, and matplotlib.

Moreover, the study was run on a personal computer (PC) with windows 7 or later version ultimate operating system having core i5 Intel processor and 4 Gig random access memories (RAM).

3.8.3 Data collection

The dynamic nature of MANETs makes normal nodes in the network vulnerable to attacks by malicious nodes, this leads to them being susceptible to adversary attacks and become compromised as a result. In order to build the models, we used historical data about iris flower dataset that has been widely used for classification and regression analysis [24]. Though the data is not related to MANETs, it has been used for typical test cases of many statistical classification techniques in machine learning that provide classification techniques for application fields in data mining, specific examples include software defect prediction, face recognition, and voice recognition [27, 63]. It is also used for cluster classification and regression analysis since it contains two clusters that are easily separated [91].

The dataset used is based on the iris flowers which is collected from the source "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.dat"[91]. This dataset contains 150 observations of the iris flowers. This includes the four columns of the measurements of the flowers in centimetres. It also contains the fifth column which is the species of the flower that is observed and all these belong to one of the three species. It is

sometimes referred to as Anderson's Iris data set because the collected data were used to quantify the structure of the iris flowers of three related species.

There are 50 samples from each of the three species of Iris which are Iris setosa, Iris virginica, and Iris versicolor. This consists of four features that were measured from each of the samples. These samples consist of the length and width of the sepal and petal that were measured in centimetres. The dataset is widely used for a typical test case for many statistical classification techniques in machine learning. The dataset is also widely used for cluster analysis since it contains two clusters that are easily separated. The first cluster contains the Iris setosa and the other contains both the Iris virginica and Iris versicolor.

The justification for using this dataset is that we were not able to get enough data about PDER and PMMR from the network set. The assumption is made that the size of data collected represents the size of data packets travelling in a MANET. However, when an ad hoc network is formed these data packets in MANETs can communicate with each other wirelessly. During the communication, data packets are delivered for communication from the source to destination and this is referred to as PDER metric. However, while communicating some data packets are modified and misrouted by malicious attackers in MANETs and these are referred to as PMMR.

3.8.4 Iris dataset manipulation

As stated in the above section, the Iris data used in this research are not related to MANETs and the nature of the data values is quite different from the values of PDER and PMMR. Thus, to effectively use this data, we have to make some adjustments to them in order to get the data packet values closer to MANETs the nature of data values or the size of packets used by the MANETs network.

In this case, for the three classes which are the *Iris setose* (50), *Iris versicolor* (50) and *Iris virginica* (50) having four features each: the sepal length, sepal width, petal length and petal width, we performed average computation for sepal length and sepal width as well as for petal length and petal width.

Thus,

Av_SepalIris = (Sepal length + Sepal width) / 2

Av_PetalIris = (Petal length + Petal width) / 2

Hence the size of metrics is generated to equal the size of packets delivered, packet misroute rate and packet modification. This is important as a representation of data collected and used as packets in the network. Hence, we come up with the following conversions:

Av_SepalIris / 2 = PDER, and
Av_PetalIris / 2 = PMMR

Also to build the model, Iris_setosa (50), Iris_versicolor (50) which will be labelled as Iris setosa "1" or "Normal" and Iris_versicolor "0" or "Intrusion" will be used as train data for the developed model.

3.8.5 Model parameters and inputs

We read the input data as metrics PDER and PMMR into a powerful Python toolkit (Pandas) using two methods for randomly training data packet and testing phase for prediction integrating with machine learning algorithms LR and SVM. In addition, the data packet is fitted to the separated classes such that the function of metric f *(PDER & PMMR)* = (Iris _setosa or Iris_versicolor) is predicted using labelled data packet.

3.9 Model operations

When running the LR and SVM experiments, the information gained based on PDER and PMMR will be collected for each model and analyzed as follows:

1) The performance classification models of LR and SVM are analyzed on data packets for which the true values are known. The actual and predicted conditions; true positive, true negative, false positive and false negative are collected for comparison.
2) The classification accuracy for both models: the percentage of correct predictions of data packets which was calculated using the equation (TP +TN) / (TP + TN + FP + FN) was tested.

3) The classification error of misclassification performance for LR and SVM with the equation (FP + FN) / (TP + TN + FP + FN) was determined.
4) The sensitivity of the algorithms which are calculated with the equation TP / (TP + FN) was tested, i.e. when the actual value of data packet was positive, how often was the prediction correct.
5) The specificity for both models of data packet which is calculated with the equation {TN / (TN + FP)} was looked into, i.e. how often did it predict no, when the actual value for data packet was false.
6) False positive rate, that ask the question when the actual value is negative how often does it predict incorrect {FP / (TN + FP)} was computed.
7) Precision: when the positive value was predicted accurate, how often was the prediction correct; {TP / (TP + FP)}.

The process involved is captured in Figure 3.3

3.10 Chapter summary

The chapter began with presentation strategies of how the model is established to achieve the objective of this research. Explanations of machine learning as a widely used tools/techniques for detection of meaningful patterns in dataset were also discussed. The two research methods, how qualitative and quantitative methodology were used and explained to achieving the research objectives. Next, the measurements setup and research tools such as Python, anaconda, numpy, scipy, matplotlib, sklearn and others were also presented as part of the research. The chapter also presented the experimental design process. Comparison and evaluation of learning algorithms based on the training observations by dividing data into two phases namely: were identified as training and testing.

Chapter 4

Discussion of Results

4.1 Chapter Outline

This chapter presents a detailed discussion and interpretations of results obtained from the experimental set-up reported in chapter 3. It begins with an explanation of the classification results from the models to detect compromised data packets within the network as well as to give the comparative analysis of how the designed models and algorithms work together to achieve the main goal of this research. Moreover, it proposes a framework for compromised nodes detection using the designed models in MANETs.

4.2 Overview

In this chapter, we discuss and analyse the results from the method of identification and alleviation of compromised nodes in MANETs in order to thwart the malicious attacks using SVM and LR algorithms. The essence is to identify which of the designed models is more efficient in detecting compromised nodes and then develop a framework for automatic detection of attacks on MANETs.

To achieve the outlined objectives, we employed supervised machine learning algorithms to perform classification on existing historical data [91] to determine accuracy, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision. In this case, any model found to be effective in terms of high accuracy in the classification will be employed in the development of a detection framework to automatically detect patterns as either normal or abnormal packets in MANETs.

The choice of SVM and the LR as supervised machine learning algorithms is based on their widespread acceptance as the best approach for data modelling that provides a good generalization performance [100]. They also very effective in classification and analysis of datasets [72].

Thus, this chapter will perform classification on the assumed data packet which will be analysed and evaluated for the development of a detection framework for identification of normal and abnormal packets.

4.3 Classification Results

This section presents the actual results of the classification performed using the chosen SVM and LR as discussed in chapter 3. With these models, we utilized the data packets as model metrics and other variables to generate the performance results.

Given the results obtained using python, different results were generated for SVM and LR algorithms when utilizing the data packets. These results of the classification were then analysis and compared. To this end the data packets used were classified as normal packets (50) and abnormal packets (50).

We evaluated the performance of both SVM and LR in three iterations and the analysed them in terms of accuracy, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision. The goal is to find the best model that accurately classify data packets as either normal or abnormal packets, and how well the such model will operate in future.

4.3.1 Support Vector Machine Classification

The results of the classification based on SVM are presented in this section. To determine the effectiveness in terms of the accuracy of intrusion detection, we measure PDER and PMMR using the following software packages: pycharm-professional-5.0.3 python, anaconda2-2.4.1, scipy, pandas, numpy, and matplotlib.

The proposed algorithm for SVM is shown on Appendix A and the results obtained are as follows.

a) Scatter plot as displayed in Figure 4.1.
b) Confusion matrices – for first, second and third iteration as shown in Figures 4.2, 4.3, and 4.4.
c) The results for the iterations were recorded in Tables 4.1, 4.2 and 4.3, respectively.

1) Scatter plot

This section presents the actual results of the classification performed using the chosen SVM algorithm, the scatter plot of the classification results for SVM model is captured in Figure 4.1.Two classes of data packets – were obtained using PDER and PMMR as metrics. The red dots represent the PDER, and the blue represents the PMMR.

Figure 4.1 illustrates the classification results of the two metrics utilized: PDER and PMMR which were used to analyse the classification performance. In Figure 4.1, we observed that SVM produced an optimal separating hyperplane or a decision boundary between the normal and abnormal data packets. The separation is based on the linear classifier method which is also based on a linear kernel function of binary metrics. The results show excellent performance in the classification without misclassification or under fitting between two categories. In this case, when data packets travel in the network, the data packets received were classified in real time as either normal or abnormal. The results show that SVM is excellent in classifying data packets. This is very important because knowing early when a node is compromised is critical to devising a means to mitigate the situation.

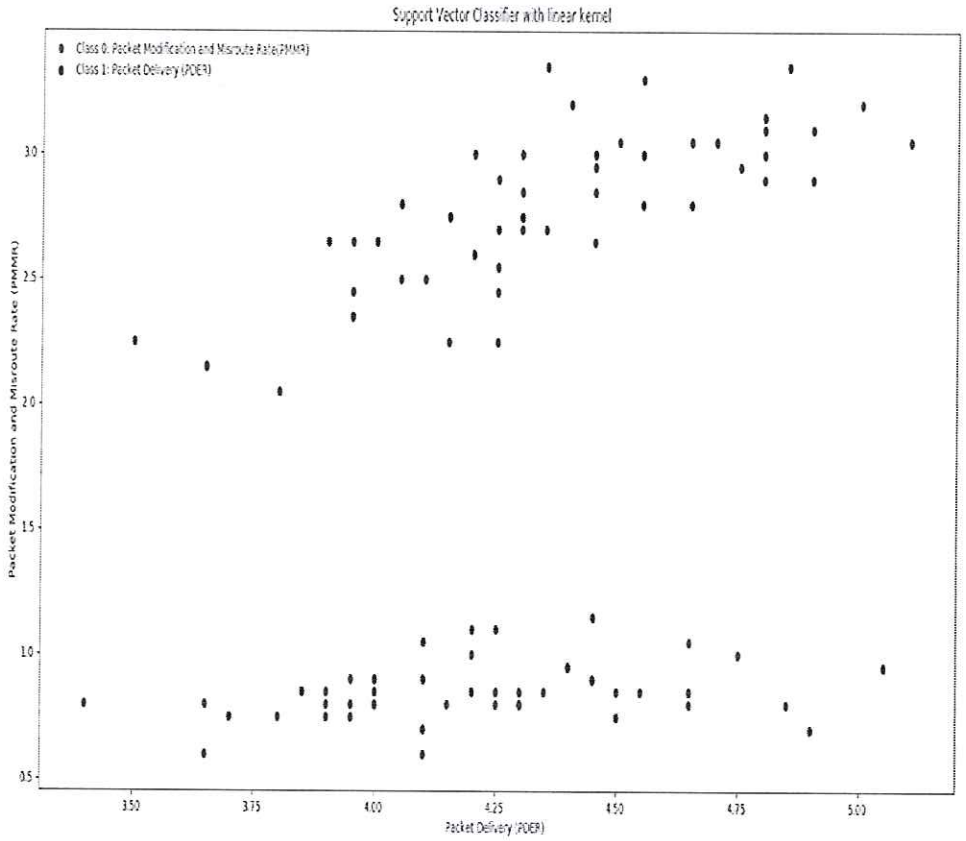

Figure 4.1. SVM classification plot

2) Confusion matrix

A confusion matrix shows the accuracy of the prediction summary results on a problem of the classification of data packet. Based on the data packets utilized, the classification for SVM is presented based on the following evaluation criteria: accuracy, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision. Data packets are evaluated and labelled as true positive by the algorithm when the data packet attack detected is actually data packet attack, on the other hand, true negative is when the normal data packet detected is actually normal data packet. False positive is the situation when normal data packet is detected as data packet attack, and false negative is when data packet attack is detected as normal data packet. Accuracy refers to the quality classification of data packet of being correct or precise. Sensitivity refers to the proportion of

identifying normal data packet that is correctly classified. Specificity is the proportion of abnormal data packets that are correctly identified. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets. Precision refers to the differentiation of random errors and accuracy between the normal and abnormal data packets. To record the performance of data packets in the network, the confusion matrix and cross validation were observed as follows in three iterations.

a) First iteration SVM classification

Figure 4.2 is the confusion matrix and cross validation results obtained from the classification of data packets in the first iteration. The false negative is equal to the false positive – which is 0 % as shown in Table 4.1, meaning data packets travelled from the source to the destination. This allows the data packets to move from source to destination thus letting services to work properly. The higher the number of classification accuracy of data packets the higher the number of packets moving freely in the network without data packets being dropped. However, for communication and delivery of data packets in the network, the percentage of true positive – of 12 % and true negative – of 13 % as shown in Table 4.1, means that as the data packets are delivered, modification has not taken place – and, as a result, this cannot affect and degrade the network performance.

