AI-BASED INTRUSION DETECTION SYSTEM IN CLOUD COMPUTING

A Thesis

Submitted

In Partial Fulfilment of the Requirements

For the Degree of

MASTER OF TECHNOLOGY

In

ADVANCED COMPUTING AND DATA SCIENCE

Submitted by

Srijan Mishra

(2001209003)

Under the Supervision of

Dr. Shashank Singh

(Assistant Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

JULY, 2022

INTEGRAL UNIVERSITY
इंटीग्रल विश्वविद्यालय
Accredited by NAAC. Approved by the University Grants Commission under Sections 2(f) and 12B of the UGC Act, 1956, MCI, PCI, IAP, BCI, INC, CoA, NCTE, DEB & UPSMF. Member of AIU. Recognized as a Scientific & Industrial Research Organization (SIRO) by the Dept. of Scientific and Industrial Research, Ministry of Science & Technology, Government of India.

CERTIFICATE

This is to certify that Mr. /Mrs. **Srijan Mishra** (Roll. No. 2001209003) has carried out the research work presented in the thesis titled "**AI-Bases Intrusion Detection System in Cloud Computing**" submitted for partial fulfillment for the award of the Degree of **Master of Technology in Computer Science & Engineering** from **Integral University**, **Lucknow** under my supervision.

It is also certified that:

(i) This thesis embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The thesis fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

(iv) No published work (figure, data, table etc) has been reproduced in the thesis without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

Dr. Shashank Singh
Dissertation Guide
(Assistant Professor)
Department of CSE,
Integral University, Lucknow
Date:
Place: Lucknow

DECLARATION

I hereby declare that the thesis titled "**AI-Based Intrusion Detection System in Cloud Computing**" submitted to the Computer Science and Engineering Department, Integral University, Lucknow in partial fulfillment of the requirements for the award of the Master of Technology degree is an authentic record of the research work carried out by me under the supervision of **Dr. Shashank Singh**, Department of Computer Science & Engineering, Integral University, Lucknow. No part of this thesis has been presented elsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the thesis. I further certify that I have not will fully take other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites without their permission, and have not included those in this MTech. thesis citing as my own work. In case, this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

Date:

Srijan Mishra

Signature

Srijan Mishra

Enrol. No- **2000102246**

RECOMMENDATION

On the basis of the declaration submitted by “**Srijan Mishra**”, a student of M.Tech CSE (Advanced Computing and Data Science), successful completion of Pre presentation on 24/06/2022, and the certificate issued by the supervisor, “**Dr. Shashank Singh**”, Assistant Professor, Computer Science, and Engineering Department, Integral University, the work entitled “AI-Based Intrusion Detection System in Cloud Computing”, submitted to the department of CSE, in partial fulfillment of the requirement for the award of the degree of Master of Technology Advanced Computing and Data Science, is recommended for examination.

Program Coordinator Signature	HOD Signature
Dr. Faiyaz Ahmad	Mrs. Kavita Agarwal
Dept. of Computer Science &Engineering	Dept. of Computer Science & Engineering
Date:	Date:

COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: “**AI-Based Intrusion Detection System in Cloud Computing”**

Candidate Name: **Srijan Mishra**

The undersigned hereby assigns to Integral University all rights under copyright thatmay exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award of the Master of Technology Advanced Computing and Data Science degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University’s copyright notices are indicated.

SRIJAN MISHRA

ACKNOWLEDGEMENT

I am highly grateful to the Head of the Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.
It gives me great pleasure to express my deep sense of gratitude and indebtedness to my guide **Dr. Shashank Singh, Assistant Professor, Department of Computer Science and Engineering,** for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me with this opportunity to carry out the ideas and work during my project period and for helping me to gain the successful completion of my Project.
I am also highly obliged to the Head of Department, **Mrs. Kavita Aggarwal (Head of Department of Computer Science and Engineering),** and PG Program Coordinator **Dr. Faiyaz Ahamad, Assistant Professor, Department of Computer Science and Engineering,** for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly to work hard in this project. I pay my respect and love to my parents and all other familymembers and friends for their help and encouragement throughout this course of project work.

Date:
Place:

Table of Content

CONTENT **PAGE NO.**

List of Figures

List of Tables

ABSTRACT

The intrusion detection system, often known as an IDS, is an essential component in the process of keeping a network secure. In addition, because the cloud platform is rapidly advancing and becoming more use in our day-to-day lives, it is both helpful and vital to developing an efficient IDS for the cloud. However, when applied on a cloud platform, conventional intrusion detection methods are likely to confront a number of obstacles. Due to the additional detection overhead, a portion of the cloud may become overloaded by the pre-determined IDS design. As a distributed system with an adaptive design, this thesis offers a neural network-based IDS that makes maximum use of available resources without overtaxing any one cloud computer. To further enhance its capacity to identify new threats, the suggested IDS employs neural network machine learning. On a physical cloud testbed, the suggested IDS was tested with the KDD dataset and shown to be a promising way to detecting cloud infrastructure threats.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The ever-increasing demand to utilize, connect with, and use other features that are linked to the Internet led to the development of cloud computing. Cloud computing frequently implies the provision of dynamically expanded Internet service by means of the utilization of resources that have been virtualized. The phrase "the cloud" is frequently used as a metaphor for referring to many computer networks as well as the internet. In the past, representations of clouds were widely used whenever there was a need to convey the concept of telecommunication networks. The term "cloud imagery," on the other hand, is being used increasingly frequently these days to describe to an abstraction of the Internet and the underlying infrastructure (Shiravi et al., 2012). The phrase "cloud computing" most commonly refers to a paradigm of information technology (IT) infrastructure that is based on renting and using resources rather than purchasing them. It indicates that the necessary resources are received through the network in compliance with principles such as being simple to grow and being available whenever they are needed; The model of computing that is referred to as "rent and use" is what is meant when we talk about "generalized cloud computing." This category of services may refer to those that are associated with information technology and software as well as the Internet, but it may also refer to other kinds of services. It suggests that the ability to compute might be considered a sort of commodity and traded via the internet in the same manner as other types of utility, such as water, gas, electricity, and so on. Cloud computing has the following main features:

1. The distribution of resources in a dynamic fashion. Cloud computing has the ability to dynamically divide and release various physical and virtual resources in response to the real-time demands of customers. When a request is made, the cloud will quickly fulfil it by increasing the number of resources that are now accessible in order to achieve elasticity of resources. If the user decides that

they do not require this particular portion of the resources anymore, they are free to release it. Therefore, cloud computing is seen as the combination of an endless number of resources, which enables the scalability of information technology resources.

2. Self-service options available on demand Users are able to obtain resources automatically through the usage of cloud computing's self-service mode, which eliminates the need for them to communicate with service providers. In addition, the cloud computing system gives users access to a directory of available application services, from which they can choose the one that best suits their requirements.
3. The ease with which one can enter the network Users have the ability to connect to the network through a variety of terminal equipment, which enables the network to be accessed from any location.
4. A service that can be measured. In cloud computing, the allocation of resources can be maximized, and the resources themselves can be automatically regulated depending on the sort of service being provided. It is similar to a pay-as-you-go model of providing services.
5. Virtualization. It is possible, thanks to the technology of virtualization, to reorganize computing resources that are located in various locations in order to realize the concept of shared infrastructure.

Computing in the cloud is rapidly expanding in today's world, and an increasing number of people are becoming aware of it as a result of the numerous advantages it provides, which include high scalability, high flexibility, and low operational cost. Users of cloud services are often exempt from the requirement that they be familiar with the inner workings of the cloud-based software or platform they make use of (Shiravi et al., 2012). They are just required to send their requests to the cloud provider, and then wait for the results, which is a substantially simpler and more effective way to acquire access to the required computing resources. Instead, they are required to send their requests to the cloud provider.

On the other hand, the cloud systems that are now accessible suffer from a few drawbacks (Brown et al., 2009). According to the research, the majority of cloud users are most concerned about potential security concerns, such as the unauthorized access, erroneous data, and information leaking. Other issues, including user friendliness, support systems, and stable operations, have not been given as much consideration as they could have been. The implementation of a distributed intrusion detection system (IDS) on the cloud as a means of shielding virtual machines (VMs) and virtual networks from potential threats appears to be the most logical solution for the problem of insufficient security that is present in cloud computing. This would solve the problem of cloud computing being insecure (Axelsson, 1998). When monitoring the functioning of a system, it is common practice to make use of a piece of software known as an intrusion detection system. This is done with the intention of preventing behaviors that were not expected and delivering a report to the manager. Real-time monitoring is a type of system that is utilized for network transmissions, and one such system is known as an intrusion detection system. It would either sound an alarm or take the initiative to react in order to defend the network from being attacked when it found acts that were suspicious. This was done in order to protect the network from being breached. When compared with other kinds of network security systems, the intrusion detection system provides a method that is more effective for ensuring the safety of the entire network. This is the most important distinction that can be made between the two. In addition, the construction of an IDS can often be accomplished using one of two major strategies. The first technique is designed to identify unusual occurrences. Searching for unexpected behaviors or data that do not conform to an existing model is an example of what is meant by the term "anomaly detection." These actions or data that do not comply with the rules are not acceptable. of the model are usually referred to as anomalies, abnormal values, the disharmony of observation, exceptions, aberrations, surprise, quirks, or contaminants in a number of different application areas. Other names for these things include abnormalities, abnormal values, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, the disharmony of observation, In this method, the intrusion detection system, or IDS, does not have any established normal activities set into it; rather, the IDS will be designed to learn what kinds of actions are malicious and what kinds of actions are normal based on a well-planned training programmed that utilizes a substantial amount of data. The IDS will be designed to learn this information based on the fact that it will be designed to learn what kinds of actions are normal based on the fact that it will be designed to learn what kinds of actions are. The IDS will be designed to learn what kinds of actions are malicious and what kinds of actions are normal based on this well-planned training programme. The IDS will be able to better identify possible threats as a result of this. Any form of technology that detects anomalies requires a specific kind of data to be entered, and this particular kind of data is a crucial component. These actions or data that do not comply with the rules are not acceptable.

The general input data is always a collection of the data instance, which is also known as the object, record, point, vector, pattern, case, sample, observation, and entity (Caswell & Beale, 2004). It is possible for this method to cause a lot of inaccurate judgments, such as sounding an alarm when the network is functioning normally or ignoring an attack because it considers the attack to be a normal action. Both of these scenarios are examples of how this method could lead to inaccurate judgments. This approach has the benefit of being able to test out new kinds of assaults, which is a distinct advantage. On the other side, it could lead to a significant number of incorrect conclusions. The second approach to constructing an intrusion detection system (also known as an IDS) is called signature-based detection, and it is dependent on a knowledge base. These actions or data that do not comply with the rules are not acceptable.

The signature-based detection method is highly useful due to the fact that it is very effective at detecting known threats by utilising signatures of observed events in order to determine probable attacks. This makes the signature-based detection method a particularly important technique. Because of this, the method of detection that relies on signatures is extremely helpful (Carpenter et al., 1992). The general input data is always a collection of the data instance, which is also known as the object, record, point, vector, pattern, case, sample, observation. It is possible for this method to cause a lot of inaccurate judgments, such as sounding an alarm when the network is functioning normally or ignoring an attack because it considers the attack to be a normal action. Both of these scenarios are examples of how this method could lead to inaccurate judgments. This approach has the benefit of being able to test out new kinds of assaults, which is a distinct advantage. On the other side, it could lead to a significant number of incorrect conclusions. The second approach to constructing an intrusion detection system (also known as an IDS) is called signature-based detection, and it is dependent on a knowledge base. These actions or data that do not comply with the rules are not acceptable. The signature-based detection method is highly useful due to the fact that it is very effective at detecting known threats by utilizing signatures of observed events in order to determine probable attacks. This makes the signature-based detection method a particularly important technique. Because of this, the method of detection that relies on signatures is extremely helpful (Carpenter et al., 1992). In addition to this, it is crucial for the intrusion detection system (IDS) to be able to identify attacks in the cloud that had not been seen or known about before. As a consequence of this, anomaly detection will be favoured, despite the fact that it may place a greater demand on the resources that are now accessible (Chebrolu et al., 2005). Therefore, in order to satisfy cloud clients and provide a suitable level of performance for intrusion detection at the same time, a balance 6 needs to be achieved as soon as possible.

1.2 Existing IDS architecture and algorithms

A hybrid intrusion detection system that combines K-Means. They use K-nearest neighbor and Naive Bayes as the two key factors for anomaly detection. An entropy-based algorithm is used to select the important attributes and removes the redundant attributes, a misuse intrusion detection system is founded by a genetic algorithm that based on the knowledge of a set of intrusion behavior classification rules (Chittur, 2001). An adaptive network intrusion detection system which uses a two-stage architecture. In the first stage, some possible malicious connections in the traffic are detected by a probabilistic classifier. In the second stage, the authors try to minimize the possible IP addresses of attacks through an HMM based traffic model. In Ref. (Chittur, 2001), the researchers propose a distributed IDS by using multi-agent methodology which is combined with accurate data mining techniques. Those intelligent agents are responsible for collecting and analyzing the network connections, and the performance is really good. The authors in Ref. (Crosbie & Spafford, 1995) use evolution theory to explain the evolution of data and connections in the network and thus reduce the complexity. The proposed Intrusion Detection System (IDS) is based on the theory discusses an IDS with a new alert clustering and analyzing facility. This mechanism could help all cooperating nodes get a better understanding of whole system which helps them to find false alarms and detect those damaged nodes in the system. Attacks in one node will spread to the network to alert other cooperating nodes to update themselves about new attack patterns. This will lead to early detection and prevention of attacks. In Ref. (Daumé, 2012), the authors choose 19 key features to describe all the various network visits. Then they use a gradual feature removal method and combine it with a clustering method, ant colony algorithm and support vector machine (SVM) to build an intrusion detection system to determine whether a visit in the network is normal or not. It is shown that a high-throughput intrusion detection system (IDS) is represented. This IDS is based on a comparison architecture. It includes a bloom filter-based header comparison and parallel pattern matching method which means it can parallel sequence compare packet content with the Snort rules. It uses a Particle Swarm Optimization (PSO) as a feature selection algorithm and a decision tree as a classifier. This would help accelerate the speed of detection and make the result more accurate. In Ref. (Chittur, 2001), authors claim that the Hidden Naïve Bayes (HNB) is a data mining model that relaxes the Naïve Bayes method's conditional independence assumption which could help to solve intrusion detection problems as it has attributes such as dimensionality, highly correlated features and large stream volumes.

1.3 Related works

Using machine algorithms in a big data environment, Saud Mohammed Othman and Fadl Mutaher Ba-Kiwi (Shiravi et al., 2012) presented their work on an intrusion detection model. According to this study, Big data ever-increasing volume has shifted the significance placed on data security and analytic technologies. An IDS monitors and analyses data in order to discover any system or network intrusions that may have occurred. Traditional strategies for detecting network assaults have become increasingly complex due to the volume, diversity, and speed of data created in the network. IDS uses Big Data approaches to analyze Big Data in an accurate and effective manner.By utilizing the Spark-Chi-SVM architecture, they were able to construct an intrusion detection model that is capable of managing enormous amounts of data. The Spark Big Data platform was utilized in the recommended approach in order to facilitate the handling and analysis of data in a timely manner. The large dimensionality of big data adds to the difficulty and length of the categorization process. A Survey of Intrusion Detection Systems utilizing Machine Learning Techniques was provided by Sharmila Wagh and Vinod K. Pachghare (Axelsson, 1998). Computers and network-based technologies are becoming more and more commonplace in today&#39; s environment, according to the authors. The importance of network security cannot be overstated in computer age. Detection of system assaults and classification of system activity into normal and abnormal forms are the goals of an Intrusion Detection System (IDS). IDS systems that use machine learning to identify intrusions have become increasingly commonplace.