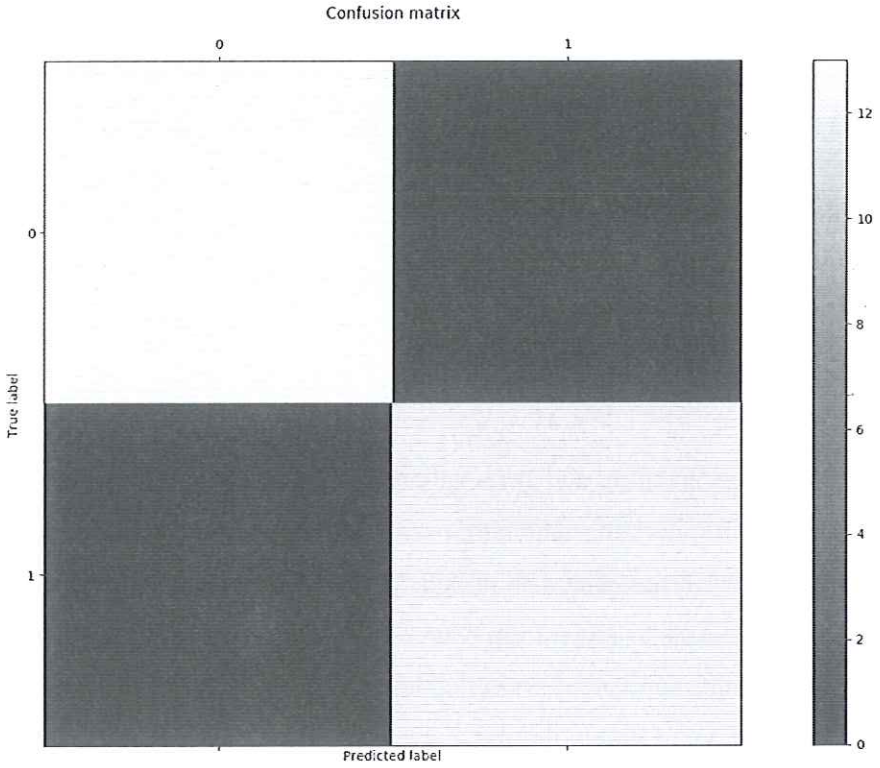

Figure 4.2: first iteration SVM classification

Table 4.1 shows that the functional classification of communication accuracy between the data packets in the MANET is 100 %. The colours represent the labelled true positive, true negative, false positive and false negative. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 0 %, showing – that the data packet is classified correctly. Sensitivity shows the proportion of identifying normal data packets that are correctly classified – and it is 94 %. This shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of abnormal data packets that are correctly identified – which is 84 %, shows the percentage of normal data packets that are correctly identified as abnormal data packets. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets – and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets – and it is 92 %, indicating the difference between a measured normal and abnormal data packet.

The results of the first iteration are recorded in Table 4.1.

Table 4.1: Confusion matrix & cross validation evaluation of SVM

Evaluation Parameter	First iteration (%)
Accuracy: (% of correct prediction)	100
True Positive	12
True Negative	13
False Positive	0
False Negative	0
Misclassification	0
Sensitivity	94
Specificity	84
False Positive Rate	0
Precision	92

b) Second iteration SVM classification

Figure 4.3 is the confusion matrix and cross validation results obtained from the classification of data packets in the second iteration. The false negative is equal to the false positive – which is 0 % as shown in Table 4.2, meaning data packets travelled in the network from the source to the destination. This allows the data packets to move from source to destination thus letting services to work properly. The higher the number of classification accuracy of data packets the higher the number of packets moving freely in the network without data packets being dropped. However, for communication and delivery of data packets in the network, the percentage of true positive – of 9 % and true negative – of 16 % as shown in Table 4.2, means that as the data packets are delivered, modification has taken place – and, as

a result, this can affect and degrade the network performance and propagate incorrect routing information.

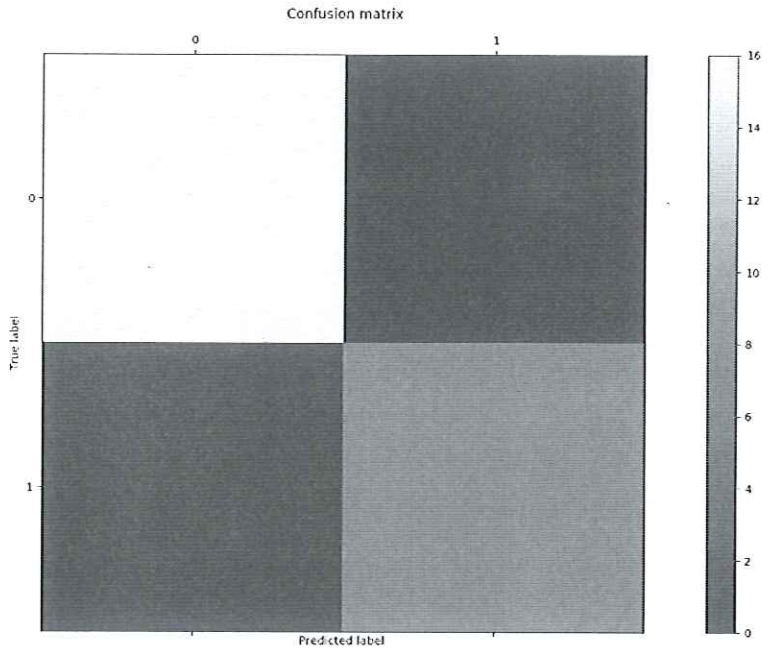

Figure 4.3: second iteration SVM classification

Table 4.2 shows that the functional classification of communication accuracy between the data packets is 100 %. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal data packet, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 0 %, showing – that the data packet is classified correctly. Sensitivity shows the proportion of identifying normal data packets that are correctly classified – and it is 94 %. This shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of abnormal data packets that are correctly identified – which is 86 %, shows the percentage of normal data packets that are correctly identified as abnormal data packets. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets – and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets – is 90 %, indicating the difference between a measured normal and abnormal data packet.

The results of the second iteration are recorded in Table 4.2.

Table 4.2: Confusion matrix & cross validation evaluation of SVM

Evaluation Parameter	Second iteration (%)
Accuracy: (% of correct prediction)	92
True Positive	9
True Negative	16
False Positive	0
False Negative	0
Misclassification	0
Sensitivity	94
Specificity	86
False Positive Rate	0
Precision	90

c) Third iteration SVM classification

Figure 4.4 shows the results of the classification obtained in the third iteration. The prediction of false negative – of 1 % slightly differs from that of false positive – which is 0 % as shown in Table 4.3, which means there are normal packets that have been modified and dropped in the network – and hence intrusion has been detected. Also, the percentage of true positive – which is 14 % and true negative – which is 11 % as shown in Table 4.3, imply that the normal data packets have been modified and predicted abnormal, as a result this could degrade the network performance or prevent other services from working properly.

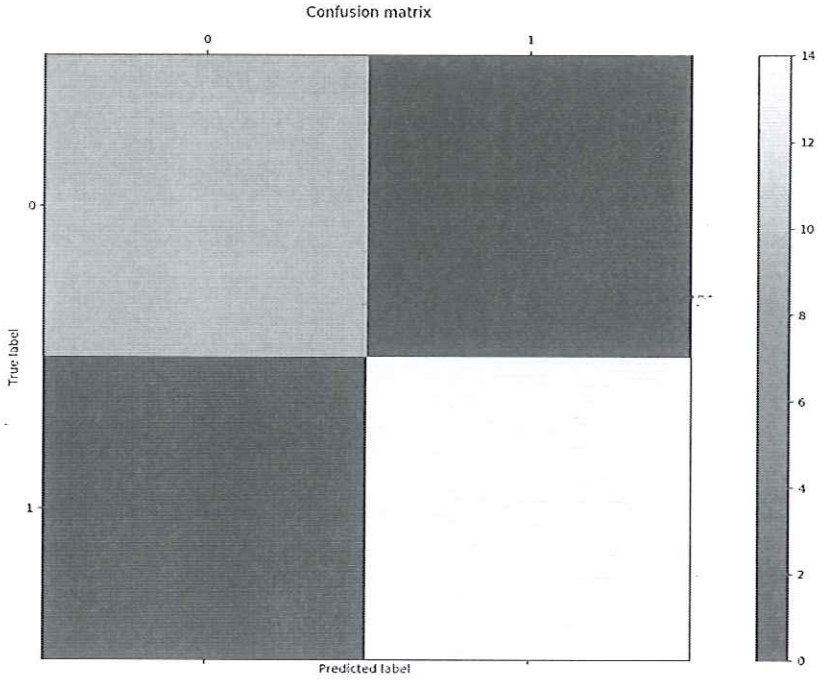

Figure 4.4: Third iteration SVM classification

Table 4.3 shows that the functional classification of communication accuracy between the data packets in the MANET is 94 %. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 4 %, showing – that the abnormal data packet is noticed and classified. Sensitivity shows the proportion of identifying normal data packets that are correctly classified – and it is 93 %. This shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of abnormal data packets that are correctly identified – which is 90 %, shows the percentage of normal and abnormal data packets that are correctly identified. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets –and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets – is 90 %, indicating the difference between a measured normal and abnormal data packet.

The results of the third iteration are recorded in Table 4.3.

Table 4.3: Confusion matrix & cross validation evaluation of SVM

Evaluation Parameter	Third iteration (%)
Accuracy: (% of correct prediction)	94
True Positive	14
True Negative	11
False Positive	0
False Negative	1
Misclassification	4
Sensitivity	93
Specificity	90
False Positive Rate	0
Precision	90

Table M captured the information collected randomly SVM model while utilizing data packet. Table M shows information about the classification as either normal or abnormal data packet when delivered in the network. The confusion matrix captured in Figure 4.2, 4.3 and 4.4 respectively and generated in SVM and is shown in Table M.