Cloud-based distributed machine learning-based intrusion detection system. Edge network components from Cloud providers are to accompany the proposed system in the Cloud. Incoming network communication may be intercepted by the edge network routers on the physical layer as a result. Before being transmitted to a module that employs the NaiveBayesclassifierr to identify anomalies, the network data collected by each Cloud router is preprocessed using a time-based sliding window technique. When there is a buildup of network congestion, server nodes that are powered by Hadoop and MapReduce are made accessible to each anomaly detection module. A server for synchronizing anomalous network traffic data is assigned to each time frame. This server collects anomalous network traffic data from each router in the system. Following this, Random Forest classifiers are applied to each attack in order to establish the kind of assault that was committed.

Hassan Musafer and Ali Alessa; s study, titled & quot; Machine Learning-Based Network Intrusion Detection Detection: Dimensionality Reduction Approaches, & quot; was recently published (Caswell & Beale, 2004). They claim that all parties, including consumers, businesses, and governments, are worried about the safety of computer networks. As it becomes increasingly difficult to safeguard networked systems against assaults, the strategies that attackers use to carry out such assaults also evolve. A portion of the solution is in making the existing intrusion detection systems more effective. The use of machine learning to create intrusion detection systems is an approach that is becoming more popular because of its efficiency (IDS). When improvements are made to IDS qualities like as discrimination and representation, there is a considerable increase in the system overall performance. Through the use of Deep Learning Auto-Encoder (AE) and Principal Component. Using Principal Components Analysis (PCA), the dimensionality of the features was reduced for the purpose of this inquiry (PCA). It is possible that the combination of these two approaches will result in the acquisition of low-dimensional attributes that can be utilised in the building of classifiers such as Random Forest (RF), Bayesian Network, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).

1.4 Objective of the project

In this paper, a distributed IDS architecture is proposed. This design is based on the difficulties that were outlined before and consists of nodes running backpropagation (BP) based ANNs on the cloud platform. It is anticipated that, as a result of its architecture, it will gain more flexibility, scalability, and performance. The proposed IDS system has two main characteristics:

1) It has a flexible distributed architecture that is capable of adjusting its configuration depending on real-time information regarding resource usage in order to prevent any node in the cloud from becoming overloaded.

2) It offers results on various dimensions, which can be utilised not only to spot malicious activity but also to discover what kinds of malicious activities are going place. After the IDS has been constructed, it will be put through a variety of tests to evaluate its performance. The results of these

tests will indicate whether or not this IDS is adequate and how it can be improved in the near and distant futures.

Chapter- 2

LITERATURE REVIEW

2.1 Outline

There are three sections to this chapter. Understanding neural networks and backpropagation algorithms, which are key concepts in this investigation, may be found in the first two sections. Among neural network-based algorithms, the backpropagation method is one of the most widely used that incorporates feedback. To alter the backpropagation algorithm in the cloud-based IDS, see the third section.

2.2 Artificial intelligence

The process of reproducing human cognition in machines via the use of computer programming that enables the machines to think and behave in the same manner as humans is referred to as "artificial intelligence" (AI), which is an abbreviation for the word artificial intelligence. The expression can also be used to refer to any kind of machine that displays characteristics that are generally associated with the human mind, such as the capability to learn and find solutions to problems.

One of the most desirable characteristics of artificial intelligence would be the ability to think and act in a way that maximises the likelihood of achieving a specific goal as a result of the decisions they make over what courses of action to pursue. Machine learning is the process by which computer programmes may automatically learn from and adapt to new data without the intervention of humans. This notion is known as "machine learning." The term "artificial intelligence" (AI) is used to describe the notion, and this is a subset of it. Deep learning techniques, which require the consumption of large amounts of unstructured data such as text, images, or video, are what make this form of autonomous learning possible. Examples of this type of data include text, photos, and video.

2.3 History of artificial intelligence

The following is a condensed version of a timeline that details the development of AI from its creation during the past sixty years. The term "artificial intelligence" was first used in 1956 by John McCarthy, who also organized the first conference on the subject.1969 saw the construction of Shakey, the world's first mobile robot with a general function. It is now able to perform tasks with a goal in mind rather than simply following a set of instructions.1997 saw the creation of the supercomputer known as "Deep Blue," which went on to claim victory against the reigning world chess champion. The development of such a big computer by IBM was a significant step forward. 2002 saw the debut of the world's first robotic vacuum cleaner to achieve widespread commercial success. 2005 - 2019 - Today, we have speech recognition, robotic process automation (RPA), a dancing robot, and other inventions making their debut. In addition, smart houses are becoming increasingly common.