Table M: Confusion matrix for first, second and third iteration SVM

Evaluation Parameter	First iteration (%)	Second iteration (%)	Third iteration (%)
Accuracy: (% of correct prediction)	100	92	94
True Positive	12	9	14
True Negative	13	16	11
False Positive	0	0	0
False Negative	0	0	1
Misclassification	0	0	4
Sensitivity	94	94	93
Specificity	84	86	90
False Positive Rate	0	0	0
Precision	92	90	90

4.3.2 Logistic Regression Classification

The results of the classification based on LR are presented in this section. To determine the effectiveness in terms of the accuracy of intrusion detection, we measure PDER and PMMR using the following software packages: pycharm-professional-5.0.3 python, anaconda2-2.4.1, scipy, pandas, numpy, and matplotlib.

The proposed algorithm for LR is shown on Appendix B and the results obtained are as follows.

a) Scatter plot as displayed in Figure 4.5.
b) Confusion matrices for first, second and third iterations as shown in Figures 4.6, 4.7, and 4.8.
c) The results for the iterations were recorded in Tables 4.4, 4.5 and 4.6, respectively.

1) Scatter plot

This section presents the actual results of the classification performed using the chosen LR algorithm, – as captured by the scatter plot in Figure 4.5.Two classes of data packets – were obtained using PDER and PMMR as metrics. The red dots represent the PDER, and the blue represents the PMMR.

In particular, LR provides knowledge of the relationships and strength among the data packets travelling in the network. In terms of predictions, the presence or absence of characteristics, LR classified excellently the data packets into specific categories by learning from the delivered data packets. In this case, normal or abnormal data packets classification presents 100 % accuracy. This indicates that data packets were delivered as normal without intrusion or degradation on the network performance, no incorrect routing information was propagated, and there was no prevention of services from working properly. Abnormal packets in the MANETs network were noticed when normal data packets were being compromised.

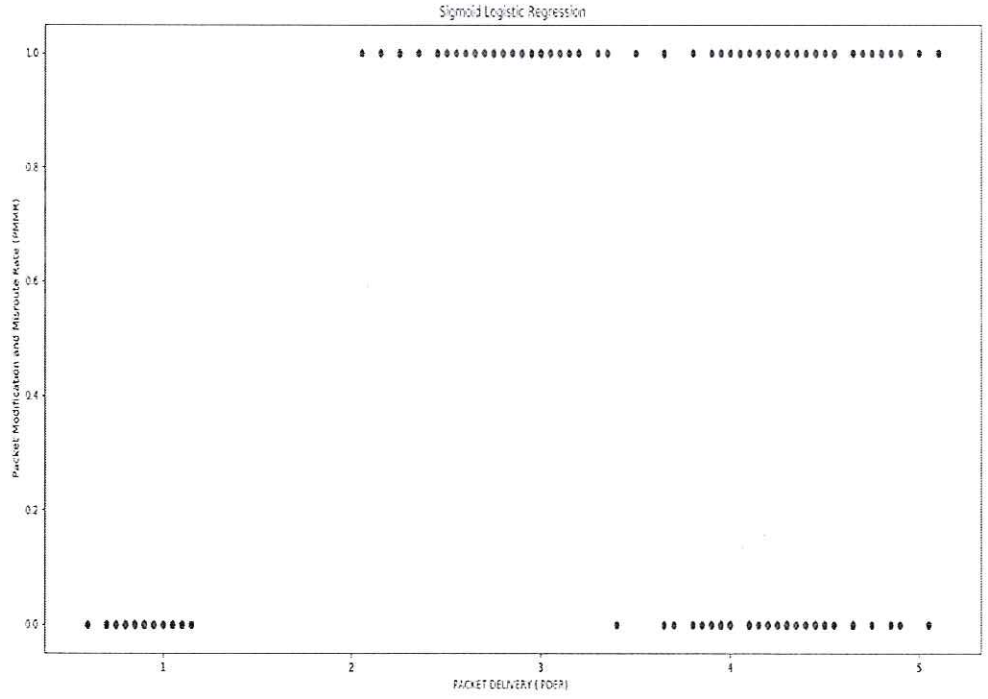

Figure 4.5. LR classification plot

2) Confusion Matrix

Figure 4.6 shows the classification in terms of accuracy using the data packets for LR classification. The same parameters are utilized as in the SVM evaluation. Thus, data packets are evaluated when transmitted from the source to the destination by considering the validity between packets as: accuracy, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision. Data packets are evaluated and labelled as true positive by the algorithm when the amount of attack data detected is actually attack data, on the other hand, true negative is when the amount of normal data packet detected is actually normal data. False positive is the situation when normal data packet is detected as attack data packet, and false negative is when attack data packet is detected as normal data packet. Accuracy refers to the quality classification of data packet of being correct or precise. Sensitivity refers to the proportion of identifying normal data packets that are correctly classified. Specificity is the proportion of abnormal data packets that are correctly identified. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets. Precision refers to the differentiation of random errors and accuracy between the normal and abnormal data packets. The confusion matrix for the iterations is captured showing the accuracy of classification of the data packets transmitted in the network of MANETs as either normal or abnormal.

a) First iteration LR classification

Figure 4.6 shows the difference for the percentage between true positive – 12 % and true negative – 13% as shown in Table 4.4 is minimal meaning that the data packets are delivered without being modified in the network. First iteration shows that false negative – of 0 % is equal to false positive – of 0 % as shown in Table 4.4 which makes data packets to travel safely in the network without being modified or degrading the network perfomance.

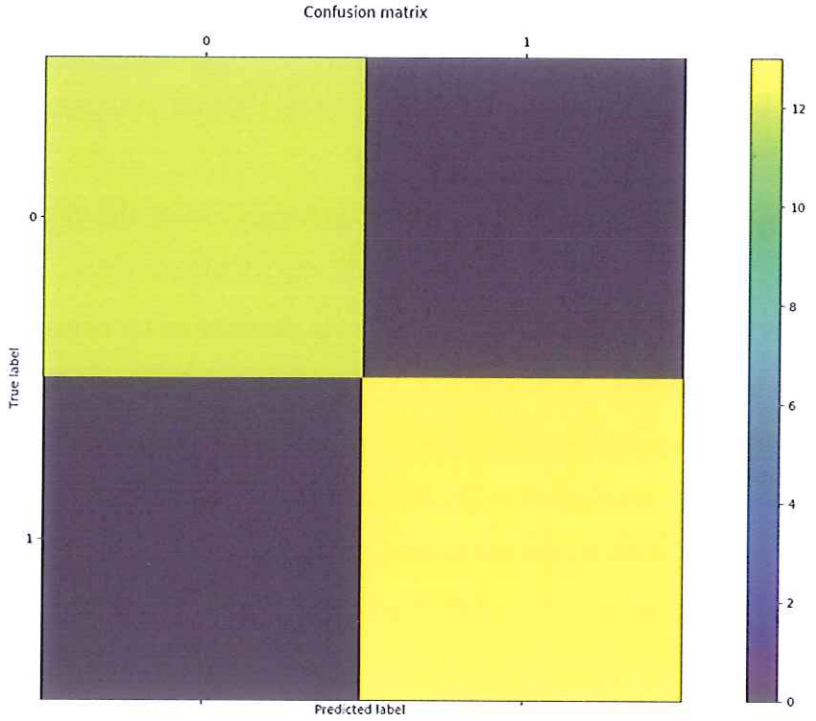

Figure 4.6: First iteration LR classification

Table 4.4 shows that the functional classification of communication accuracy between the data packets is 100 %. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 0 %, showing – that the data packet is classified correctly. Sensitivity shows the proportion of identifying normal data packets that are correctly classified and it is 100 %. This shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of abnormal data packets that are correctly identified – which is 100 %, also shows the percentage of normal data packets that are correctly identified. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets – and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets – is 100 %, indicating the difference between a measured normal and abnormal data packet.

The results of the first iteration are recorded in Table 4.4.

Table 4.4 Confusion matrix & cross validation evaluation of LR

Evaluation Parameter	First iteration (%)
Accuracy (% of correct predictions)	100
True Positive	12
True Negative	13
False Positive	0
False Negative	0
Misclassification	0
Sensitivity	100
Specificity	100
False Positive Rate	0
Precision	100

b) Second iteration LR classification

Figure 4.7 is the confusion matrix and cross validation results obtained from the classification of data packets in the second iteration. The false negative is equal to the false positive – which is 0 % as shown in Table 4.5, meaning data packets travel safely in the network without being modified or without degrading the network perfomance. This allows the data packets to move from source to destination thus letting services to work properly. The higher the classification accuracy of data packets the higher the number of packets moving freely in the network without modification and data packets being dropped. However, for communication and delivery of data packets in the network, the percentage of true positive – of 9 % and true negative – of 16 % as shown in Table 4.5, means that as the data packets are delivered in the network, modification of data packets have been noticed and identified in the network – and, as a result, this affect and degrade the network performance.

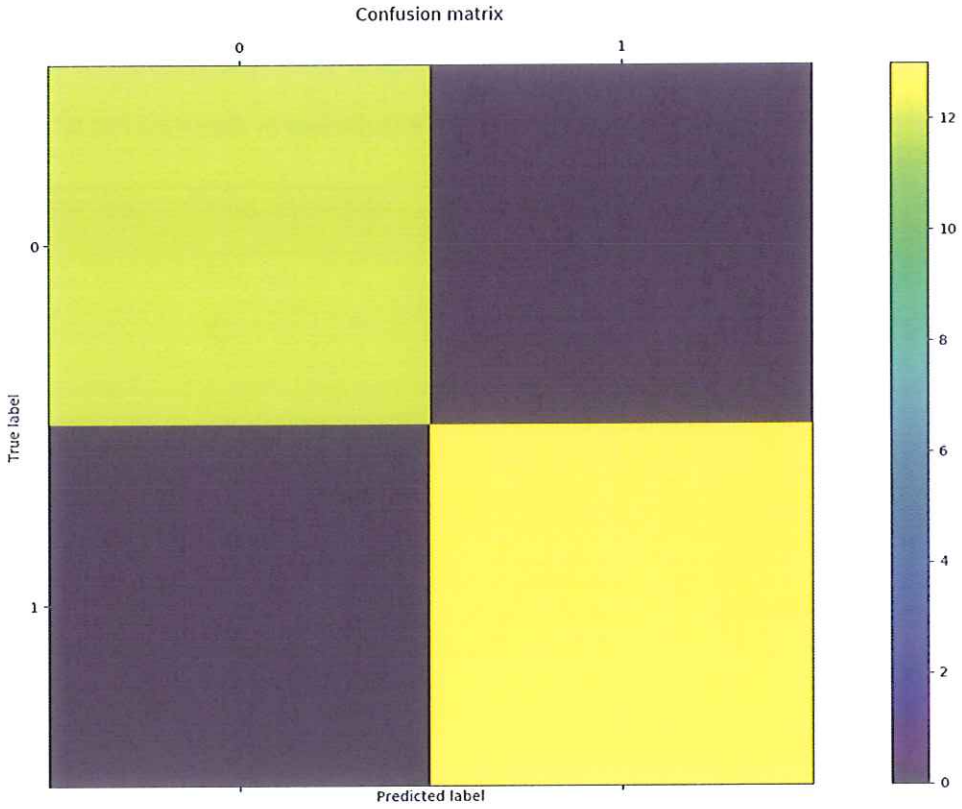

Figure 4.7: Second iteration LR classification

Table 4.5 shows that the functional classification of communication accuracy between the data packets in the MANET is 100 %. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 0 %, showing – that the normal and abnormal data packet are classified correctly. Sensitivity shows the proportion of identifying normal data packets that are correctly classified – and it is 100 %, this shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of normal and abnormal data packets that are correctly classified – which is 100 %, shows the percentage of normal and abnormal data packets that are correctly identified and classified. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets – and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets – is 100 %, indicating the difference between a measured normal and abnormal data packet.

The results of the second iteration are recorded in Table 4.5.

Table 4.5 Confusion matrix & cross validation evaluation of LR

Evaluation Parameter	Second iteration (%)
Accuracy (% of correct predictions)	100
True Positive	9
True Negative	16
False Positive	0
False Negative	0
Misclassification	0
Sensitivity	100
Specificity	100
False Positive Rate	0
Precision	100

c) Third iteration LR classification

Figure 4.8 is the confusion matrix and cross validation results obtained from the classification of data packets in the second iteration. The false negative is equal to the false positive – which is 0 % as shown in Table 4.6, meaning data packets travel safely in the network without degrading the network perfomance. This allows the data packets to move from source to destination thus letting services to work properly. The higher the classification accuracy of data packets the higher the data packets moving freely in the network without modification and being dropped. However, for communication and delivery of data packets in the network, the percentage of true positive – of 17 % and true negative – of 8 % as shown in Table 4.6, means that as the data packets are delivered, modification of data packets have been noticed and identified in the network – and, as a result, this can affect and degrade the network performance by propagate incorrect routing information and prevent services from working properly.

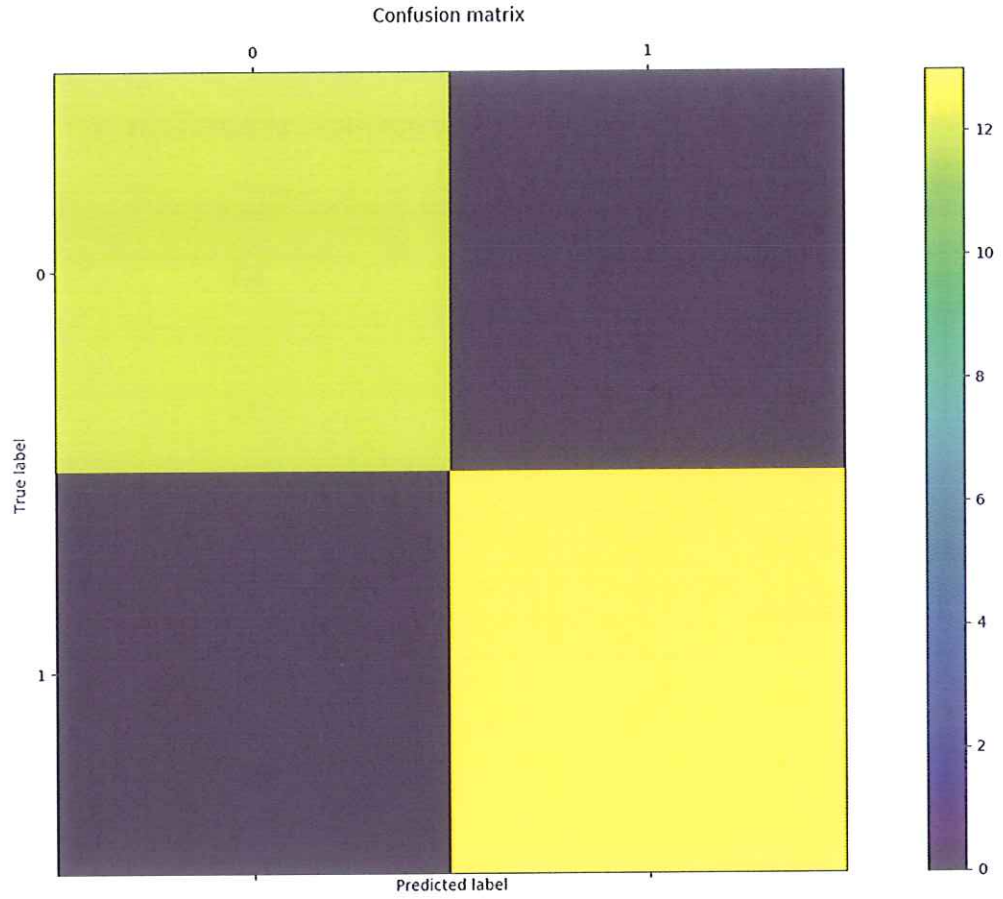

Figure 4.8: Third iteration LR classification

Table 4.6 shows that the functional classification of communication accuracy between the data packets in the MANET is 100 %. The quality of prediction between the data packets is good, that is when a data packet is actually predicted to be normal, – and how often this is done. The measurement error of wrongly assigning or misclassification of – the data packet is 0 %, showing – that the data packet is classified correctly. Sensitivity shows the proportion of identifying normal data packets that are correctly classified – and it is 100 %. This shows that the data packets are classified accordingly according to the formula TN / (TN + FP). Specificity is the proportion of abnormal data packets that are correctly identified – which is 100 %, shows the percentage of normal data packets that are correctly identified as abnormal data packets. False positive rate is the ratio of the abnormal data packets that have been wrongly categorized as normal data packets – and it is 0 %, which is what is expected. Precision which is the differentiation of random errors and accuracy between the normal and abnormal data packets and – it is 100 %, indicating the difference between a measured normal and abnormal data packet.

The results of the third iteration are recorded in Table 4.6.

Table 4.6 Confusion matrix & cross validation evaluation of LR

Evaluation Parameter	Third iteration (%)
Accuracy (% of correct predictions)	100
True Positive	17
True Negative	8
False Positive	0
False Negative	0
Misclassification	0
Sensitivity	100
Specificity	100
False Positive Rate	0
Precision	100

Table N captured the information collected randomly LR model while utilizing data packet. Table N shows information about the classification as either normal or abnormal data packet when delivered in the network. The confusion matrix captured in Figure 4.6, 4.7 and 4.8 respectively and generated in LR is shown in Table N.

Table N: Confusion matrix for first, second and third iteration LR

Evaluation Parameter	First iteration (%)	Second iteration (%)	Third iteration (%)
Accuracy: (% of correct prediction)	100	100	100
True Positive	12	9	17
True Negative	13	16	8
False Positive	0	0	0
False Negative	0	0	0
Misclassification	0	0	0
Sensitivity	100	100	100
Specificity	100	100	100
False Positive Rate	0	0	0
Precision	100	100	100

4.4 Model accuracy for SVM and LR

The analysis of the predictive models determines which of the two models has the greater predictive power. Both models render correct predictions, however they are not at all time correct and may result in some degree of incorrectness. The models provide a distinction between classification and regression analysis. For binary classification, they provide correct prediction such that the model predicts the same value as that of the dependent variable. Given a classification problem, the prediction based on binary outcome is good for representing a default status for detection of malicious MANET data. The accuracy of SVM and LR models will be tested and analyzed using the following criteria – [27, 100, 101];

1. Ability to classify data packet accurately.
2. Prediction / scoring the data packets well.
3. Expected future (new) performance of the models for data packet if the current performance is quite accurate.
4. Over fitting - that is the models necessarily adapt well to new or different circumstances.

5. The average performance reflects the performance of the model on the entire data packet.
6. Both models adapt well using confusion matrix for the values of true positive, true negative, false positive, and false negative.
7. Evaluation using cross validation for both models bring higher percentages of accuracy.

4.5 Discussions

In the above sections, we presented the results obtained from the SVM and LR models using the metrics PDER and PMMR. The main objective of the results was to identify which of the models used has high accuracy for communication and deliverance of data packets in the network. Based on the results obtained, it shows that SVM and LR classify data packets differently based on the classification and regression methods.

Table M captured the information of SVM model collected randomly while utilizing data packet based on confusion matrix for first, second and third iteration. Table M shows a decrease of SVM accuracy of normal and abnormal data packet from first iteration – 100 % to second iteration – 92 %, this shows that when true positive – 12 % and true negative 13 % the algorithm classifies the normal and abnormal data packet correctly. As compared to second iteration, when true positive – 9 %, and true negative – 16 %, the accuracy is – 92 %, this shows that normal and abnormal data packet is not correctly classified. For third iteration, when true positive – 14 % and true negative 11 %, the algorithm classifies the normal and abnormal data packet better than the second iteration in terms of accuracy. However, false positive – 0 % and false negative – 1 %, this shows that the algorithm has encountered errors between the normal and abnormal packet that are not correctly classified.

In comparison to the SVM model, is that the accuracy for normal and abnormal data packet for first iteration – 100 %, true positive – 12 % and true negative – 13 % as shown in Table N, this shows that the normal and abnormal data packet was correctly classified. Second iteration – 100%, true positive – 9 % and true negative – 16 % as compared to the second iteration, this shows the LR classifies data packet correctly when the true positive – 9 % and true negative – 16 %. The accuracy for LR algorithm remains constant when the true positive and true negative of normal and abnormal data packet increases or decreases. The rational is

that, when the percentage of normal data packet is higher, the proportion of abnormal data packet will be getting lower, and then it becomes very easy to recognize a normal and abnormal data packet. Accordingly, with the information presented in Table N the results of false positive, false negative and misclassification were all zero percentage which provides the model with the low false alarm rates. Furthermore, the SVM also showed inconsistency on the accuracy of precision, specificity, sensitivity, misclassification, and false positive rate as obtained from the confusion matrix.

In general, the SVM showed the predictive accuracy with an average rate of 92%, while LR algorithm has the predictive accuracy of 100%. This shows that the SVM algorithm had difficulty in detecting the data packets attack, as compared to the LR algorithm. Also it is clear that the LR algorithm showed the highest result of predictive accuracy when the normal and abnormal data packets classification is at 100 %. Furthermore, the LR also showed consistency with the average of 100% in terms of precision, specificity, sensitivity, misclassification, false positive rate as obtained from the confusion matrix as shown in Table N, which outperformed the SVM algorithm as shown in Table M. Our results indicate that the LR algorithm has the capability of detecting normal and abnormal data packets in MANETs and shows a good detection result of 100%. Thus, the LR model will be used in the design and development of MANET intrusion detection framework in this research.

4.6 MANET Node Intrusion Prediction and Identification Framework

This section discusses the proposed framework for predicting or identifying compromised nodes in MANETs using the LR prediction model. The model integrates packets within the MANETs network. The model uses the concepts of categorizing packets according to them being classified as abnormal and normal.

The choice of LR in the framework is informed by the fact that it has a higher predictive power as compared to SVM as demonstrated by the results. LR has consistency of producing good results in terms of precision, specificity, sensitivity, misclassification, false positive rate and confusion matrix and others.