2.4 Types of artificial intelligence

1) **Purely Reactive -**These machines are limited to a single area of work and do not possess any memory or data that they can use in their operations. Take the game of chess as an illustration: the computer analyses each move and selects one that gives it the greatest chance of success.
2) **Limited Memory -** These computers remember the information they have already gathered and continue to add more of it. They either have a sufficient amount of memory or enough experience to make sound choices, but their memory is limited. This computer, for instance, is able to recommend a restaurant based on the location information that has been gathered.

3) **Theory of Mind-** This type of artificial intelligence is able to comprehend both thoughts and feelings, as well as engage in social interaction. However, a device that is based on this type has not yet been constructed.

4) **Self-Aware-** The next generation of these cutting-edge technology will consist of self-aware machines. They are going to have intelligence, sensitivity, and awareness.

2.5 How does artificial intelligence work

AI systems accomplish their work by combining large-scale processing algorithms with clever, iterative ones. This combination enables artificial intelligence to learn from the patterns and features present in the data that is studied. An Artificial Intelligence system, once it completes a cycle of data processing, evaluates and assesses its own performance, and then makes use of the outcomes in order to acquire further knowledge.

Ways of implementing artificial intelligence

Let's investigate the following options, which will illustrate how we can put AI into action:

Machine learning

Machine learning is what provides artificial intelligence the ability to learn new things. In order to accomplish this, algorithms are utilized to search for patterns and derive insights from the data that they are presented with.

Deep learning

Deep learning, which is a subfield of machine learning, is what gives artificial intelligence the ability to simulate the neural network of a human brain. It is able to make sense of the data by identifying trends, noise, and sources of misunderstanding.

2.6 Neural network concept and applications

The process of the computer gaining knowledge through its own experiences is referred to as machine learning. This kind of education includes both learning from previous experiences and learning by analogies. Research in machine learning is frequently conducted in isolation from its actual practical applications. It is possible for a researcher to design a brand new classification method, and then compare the performance of that method (such as accuracy or AUC) with the performance of an already existing data set of classification models that is available to the general public in order to evaluate the method's usefulness (Chouhan & Khan, 2019). When a computer possesses the capacity for machine learning, it is automatically able to adjust itself to the complexities of its surroundings; this capacity for machine learning can be expanded with increased experience and exposure to more cases. The neural network and the genetic algorithm are the two types of machine learning software that are currently the most widely used.

The goal of the creation of artificial neural networks is to mimic the function of biological neural networks. Without first building a model of a real system (Kanimozhi & Jacob, 2019), every neuron in the network has been carefully trained to work together to tackle challenges involving artificial intelligence. These problems can be solved by the network as a whole. It is possible to define a neural network as a system that is based on a model of the structure of the human brain. The human brain is made up of a large number of nerve cells that are capable of communicating with one another. Neurons are the core units of the information processing mechanism in the brain, and these nerve cells are the fundamental components of that mechanism. Nearly 10 billion neurons can be found in a typical human brain, with an additional 60 trillion synapses and connections between these neurons. The information processing capability of the human brain is significantly faster and more powerful than that of any computer that currently exists. The neurons in the brain are responsible for this. Despite the fact that the structure of each neuron is quite simple, a sufficient number of neurons working together can generate a complex and highly developed processing system. A neuron is composed of a cell body, a soma, a great deal of fibre that is referred to as axons, and a lot of dendritic, and something called an axon, which is a long fibre. One way to think about the human brain is as a very complex, nonlinear, and parallel information processing system. Both the information processing and the information storage methods that take place in the brain are not entirely distinct from one another; rather, they may take place concurrently in the same neural network. To put it another way, these two processes are carried out not locally but rather globally throughout the neural network. The capacity for learning is the most prominent feature of a biological neural network; hence, based on this concept, computers are employed to replicate the processes of biological learning in order to achieve the functions that are required.

.