4.7 The Intrusion Identification Framework

This section discusses the proposed framework for the prediction and identification of compromised nodes in MANET using the LR model which is integrated with packets in the MANETs network. The framework provides a regression method that is used to make predictions and probability based classification analysis [102].

The framework will provide the knowledge about the relationships and strength among the MANET data packets travelling in the network. In terms of predictions, the presence or absence of characteristics, LR predicts or classifies the MANETs data packets based on their category by learning from the trained and delivered packets data.

4.7.1 Framework Components

a) MANET Network

MANET network is set up as shown in Figure 4.9 having several MANET data packets. Each node of MANET receives and sends number of mobile packets when communicating with other packets. These packets dynamically establish paths among one another and these MANET packets are computed as either PDER or PMMR in the network. In order words, PDER and PMMR are the important metrics which are used in the computation of packet's delivery and modification rates in each node. Thus, these measures will be used to identify if an intrusion occurs in the network and which packets was compromised.

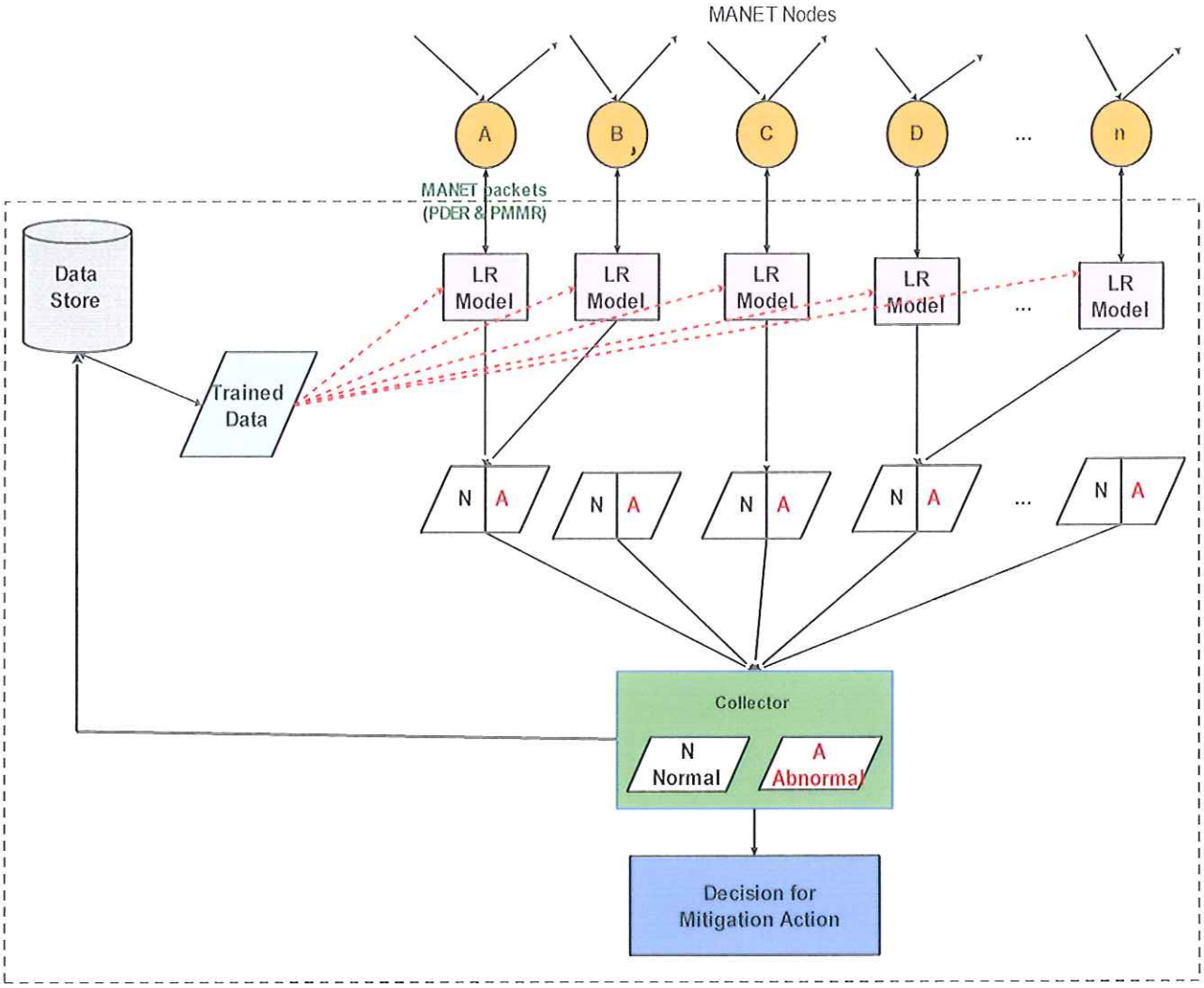

Figure 4.9 LR intrusion detection framework

b) LR model

Shown in Figure 4.9, to automatically identify in real-time if an intrusion occurred or not, is a predictive model using LR. This model is built using trained data which are MANETs historical data packets. Thus, the inputs to the model are the MANETs packet data computed as PDER and PMMR. By using the trained data, once packets are received in the nodes, the packets in the form of PDER and PMMR are used to fit the model. The model operates by learning the trained information to classify the packets according to their categories as either NORMAL or ABNORMAL. The operation of the LR model is shown in Figure 4.10.

c) Collector

This stage is the receiver of the output of the LR model. Once the MANETs packets are categorized into two classes: NORMAL or ABNORMAL, there are collected at this point. The classified data are then stored and reused when needed.

d) Data store

Prediction data or data that has been categorized are stored in the data store and are used as input for training the LR model in future.

e) Decision

Once the LR model output is received, decision can then be made as to which node was compromised. This is then followed by a mitigation action to identify the source of the attacks and to guard against further intrusion.

4.8 Framework Operations

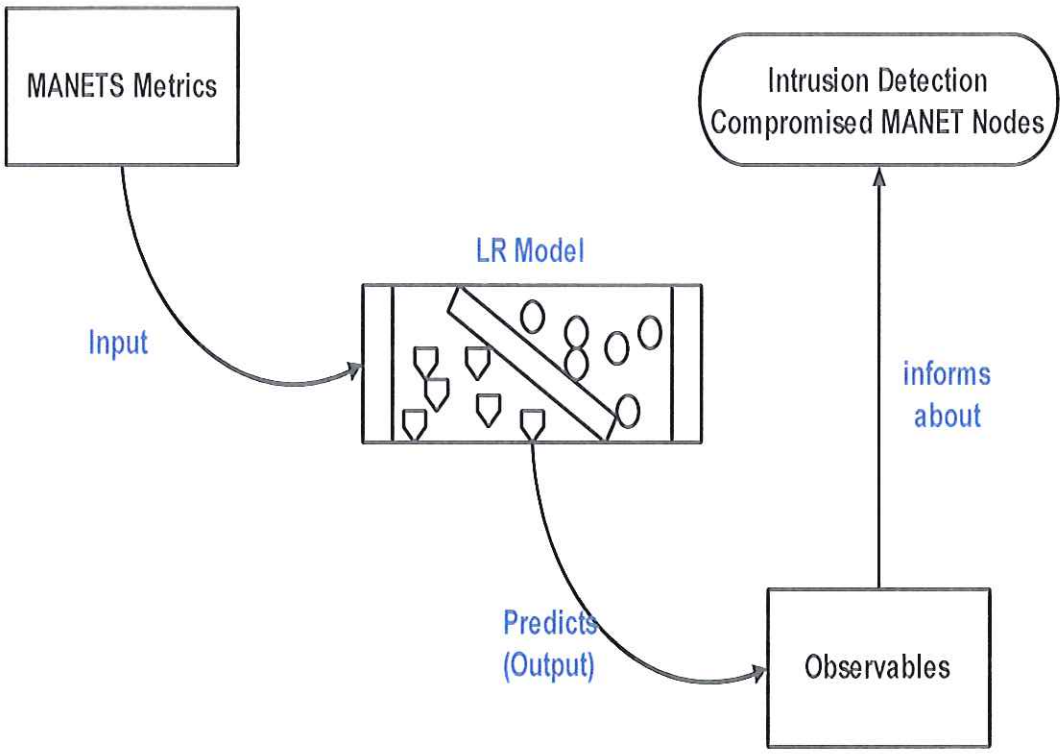

Figure. 4.10 LR model process

However, to construct the model using the iris flower dataset [91] collected in this research; Table X shows the descriptive statistics of the data, Table A and Table B shows the parameters for the model obtained using Statistical Package for the Social Sciences (SPSS) tool. Two possible models, LRM1 and LRM2 were constructed. In SPSS, we computed the binary LR parameter for LRM1 and LRM2 using a two-step process: (1) enter method and (2) perform the forward stepwise method with a cut-off value of 0.5 in 20 iterations. The statistics computed are the R-squared measures (R^2), the regression co-efficient (β), statistical significance value (ρ-value), the odds ratio (Exp(β)), negative 2 (-2) Log likelihood and the constant (α).

Table X. Descriptive statistics

Metric	N	Min	Max	Max	Std. Deviation
PDER	100	3.40	5.10	4.2825	0.35974
PMMR	100	0.60	3.35	1.8235	1.00311
Intrusion	100	0	1	0.50	0.503

Table A. LRM1 enter method

Metric	B	ρ-value	Exp(β)
PDER	-13.270	0.999	0.000
PMMR	31.253	0.995	3.739E+13
Constant (α)	2.982	1.000	19.727
-2 Log likelihood	0.000		
R^2	0.750		

Table B. LRM2 Forward Stepwise method

Metric	B	ρ-value	Exp(β)
PMMR	36.231	0.994	5.433E+15
Constant (α)	-58.173	.994	0.000
-2 Log likelihood	138.629		
R^2	0.750		

The classification results for the LRM1 and LRM2 are captured in Table C.

Table C. Classification Results

	Predicted Intrusion			
	Non-Intrusion		Intrusion	
LR Model	LRM1	LRM2	LRM1	LRM2
Non-Intrusion	50	50	0	0
Intrusion	0	0	50	50

Moreover, the predictors for LRM1 and LRM2 are shown in Table D.

Table D – Predictors

Model	PDER (β)	PMMR (β)	Constant (α)
LRM_1	-13.270	31.253	2.982
LRM_2	-	36.231	-58.173

Therefore, based on the Table D, our LR or predictive model for instruction detection is the formulated as follows:

For LRM1, the model is:

$$(1/MANET\_INTRUSION) = 2.982 + (-13.270)\ PDER + (31.2553)\ PMMR$$

For LRM2, the model is:

$$(1/MANET\_INTRUSION) = -58.173 + (36.231)\ PMMR$$

The choice of any of these models, LRM1 and LRM2 will depend on the R^2 and the log likelihood statistics. Prediction is done based on the following. At threshold or cut-off value of 0.5, packets are classified as:

NORMAL if $MANET\_INTRUSION \geq 0.5$

Otherwise:

ABNORMAL if $MANET\_INTRUSION < 0.5$.

The overall algorithm is shown on Figure. 4.11. As shown in Figure.4.11, "authentic node" means that the node is not compromised in any way in the network.