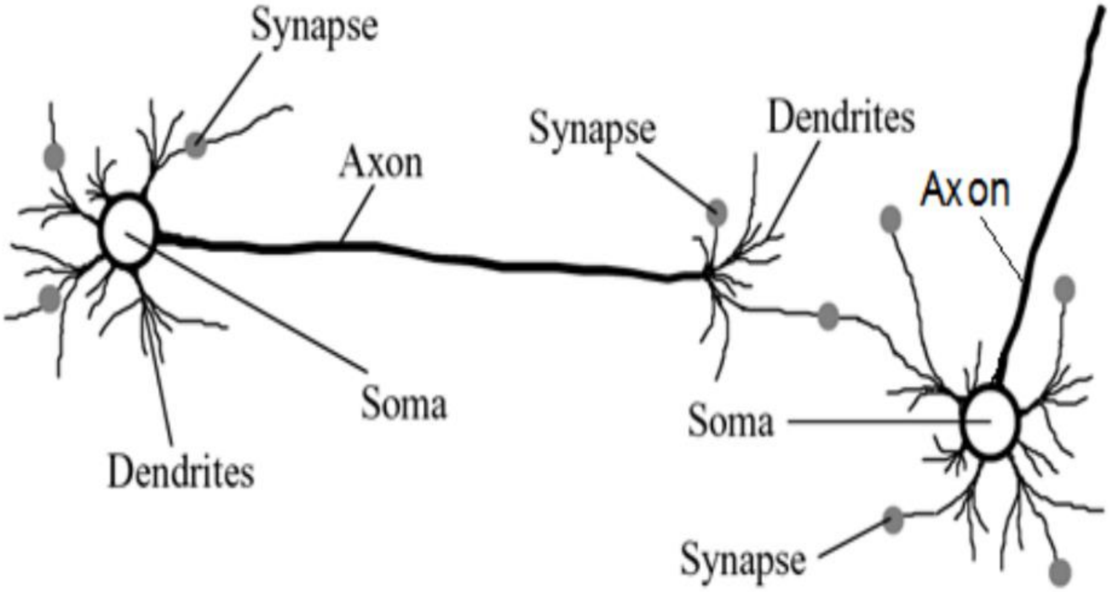


Figure.1. Biological Neural Network

An artificial neural network is made up of a number of very basic processors known as neurons. These artificial neurons are analogous to the biological neurons that are found in the human brain. The neurons in the network are connected to one another by means of weighted linkages, and it is these links that bring the network as a whole together. Signals are transmitted in this manner from one neuron to the subsequent neuron in the chain. When the neurons send out a signal, the signal will be split into a very large number of branches, all of which will send the identical signal. The incoming connections of the other neurons in the network serve as the terminal locations for the outbound branches of the network. Neural networks have been put to a significant amount of use up to this point. There are many distinct sorts of algorithms and works in the field of computer science that are based on neural networks. These algorithms and works have been applied to a broad variety of various domains over the years. If there are a number of different variables that appear at random and the user needs to determine or clarify something based on those variables, then a neural network would be a smart answer because of its potential to learn on its own and organise itself. This ability to learn and organise oneself. The following are some examples that illustrate this point: interesting research is done by contributing to the development of a model of radial basis (exact fit) artificial neural network for estimating the shelf life of burfi stored at 30 degrees Celsius (Caminero et al., 2019). The output value of this model will be the acceptability, and the results turn out quite well. The research described in reference (Raiyn, 2014) emphasises the potential of ANN models that are based on the Cascade Backpropagation algorithm for determining the shelf life of processed cheese that has been stored at a temperature of 30 degrees Celsius. This is an additional work that was completed that was pertinent to the situation. Cascade backpropagation algorithm (CBA), which was utilised by the authors in order to accelerate the learning process in ANNs; Bayesian regularisation strategy, which was utilised in order to train the network. The propose a new method for analysing road photos that frequently include automobiles and extracting licence plate (LP) data from natural characteristics by locating vertical and horizontal boundaries. This method can be used to analyse road photos that frequently include automobiles. This technique is used to analyse photographs of roads that include moving automobiles. An algorithm based on the idea of artificial neural networks is used in this paper to handle the recognition of Korean plate characters. This algorithm was developed specifically for this paper (ANN). The research presented in reference, which provides a radial basis function artificial neural network, makes use of a multilayer feed forward network to handle hydrological data. This research was carried out in the United Kingdom (Satpute et al., 2013). The values of the spread and the centre are considered model parameters in RBFANN, and their estimation is accomplished by inducing the relevant weights. After that, one can attempt to create predictions based on these values. The authors (Satpute et al., 2013)are interested in the problem of the uncertainty of the calorific value of coal. They suggest using the RBF neural network as the foundation for a soft measurement model that may be used to determine the calorific value of coal. The purpose of this model is to make the calorific value of coal more predictable by reducing the amount of uncertainty associated with it.

In addition, the evolutionary algorithm produced a fitness function in order to maximise the RBF network parameters when k-cross validation was taken into consideration. This was done in order to find the optimal solution. The article presents and discusses an efficient method for managing the traffic flow problem, which is a complicated non-linear prediction of a large-scale system(Garcia-Teodoro et al., 2009). This method may be found in the article. Backpropagation and neural networks are the foundations of this technique. The fact that the Neural Network Model is capable of auto-learning as well as adaption is one of the factors that adds to the efficiency of this. In addition, the numerous target sorts (vehicles) that may be present in an Intelligent Transport 18 System are categorised, which may be found in the aforementioned article. As a method of classification, the use of a Supervised Artificial Neural Network as a tool of soft computing is currently being explored. The quantity of energy that is returned to the radar or the Radar Cross Section (RCS) measurements that are taken at a variety of aspect angles are used here to classify the targets as one of several different categories. The authors (KR & Indra, 2010) carry out an investigation and get to a conclusion on the factors that have an impact on the utilisation of lifeless-repairable spares. Then employ a genetic algorithm that has the ability to optimise the weights and thresholds of the BP neural network, and they combine that with the BP neural network, which is used to forecast consumption. The study that is being conducted with the intention of developing methods that can be used as references for measuring the image quality of JPEG files. Additionally, an Elman neural network has been built in order to categorise photos according to the quality of the image. This was done in order to facilitate the classification process. A innovative approach is described in the article referred to as Ref. (SHAAR & Ahmet, 2018), and it takes use of the Modular Radial Basis Function Neural Network (M–RBF–NN) technology. In order to increase the performance of rainfall forecasting, this strategy is employed in conjunction with appropriate data–

preprocessing techniques, such as Singular Spectrum Analysis (SSA) and Partial Least Square (PLS) regression. Another example of a potentially beneficial application of the neural network technology is seen in Inference and Decision Systems, abbreviated as IDS. In the following section of this thesis, a backpropagation algorithm will be explained. This technique is based on neural networks and is used in intrusion detection systems (IDS).