LR intrusion identification algorithm

```
Collect all the MANET data packets in the network computed
as PDER and PMMR
Based on trained data, fit the model with PDER and PMMR
   If (MANET_INTRUSION ≥ 0.5) then
           Node is authentic (NORMAL)

           else (MANET_INTRUSION < 0.5)

     Node is suspicious (ABNORMAL)
```

Fig. 4.11 LR model intrusion identification algorithm

4.9 Theoretical Framework Evaluation

For evaluation of the LR intrusion identification framework model explained in Figure 4.9, the parameters recorded in Table 4.7 were used taken from the literature such as proposed methods, algorithms, and the type of attacks, the behavioural metrics and others for detection of malicious and compromised nodes. The information gained from the literature reviews for the developed frameworks were used to validate the LR model for performance and identification of malicious attacks and compromised nodes in the MANET.

Table 4.7 LR Theoretical Framework Evaluation

	Author	Proposed Method	Algorithm	Attacks	Layers	Metrics	Protocol
1	Nikos et al [12]	Two phase detection based on zero knowledge techniques	Not based on symmetric or asymmetric encryption algorithms	Unauthorized nodes	Link and Network		
2	Patel et al [23]	Detection and identification of malicious nodes	Support vector machine	Packet dropping nodes	Network and Link	PDER, end to end delay, average throughput, normalized routing overhead and average energy consumption	AODV
3	Sukla et al [78]	finding chain of cooperating misbehaviours nodes Mechanisms	algorithm	Black and gray hole	Network		
4	Meenakshi et al [20]	Automatic security mechanism based in machine learning	Support vector machine	DoS, black and grey hole	Network	PDER, PMOR and PMISR	AODV

4.10 Benefits of the Framework

For the benefits of the LR intrusion identification framework, it provides security for MANET network data packets before any attack takes control and makes malicious node break the security of the network. Due to the rapid changes in the configuration of MANET

and the accompanying features, its services are faced by a number of challenges. Hence, for securing MANET network, the trade-off between the services is provided, as the results when one service guarantees without informing other services the security fails. And this security services is only depended on the applications of the network.

It also provides services to all available nodes in MANET whenever required. However, it provides with assurance of survival without consideration of denial of services (DoS) which can be launched at any layer in MANETs. As shown in this research, the framework helps with the identification of compromised node in MANETs, and potentially offers overall security benefits where the end users of the network can share information among the mobile devices. Hence the end user in turn sees the convenience benefits from having data packet transmission and other services available from any location of MANET network, and not accessing bogus information that propagates routing information and prevents services from operating properly. The last point is crucial to many users in that it provides with a large amount of performance and flexibility of mobile devices to operate in a clear secure environment applications like military applications, educations, commercial sector, persona area network and others.

The framework also provide the users with protection of information of mobile nodes in the MANET network from being exposed to malicious entities (confidentiality), so that the malicious adversary attack cannot easily eavesdrop the information being routed along the network. Hence the leakage on such information is notice before the malicious attack does damage to normal node and not lead to devastation consequences of the network.

Again, the framework also provides with assurance of communication between users of mobile nodes in the network. This will serve as to provide with assurance or confirmation of the origin of a communication. Hence prevent the attackers from compromising normal mobile nodes so that the resources and sensitive information in the network is not accessible by unauthorized members. Furthermore, it also restricts unwanted access from interfering with operation of other nodes in the MANET.

4.11 Chapter Summary

This chapter presents a discussion and interpretation of the experimental settings done in Chapter 3. The results achieved for both models, and described how the metrics and algorithms work together to perform classification and regression analysis using machine learning techniques in order to secure the MANET environment.

Chapter 5

Summary and Conclusion

5.1 Chapter overview

This chapter presents a brief summary of this research that was conducted and shows how the aim and objectives were achieved. It also discusses the recommendations and proposed direction for future work.

5.2 Conclusion

The summary of what was done in this study is presented in this section. In Chapter 1, we discussed the introduction of this research. We presented the problem statement, aim and objectives of study. In addition, we further outlined the research questions and the methods of investigation utilized. In Chapter 2, we performed an in-depth literature survey of MANETs in the perspective of intrusion detection using machine learning techniques, security and many other requirements into considerations. After carefully considering related and previous works, we were persuaded that the MANET technical issues or challenges can be fixed or solved, the most significant challenge was on MANET packets.

Chapter 3 proposed the method of identification and alleviation of compromised nodes in MANETs in order to thwart the malicious adversary attacks using SVM and LR by choosing two of the available machine learning tools for evaluation. The essence was to identify which of the models was more efficient in the detection of compromised nodes. Although several machine learning algorithms have been applied to automate the detection of meaningful pattern for intrusion detection in MANETs [20, 103, 104], in this study, machine learning algorithms of SVM and LR were used to categorize MANET generated data packets as either normal or abnormal in the network. The machine learning algorithms are far more resilient to the changes in MANETs such as those due to malicious nodes changing their patterns over time or rapid changes in environmental factors, as well as set up in the experiment to compare the performance of the algorithms.

In Chapter 4 we compare the results of our schemes and presented the results of the evaluation of the two chosen methods of identification and alleviation of compromised nodes in MANETs. The essence was to identify which of the models was more efficient in the

detection of compromised nodes and then develop a framework for automatic detection of attacks on MANETs. To achieve the outlined objectives, we designed a security model using the appropriate supervised machine learning tools to perform classification on existing historical data [91] to determine their accuracy, sensitivity, specificity, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision. In our case, the LR model was found to be effective in terms of high accuracy in the classification was employed in the development of a detection framework to automatically detect meaningful patterns (i.e. normal and abnormal data packets) in MANETs. The choice of supervised machine learning algorithms such as the SVM and LR is based on their widespread acceptance as the best approach of data modelling that provide a good generalization performance of learning from a model. They also play a significant role in cases where classification and analysis of datasets is to be performed. Thus, the chapter performed classification on the existing MANET data packets which were analysed and evaluated for the development of a detection framework for identification of normal and abnormal packets. It further explained how the LR algorithm outperformed the SVM algorithm in terms of accuracy, sensitivity, specificity, true positive, true negative, false positive, false negative, misclassification, sensitivity, specificity, false positive rate and precision to secure the performance of data packets in the network. It also explains how the LR algorithm was chosen for classification of data packets to ensure secure multihop data packets transmission against mobile attacks.

Finally, we presented a design of LR framework for identification of compromised nodes in MANETs in terms of analysis, design and operations. A theoretical evaluation was also performed to assess the effectiveness of the proposed LR model which we found to be effective, reliable and standardized. Based on the operation of LR model, we therefore conclude that introducing the LR algorithm to MANET intrusion detection system; will go a long way in making MANETs resilient to failure.

5.3 Recommendation and Future work

While it is important to understand how to minimize the malicious attacks in MANET environment, it is also crucial to consider security countermeasures for securing the network, and also to provide with practice of classification and regression analysis. However, to obtain a clear picture of identifying compromised nodes and securing the network against malicious

attacks, and to understand the potential role of security measures in providing secure packet deliveries, a more comprehensive analysis is required. To minimize the percentage of cooperative nodes being affected by malicious attacks, MANET environments should be provided with tighter security.

The dynamic nature of the network has been attractive to many different application areas such as military tactical networks, wireless sensor networks, and others. These applications have in turn introduced some design issues and challenges that need to be overcome. Future work could look at issues and challenges such as increasing bandwidth, finding lasting battery power and optimal computational power, as well as security issues.

Another future direction could be to implement the model designed in this dissertation in a real–world MANET in order to assess its effectiveness and performance. In addition, more MANET metrics could be utilized on more machine learning algorithms to identify best performing algorithms.

5.4 Appendix A: Support Vector Machine algorithm

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import  style
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn import metrics


X = np.array

([[4.3,0.8],[3.95,0.8],[3.95,0.75],[3.85,0.85],[4.3,0.8],[4.65,1.05],[4,0.85],[4.2,0.85],[3.65,0.8
],[4,0.8],[4.55,0.85],[4.1,0.9],[3.9,0.75],[3.65,0.6],[4.9,0.7],[5.05,0.95],[4.65,0.85],[4.3,0.85],
[4.75,1],[4.45,0.9],[4.4,0.95],[4.4,0.95],[4.1,0.6],[4.2,1.1],[4.1,1.05],[4,0.9],[4.2,1],[4.35,0.85
],[4.3,0.8],[3.95,0.9],[3.95,0.9],[4.4,0.95],[4.65,0.8],[4.85,0.8],[4,0.8],[4.1,0.7],[4.5,0.75],[4,0
.8],[3.7,0.75],[4.25,0.85],[4.25,0.8],[3.4,0.8],[3.8,0.75],[4.25,1.1],[4.45,1.15],[3.9,0.85],[4.45,
0.9],[3.9,0.8],[4.5,0.85],[4.15,0.8],[5.1,3.05],[4.8,3],[5,3.2],[3.9,2.65],[4.65,3.05],[4.25,2.9],[
4.8,3.15],[3.65,2.15],[4.75,2.95],[3.95,2.65],[3.5,2.25],[4.45,2.85],[4.1,2.5],[4.5,3.05],[4.25,2.
45],[4.9,2.9],[4.3,3],[4.25,2.25],[4.2,3],[4.05,2.5],[4.55,3.3],[4.45,2.65],[4.4,3.2],[4.45,2.95],[
4.65,2.8],[4.8,2.9],[4.8,3.1],[4.85,3.35],[4.45,3],[4.15,2.25],[3.95,2.45],[3.95,2.35],[4.25,2.55
],[4.35,3.35],[4.2,3],[4.7,3.05],[4.9,3.1],[4.3,2.85],[4.3,2.7],[4,2.65],[4.05,2.8],[4.55,3],[4.2,2.
6],[3.65,2.15],[4.15,2.75],[4.35,2.7],[4.3,2.75],[4.55,2.8],[3.8,2.05],[4.25,2.7]])

y = np.array

([0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1])

X_train, X_test, y_train, y_test = train_test_split(X, y)


clf = svm.SVC(kernel='linear')
y_pred = clf.fit(X_train, y_train).predict(X_test)

# Compute confusion matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

print(y_test)
```

```
# PRINT THE X_TRAINING AND X_TESTING VALUES

print('X_training values are : ', X_train)
print('X_testing values are : ', X_test)

# PRINT THE ACTUAL AND PREDICTED VALUES

print('The actual y_test true values are :', y_test)
print('The y_predicted values are :', y_pred)

# PRINT THE PERCENTAGE AND CORRECT PREDICTIONS

print('Classification accuracy: percentage of correct predictions = ',
metrics.accuracy_score(y_test, y_pred))


confusion = metrics.confusion_matrix(y_test, y_pred)

TP = confusion[1, 1]
TN = confusion[0, 0]
FP = confusion[0, 1]
FN = confusion[1, 0]

print('The value of True Positive: ', TP)
print('The value of True Negative: ', TN)
print('The value of False Positive: ', FP)
print('The value of False Negative: ', FN)

# THE OVERALL CLASSIFICATION ACCURACY
print('Classification Accuracy: Overall : (TP + TN)/ float(TP + TN + FP + FN)) = ',(TP +
TN)/ float(TP + TN + FP + FN))

# CLASSIFICATION ERROR: OVERALL, HOW OFTEN IS THE CLASSIFIER
INCORRECT?
# MISCLASSIFICATION RATE

print('Classification Error of misclassification: Overall : (FP + FN)/ float(TP + TN + FP +
FN)) = ',(FP + FN)/ float(TP + TN + FP + FN))

# SENSITIVITY: When the actual value is positive, how often is the prediction correct?
# How "sensitive" is the classifier to detecting positive instances?
# Also known as "True positive rate" or "Recall"

print('Sensitivity {TP / float(TP + FN)}  is :', TP / float(TP + FN))
```

 # SPECIFICITY!!
 # When the actual value is negative, how often is the prediction correct?

print('Specificity : {TN / float(TN + FP)} is :', TN / float(TN + FP))

 # WHEN THE ACTUAL VALUE IS NEGATIVE, HOW OFTEN IS THE PREDICTION INCORRECT?
print('False Positive Rate: {FP / float(TN + FP)} is :', FP / float(TN + FP))

 # PRECISION
 # When a positive value is predicted, how often is the prediction correct?

print('Precision : {TP / float(TP + FP)} is :', TP / float(TP + FP))

 # Show confusion matrix in a separate window
plt.matshow(cm)
plt.title('Confusion matrix')
plt.colorbar()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()

plt.scatter(X[:50, 0], X[:50,1], color = 'green', marker = 'o', label = 'Prediction')
plt.scatter(X[:50, 0], X[:50,1], color = 'red', marker = **'o'**, label = 'Class 0: Packet Modification and Misroute Rate(PMMR)')

plt.scatter(X[50:100, 0], X[50:100,1], color = 'blue', marker = 'o', label = 'Class 1: Packet Delivery (PDER)')

ho = plt.plot(xx, yy, color='k', label ="Optimal Separating Hyperplane")

 # plot the parallels to the separating hyperplane that pass through the
 # support vectors

b = clf.support_vectors_[1]
*yy_down = a * xx + (b[1] - a * b[0],)*
b = clf.support_vectors_[-1]
*yy_up = a * xx + (b[1] - a * b[0])*

```
print(b)


# plot the line, the points, and the nearest vectors to the plane


plt.plot(xx, yy, 'k--', color = 'k')
plt.plot(xx, yy_down, 'k', color = 'red')
plt.plot(xx, yy_up, 'k', color = 'red')

plt.xlabel('Packet Delivery (PDER)')
plt.ylabel('Packet Modification and Misroute Rate (PMMR)')
plt.title('Support Vector Classifier with linear kernel')
plt.legend()
plt.show()
```

5.5 Appendix B: Logistic Regression algorithm

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import  style
from sklearn.cross_validation import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn import model_selection


X = np.array
([[4.3,0.8],[3.95,0.8],[3.95,0.75],[3.85,0.85],[4.3,0.8],[4.65,1.05],[4,0.85],[4.2,0.85],[3.65,0.8
],[4,0.8],[4.55,0.85],[4.1,0.9],[3.9,0.75],[3.65,0.6],[4.9,0.7],[5.05,0.95],[4.65,0.85],[4.3,0.85],
[4.75,1],[4.45,0.9],[4.4,0.95],[4.4,0.95],[4.1,0.6],[4.2,1.1],[4.1,1.05],[4,0.9],[4.2,1],[4.35,0.85
],[4.3,0.8],[3.95,0.9],[3.95,0.9],[4.4,0.95],[4.65,0.8],[4.85,0.8],[4,0.8],[4.1,0.7],[4.5,0.75],[4,0
.8],[3.7,0.75],[4.25,0.85],[4.25,0.8],[3.4,0.8],[3.8,0.75],[4.25,1.1],[4.45,1.15],[3.9,0.85],[4.45,
0.9],[3.9,0.8],[4.5,0.85],[4.15,0.8],[5.1,3.05],[4.8,3],[5,3.2],[3.9,2.65],[4.65,3.05],[4.25,2.9],[
4.8,3.15],[3.65,2.15],[4.75,2.95],[3.95,2.65],[3.5,2.25],[4.45,2.85],[4.1,2.5],[4.5,3.05],[4.25,2.
45],[4.9,2.9],[4.3,3],[4.25,2.25],[4.2,3],[4.05,2.5],[4.55,3.3],[4.45,2.65],[4.4,3.2],[4.45,2.95],[
4.65,2.8],[4.8,2.9],[4.8,3.1],[4.85,3.35],[4.45,3],[4.15,2.25],[3.95,2.45],[3.95,2.35],[4.25,2.55
],[4.35,3.35],[4.2,3],[4.7,3.05],[4.9,3.1],[4.3,2.85],[4.3,2.7],[4,2.65],[4.05,2.8],[4.55,3],[4.2,2.
6],[3.65,2.15],[4.15,2.75],[4.35,2.7],[4.3,2.75],[4.55,2.8],[3.8,2.05],[4.25,2.7]])

y = np.array
([0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1])

#Split X into training and testing sets

X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)


# make class predictions for the testing set

y_pred = logreg.predict(X_test)
print('The classification accuracy:', metrics.accuracy_score(y_test, y_pred))
```

```
matrix = confusion_matrix(y_test, y_pred)
print(matrix)

# PRINT THE PERCENTAGE AND CORRECT PREDICTIONS

print('Classification accuracy: percentage of correct predictions = ',
metrics.accuracy_score(y_test, y_pred))
confusion = metrics.confusion_matrix(y_test, y_pred)

TP = confusion[1, 1]
TN = confusion[0, 0]
FP = confusion[0, 1]
FN = confusion[1, 0]

print('The value of True Positive : ', TP)
print('The value of True Negative : ', TN)
print('The value of False Positive: ', FP)
print('The value of False Negative: ', FN)

# THE OVERALL CLASSIFICATION ACCURACY
print('Classification Accuracy: Overall : (TP + TN)/ float(TP + TN + FP + FN)) = ',(TP +
TN)/ float(TP + TN + FP + FN))


# CLASSIFICATION ERROR: OVERALL, HOW OFTEN IS THE CLASSIFIER
INCORRECT?
# MISCLASSIFICATION RATE

print('Classification Error of misclassification: Overall : (FP + FN)/ float(TP + TN + FP +
FN)) = ',(FP + FN)/ float(TP + TN + FP + FN))

# SENSITIVITY: When the actual value is positive, how often is the prediction correct?
# How "sensitive" is the classifier to detecting positive instances?
# Also known as "True positive rate" or "Recall"

print('Sensitivity {TP / float(TP + FN)}  is :', TP / float(TP + FN))

# SPECIFICITY!!
# When the actual value is negative, how often is the prediction correct?

print('Specificity : {TN / float(TN + FP)}  is :', TN / float(TN + FP))

# WHEN THE ACTUAL VALUE IS NEGATIVE, HOW OFTEN IS THE PREDICTION
INCORRECT?
```

```
print('False Positive Rate:  {FP / float(TN + FP)}  is :', FP / float(TN + FP))

# PRECISION
# When a positive value is predicted, how often is the prediction correct?

print('Precision : {TP / float(TP + FP)}  is :', TP / float(TP + FP))

# Show confusion matrix in a separate window
plt.matshow(matrix)
plt.title('Confusion matrix')
plt.colorbar()
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()

report = classification_report(y_test, y_pred)
print(report)
```

Reference

[1] A. O. Bang and P. L. Ramteke, "Manet: history, challenges and applications," *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* vol. 2, pp. 249-251, 2013.

[2] J. Hoebeke, I. Moerman, B. Dhoedt, and P. Demeester, "An overview of mobile ad hoc networks: Applications and challenges," *Journal-Communications Network,* vol. 3, pp. 60-66, 2004.

[3] H. Yang, H. Luo, F. Ye, S. Lu, and L. Zhang, "Security in mobile ad hoc networks: challenges and solutions," *IEEE wireless communications,* vol. 11, pp. 38-47, 2004.

[4] B. Kaur, "Security Architecture for MANET and Its Application in M-Governance," in *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*, 2013, pp. 491-496.

[5] A. Mishra, K. Nadkarni, and A. Patcha, "Intrusion detection in wireless ad hoc networks," *IEEE wireless communications,* vol. 11, pp. 48-60, 2004.

[6] R. P. Kumar, A. Excellencia, and P. Kanimozhi, "Providing a New EAACK to Secure Data in MANET||," ed: IJREAT.

[7] D. Djenouri, L. Khelladi, and N. Badache, "A survey of security issues in mobile ad hoc networks," *IEEE communications surveys,* vol. 7, pp. 2-28, 2005.

[8] V. Karpijoki, "Security in ad hoc networks," in *Proceedings of the Helsinki University of Technology, Seminars on Network Security, Helsinki, Finland*, 2000.

[9] P. Yau and C. J. Mitchell, "Security vulnerabilities in ad hoc networks," in *The Seventh International Symposium on Communication Theory and Applications, July 13–18, 2003, Ambleside, Lake District, UK*, 2003, pp. 99-104.

[10] J.-S. Li and C.-T. Lee, "Improve routing trust with promiscuous listening routing security algorithm in mobile ad hoc networks," *Computer communications,* vol. 29, pp. 1121-1132, 2006.

[11] W. Zhang, R. Rao, G. Cao, and G. Kesidis, "Secure routing in ad hoc networks and a related intrusion detection problem," in *Military Communications Conference, 2003. MILCOM'03. 2003 IEEE*, 2003, pp. 735-740.

[12] N. Komninos, D. Vergados, and C. Douligeris, "Detecting unauthorized and compromised nodes in mobile ad hoc networks," *Ad Hoc Networks,* vol. 5, pp. 289-298, 2007.

[13] L. Raja and S. S. Baboo, "An overview of MANET: Applications, attacks and challenges," *Int. J. of Comp. Sci. Mobile Comput.(IJCSMC),* vol. 3, pp. 408-417, 2014.

[14] L. Zhou and Z. J. Haas, "Securing ad hoc networks," *IEEE network,* vol. 13, pp. 24-30, 1999.

[15] R. S. Singamsetty, "Detection of malicious nodes in mobile ad hoc networks," University of Toledo, 2011.

[16] S. B. S. G. Varaprasad, "Identification of Critical Node for the Efficient Performance in Manet."

[17] B. Sivakumar and G. Varaprasad, "Identification of critical node for the efficient performance in Manet," *Editorial Preface,* vol. 3, 2012.

[18] M. Kumar, A. Bhushan, and A. Kumar, "A study of wireless ad-hoc network attack and routing protocol attack," *International Journal of Advanced Research in Computer Science and Software Engineering ISSN,* vol. 2277, 2012.

[19] K. Madhusudhananagakumar and G. Aghila, "A survey on black hole attacks on aodv protocol in manet," *International Journal of Computer Applications (0975–8887) Volume,* pp. 23-30, 2011.

[20] M. Patel and S. Sharma, "Detection of malicious attack in manet a behavioral approach," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, 2013, pp. 388-393.

[21] H. Xiao, W. K. Seah, A. Lo, and K. C. Chua, "A flexible quality of service model for mobile ad-hoc networks," in *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, 2000, pp. 445-449.

[22] D. Johnson, Y.-c. Hu, and D. Maltz, "The dynamic source routing protocol (DSR) for mobile ad hoc networks for IPv4," 2070-1721, 2007.

[23] N. J. Patel and R. H. Jhaveri, "Detecting Packet Dropping Misbehaving Nodes using Support Vector Machine (SVM) in MANET," *International Journal of Computer Applications,* vol. 122, 2015.

[24] S. R. Gunn, "Support vector machines for classification and regression," *ISIS technical report,* vol. 14, pp. 85-86, 1998.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[26] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics,* vol. 35, pp. 352-359, 2002.