2.3 Random Forest algorithm

The Random Forest approach is widely utilised by researchers in the field of data science, making it one of the most well-known methods or frameworks. The Random Forest technique is capable of accurately classifying enormous volumes of data and is considered to be one of the best classification algorithms available. When producing projections, the random forest can be a helpful tool, particularly when it is taken into consideration that these predictions do not adhere to the law of large numbers. The staking of the claim to the right to make their arbitrary classification systems suitable and manageable. Random Forests are straightforward to comprehend and easy to use in various professional contexts. Random Forest is particularly effective when applied to the analysis of complex data structures that are nested within a fundamental record that has fewer than 10,000 rows but potentially millions of columns. The part of a random forest that is located at the very top is known as the:

- Accuracy

- Effectively operates on massive database systems
- Capable of processing hundreds of input variables without deleting any variables
- Offers Efficient Techniques for Estimating Missing Data as well as Unbalanced Data Sets
- It is possible to store the generated forest for potential future use on the other data.

2.4 K-Means algorithm

In the field of group analysis, one of the approaches that sees the most action is called the K-means algorithm. This algorithm's goal is to object to 'n' data in the 'k' for the division, where each data object will belong to a group that contributes to the future's average value. The Euclidean metric is used as a benchmark to determine whether or not two things are equal. The k-means algorithm's features include the efficient processing of large data sets; however, it is restricted to dealing only with numerical values, therefore its application is limited in this respect. The fundamental k-means algorithm is:

1. Choose items with the number K as the early centre
2. Give each object that is adjacent to the centroid the following value
3. Determine anew where the centre of each group is located.
4. 'Continue to repeat steps 2 and 3 until the centroid has not altered.'

CHAPTER 3

DESCRIPTIONS OF DATASET

3.1 Outline

An introduction to cloud computing and several tools that are utilised in the construction of clouds on servers, such as Ubuntu Enterprise Cloud, Eucalyptus, and KVM, is presented in the beginning of the presentation. After that, I will discuss the CICIDS2017 dataset in the third section, where I will also briefly highlight the 11 datasets that I researched for my thesis.

3.2 Cloud computing

Using computing resources as a service is an example of "computing in the cloud," which refers to the practise of utilising cloud computing. Customers can be offered both software and hardware as a service over a network, the Internet being the most popular example of such a network. There are allegedly two distinct varieties of clouds, which are referred to as public clouds and private clouds respectively. A form of cloud computing known as a public cloud is one that offers its services on a pay-per-use basis to the general public rather than on a freemium model. An excellent illustration of this may be found in Reference, which discusses the Adobe Creative cloud: Customers are required to pay a predetermined amount in order to acquire the resources and tools that they require. And finally, a private cloud is often utilised in order to handle the private data that are held within an organisation and are not accessible to the general public. This type of cloud is known as a "closed cloud." For example, the article indicates that the IBM smart cloud is able to provide a private cloud service. This is performed by providing threat protection for each tier of the virtual infrastructure, limiting access to vital data, tracking user access, and getting reports from the virtual infrastructure. The following three traditional services are ones that users may be able to receive using cloud computing at this time: HaaS, SaaS, DaaS. A paradigm that enables consumers to purchase necessary computer resources in a pay-as-you-go method is referred to as "hardware as a service" (HaaS), and the term "hardware as a service" is abbreviated to "HaaS."

Customers are able to obtain the necessary computer resources through the use of this paradigm. The Amazon EC2 cloud serves as a great example of this category of service. Amazon Elastic Compute Cloud, also referred to as Amazon EC2 in some circles, is a web service that provides users with access to malleable computing resources that are stored in the cloud. It was designed with the goal of relieving the strain that web-scale computing places on the shoulders of software developers. The phrase "software as a service" is referred to by its abbreviation, "SaaS." When running in this mode, a piece of software or programme is provided to users in the form of a hosted service that is accessible via the Internet. Both the "Software Service" cloud service that Microsoft provides and the Google cloud platform are great examples of this type of service and respectively. Microsoft's Windows Azure is a cloud service that allows users the ability to swiftly develop, deploy, and manage applications over a worldwide network of datacenters that are managed by Microsoft. This service is offered by Microsoft and may be accessed through the company's website. When designing applications, the user has the freedom to choose the operating system, programming language, or tool to make use of. Additionally, the Google App Engine is a platform that is used to execute regular web applications in data centres that are maintained by Google. This platform is able to support multiple languages and is used by Google.

Monitoring, failover, and the deployment of application instances according to the requirements are all under its purview. It is also capable of handling the process of deploying code to a cluster. The file system is accessible to developers using App Engine, although this access is limited to a read-only mode. DaaS denotes data as a service. There are a lot of apps, like Adobe Buzzwords, that are interested in using it to access data from the cloud wherever and whenever it is necessary to do so. One of these apps is interested in making use of it. In addition, two distinct kinds of open-source software, notably OpenStack and Eucalyptus are typically utilised in order to implement all of these cloud-based services and with the end objective of creating solutions that are compatible with any cloud environment, Open Stack was developed to be user-friendly, incredibly scalable, and jam-packed with capabilities. In light of the fact that eucalyptus contributes to the formation of clouds as part of this research, the particulars of this plant will be discussed further on in this chapter.