[27] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert systems with applications,* vol. 34, pp. 366-374, 2008.

[28] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology,* vol. 12, pp. 1-7, 2010.

[29] K. Papadopoulos, T. Zahariadis, N. Leligou, and S. Voliotis, "Sensor networks security issues in augmented home environment," in *Consumer Electronics, 2008. ISCE 2008. IEEE International Symposium on*, 2008, pp. 1-4.

[30] S. Boora, Y. Kumar, and B. Kochar, "A Survey on Security Issues in Mobile Ad-hoc Networks," *IJCSMS International Journal of Computer Science and Management Studies,* 2011.

[31] M. Kumar and R. Rishi, "Security aspects in mobile ad hoc network (MANETs): Technical review," *International Journal of Computer Applications IJCA,* vol. 12, pp. 24-28, 2010.

[32] S. Kaushik and M. Kaushik, "Analysis of MANET security, architecture and assessment," *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956),* vol. 1, pp. 787-793, 2012.

[33] P. Goyal, V. Parmar, and R. Rishi, "Manet: vulnerabilities, challenges, attacks, application," *IJCEM International Journal of Computational Engineering & Management,* vol. 11, pp. 32-37, 2011.

[34] M. Frodigh, P. Johansson, and P. Larsson, "Wireless ad hoc networking: the art of networking without a network," *Ericsson review,* vol. 4, p. 249, 2000.

[35] J. Loo, J. L. Mauri, and J. H. Ortiz, *Mobile ad hoc networks: current status and future trends*: CRC Press, 2016.

[36] S. Lalar, "Security in MANET: Vulnerabilities, Attacks & Solutions," *Intational J. Multidiscip. Curr. Res,* vol. 2, pp. 62-69, 2014.

[37] D. K. Pal and P. M. Goel, "Survey on security issues in mobile Ad Hoc networks," *IJCSIT,* vol. 5, pp. 3732-3735, 2014.

[38] B. S. CHANDRA, "ISSUES IN MOBILE ADHOC NETWORKS AND ITS AVAILABLE SECURITY METHODS."

[39] S. U. Agalawe and N. R. Chopde, "Security Issues: The Big Challenge in MANET," *International Journal of Computer Science and Mobile Computing,* vol. 3, pp. 417-424, 2014.

[40] J. Singh, A. Singh, and R. Shree, "An assessment of frequently adopted unsecure patterns in mobile ad hoc network: Requirement and security management perspective," *International Journal of Computer Applications (0975–8887),* vol. 24, 2011.

[41] P. M. Jawandhiya, M. M. Ghonge, M. Ali, and J. Deshpande, "A survey of mobile ad hoc network attacks," *International Journal of Engineering Science and Technology,* vol. 2, pp. 4063-4071, 2010.

[42] M. M. Alani, "MANET security: A survey," in *Control System, Computing and Engineering (ICCSCE), 2014 IEEE International Conference on*, 2014, pp. 559-564.

[43] D. Nguyen, L. Zhao, P.-o. Uisawang, and J. Platt, "Security Routing Analysis for Mobile Ad Hoc Networks," *Interdisciplinary Telecommunications Program of Colorado University, Spring,* 2000.

[44] N. Sharma and A. Sharma, "The black-hole node attack in MANET," in *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, 2012, pp. 546-550.

[45] K. Vishnu and A. J. Paul, "Detection and removal of cooperative black/gray hole attack in mobile ad hoc networks," *International Journal of Computer Applications,* vol. 1, pp. 38-42, 2010.

[46] R. H. Jhaveri, A. D. Patel, J. D. Parmar, and B. I. Shah, "MANET routing protocols and wormhole attack against AODV," *International Journal of Computer Science and Network Security,* vol. 10, pp. 12-18, 2010.

[47] M. Abolhasan, T. Wysocki, and E. Dutkiewicz, "A review of routing protocols for mobile ad hoc networks," *Ad hoc networks,* vol. 2, pp. 1-22, 2004.

[48] A. K. S. Ali and U. Kulkarni, "Characteristics, applications and challenges in mobile Ad-Hoc networks (MANET): overview," *Wireless Networks,* vol. 3, 2015.

[49] H. Deng, W. Li, and D. P. Agrawal, "Routing security in wireless ad hoc networks," *IEEE Communications magazine,* vol. 40, pp. 70-75, 2002.

[50] A. H. A. Rahman and Z. A. Zukarnain, "Performance comparison of AODV, DSDV and I-DSDV routing protocols in mobile ad hoc networks," *European Journal of Scientific Research,* vol. 31, pp. 566-576, 2009.

[51] K. Pandey and A. Swaroop, "A comprehensive performance analysis of proactive, reactive and hybrid MANETs routing protocols," *arXiv preprint arXiv:1112.5703,* 2011.

[52] M. Kaur and A. Nayyar, "A comprehensive review of mobile adhoc networks (MANETS)," *International journal of emerging trends & technology in computer science (IJETTCS),* vol. 2, pp. 196-210, 2013.

[53] T. O. Ayodele, "Introduction to machine learning," in *New Advances in Machine Learning*, ed: InTech, 2010.

[54] H. Mannila, "Data mining: machine learning, statistics, and databases," in *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on*, 1996, pp. 2-9.

[55] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Communications Surveys & Tutorials,* vol. 16, pp. 1996-2018, 2014.

[56] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160, pp. 3-24, 2007.

[57] K.-R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine learning techniques for brain-computer interfaces," 2004.

[58] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*: Springer Science & Business Media, 2013.

[59] N. J. Patel and R. H. Jhaveri, "Detecting packet dropping nodes using machine learning techniques in Mobile ad-hoc network: A survey," in *Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on*, 2015, pp. 468-472.

[60] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials,* vol. 10, pp. 56-76, 2008.

[61] D. Janakiram, V. Reddy, and A. P. Kumar, "Outlier detection in wireless sensor networks using Bayesian belief networks," in *Communication System Software and Middleware, 2006. Comsware 2006. First International Conference on*, 2006, pp. 1-6.

[62] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 32, pp. 2106-2112, 2010.
[63] wikipedia. (2018). *disadvantages-linear-regression* [Online]. Available: https://sciencing.com
[64] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research,* vol. 13, pp. 1063-1095, 2012.
[65] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.
[66] F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.
[67] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters,* vol. 31, pp. 651-666, 2010.
[68] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB Conference*, 1994, pp. 487-499.
[69] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks,* vol. 12, pp. 783-789, 1999.
[70] K. O. Elish and M. O. Elish, "Predicting defect-prone software modules using support vector machines," *Journal of Systems and Software,* vol. 81, pp. 649-660, 2008.
[71] S. R. Gunn, "Support vector machines for classification and regression," *ISIS technical report,* vol. 14, pp. 5-16, 1998.
[72] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, pp. 199-222, 2004.
[73] Wikipedia. (2018). *Support_vector_machine* [Online]. Available: https://en.wikipedia.org
[74] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research,* vol. 96, pp. 3-14, 2002.
[75] F. Harrell, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*: Springer, 2015.
[76] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *Journal of the American Statistical Association,* vol. 73, pp. 699-705, 1978.
[77] Wikipedia. (2018). *Logistic_regression* [Online]. Available: https://en.wikipedia.org
[78] S. Banerjee, "Detection/removal of cooperative black and gray hole attack in mobile ad-hoc networks," in *proceedings of the world congress on engineering and computer science*, 2008.
[79] S. Mehfuz and M. Doja, "Swarm intelligent power-aware detection of unauthorized and compromised nodes in MANETs," *Journal of Artificial Evolution and Applications,* vol. 2008, 2008.
[80] R. H. Jhaveri, S. J. Patel, and D. C. Jinwala, "A novel approach for grayhole and blackhole attacks in mobile ad hoc networks," in *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, 2012, pp. 556-560.
[81] D. M. Shila and T. Anjali, "Defending selective forwarding attacks in WMNs," in *Electro/Information Technology, 2008. EIT 2008. IEEE International Conference on*, 2008, pp. 96-101.
[82] G. Xiaopeng and C. Wei, "A novel gray hole attack detection scheme for mobile ad-hoc networks," in *Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference on*, 2007, pp. 209-214.
[83] M. Sengar, P. P. Singh, and S. Shiwani, "Detection of Black Hole Attack In MANET Using FBC Technique," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 2, pp. 269-272, 2013.
[84] L. Tamilselvan and V. Sankaranarayanan, "Prevention of blackhole attack in MANET," in *Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007. The 2nd International Conference on*, 2007, pp. 21-21.

[85] A. Sardana, T. Bedwal, A. Saini, and R. Tayal, "Black hole attack's effect mobile ad-hoc networks (MANET)," in *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, 2015, pp. 966-970.

[86] R. Akbani, T. Korkmaz, and G. Raju, "A machine learning based reputation system for defending against malicious node behavior," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, 2008, pp. 1-5.

[87] F. Barani and S. Gerami, "ManetSVM: Dynamic anomaly detection using one-class support vector machine in MANETs," in *Information Security and Cryptology (ISCISC), 2013 10th International ISC Conference on*, 2013, pp. 1-6.

[88] wikipedia. (2017). *Qualitative_research* [Online]. Available: https://en.wikipedia.org

[89] wikipedia. (2017). *Quantitative_research* [Online]. Available: https://en.wikipedia.org

[90] Wikipedia. (2018). *common-machine-learning-algorithms* [Online]. Available: www.analyticsvidhya.com

[91] Wikipedia. (2017). *Iris flower data set* [Online]. Available: https://en.wikipedia.org

[92] W.-H. Chen, S.-H. Hsu, and H.-P. Shen, "Application of SVM and ANN for intrusion detection," *Computers & Operations Research,* vol. 32, pp. 2617-2634, 2005.

[93] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing,* vol. 52, pp. 2348-2355, 2004.

[94] S. Boyd and L. Vandenberghe, *Convex optimization*: Cambridge university press, 2004.

[95] E. A. Zanaty and S. Aljahdali, "Improving the Accuracy of Support Vector Machines," in *Computers and Their Applications*, 2008, pp. 196-202.

[96] D. Pregibon, "Logistic regression diagnostics," *The Annals of Statistics,* pp. 705-724, 1981.

[97] R. A. Giancristofaro and L. Salmaso, "Model performance analysis and model validation in logistic regression," *Statistica,* vol. 63, pp. 375-396, 2007.

[98] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, pp. 861-874, 2006.

[99] P. Zhang, "Model selection via multifold cross validation," *The Annals of Statistics,* pp. 299-313, 1993.

[100] M. Behzad, K. Asghari, M. Eazi, and M. Palhang, "Generalization performance of support vector machines and neural networks in runoff modeling," *Expert Systems with applications,* vol. 36, pp. 7624-7629, 2009.

[101] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural networks,* vol. 17, pp. 113-126, 2004.

[102] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics,* vol. 35, pp. 352-359, 2002.

[103] H. Deng, Q.-A. Zeng, and D. P. Agrawal, "SVM-based intrusion detection system for wireless ad hoc networks," in *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, 2003, pp. 2147-2151.

[104] A. Este, F. Gringoli, and L. Salgarelli, "Support vector machines for TCP traffic classification," *Computer Networks,* vol. 53, pp. 2476-2490, 2009.