3.3 Building the cloud

Figure 3-1 depicts the servers that were utilised in the construction of the cloud platform. These servers include two Dell PowerEdge R710 server machines and two Dell PowerEdge R610 server machines, each of which has a Quad-core Intel® Xeon® CPU, 20 GB RAM, and a hard disc that is 500GB in size. To simulate a cloud computing environment, 45 virtual machines, each with 256 megabytes of random-access memory (RAM), were constructed. KVM is utilised to construct instances in the cloud, while Ubuntu Enterprise Cloud is used to develop the platform that the cloud runs on.

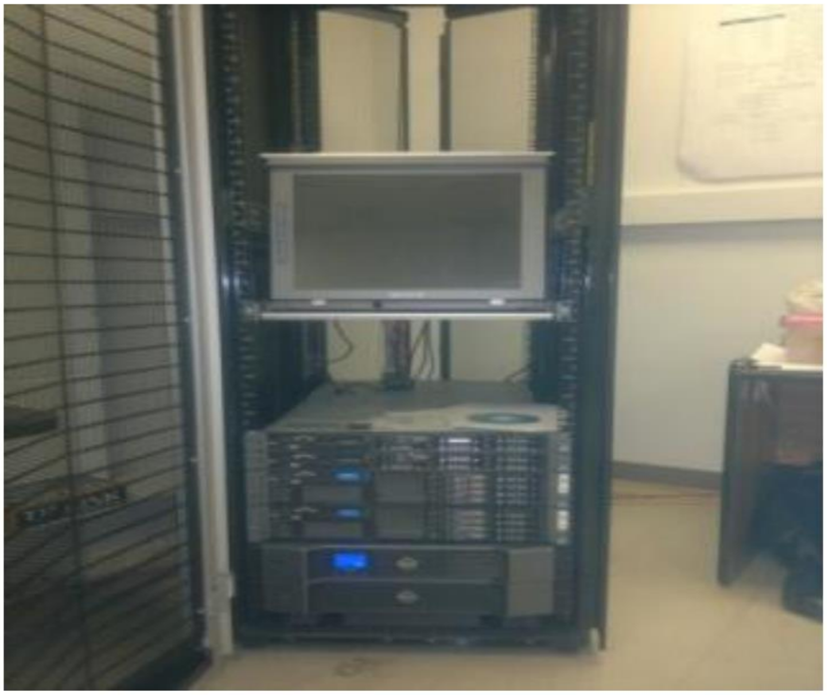


Figure. 2. The experimental cloud testbed based on Dell Power Edge R710 and R610 Servers.

3.3.1 Ubantu enterprise cloud

The Ubuntu business has come out with a brand-new category of open-source software called Ubuntu Enterprise Cloud (UEC), which is driven by Eucalyptus. The UEC is utilised in order to significantly simplify the deployment, configuration, and use of the cloud infrastructure that is based on Eucalyptus. The UEC has made the following content modifications to simplify several aspects:

1) Establish the Eucalyptus public cloud as a platform that is compatible with the Amazon EC2 infrastructure.

2) Construct a private cloud that will operate within the confines of the company's existing data centre while behind the firewall.

Up to this point, setting up and utilising Eucalyptus has been a breeze. Simply downloading a CD image of the cloud server and installing it wherever the user chooses to do so is all that is required of them. UEC is also the first open-source product that enables users to establish cloud services in a local environment and then simply access the powerful functions of the cloud. This makes UEC a pioneer in the cloud computing industry.

3.3.2 Eucalyptus

Eucalyptus, which stands for Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems, is one of the most widely used cloud platforms since it is highly developed and packed with a variety of features. Additionally, it is designed to provide an API that is interoperable with Amazon EC2. As can be seen in Figure 3-2, the Eucalyptus cloud platform is assembled from the following five primary parts:

1) The CLC, also known as the Cloud controller, is what's used to govern the virtualized resources underneath. The primary controller that is accountable for the management of the entire 29 system may be found here. It serves as the primary point of access to the cloud for all of the administrators and users. CLC will assist with the conveyance of requests to the appropriate components, then it will collect responses from those components and send them back to the customers. CLC is the portal through which the Eucalyptus cloud can communicate with the rest of the world.

2) Scalability and access control of virtual machines can be accomplished via The Walrus, which offers a service that is analogous to S3. 3) The CC, or cluster controller, is responsible for managing both the executions and the networking of the entire cluster. CC is in charge of controlling the life cycle of instances that are running on those nodes and is responsible for maintaining the information of all of the relevant Node Controllers that are operating inside the cluster. Requests from virtual instances will be forwarded to the Node Controller, along with any relevant information regarding available resources.

3) The SC, sometimes known as the Storage controller, is responsible for managing the storage in a cluster. SC and Walrus collaborate to store and retrieve user data, as well as pictures of virtual machines and kernels, RAM disc images, and virtual machine images.

4) Activities in VM instances are controlled by at least one NC, also known as a node controller. The operating system on the host computer and the hypervisor that corresponds to it, such as KVM or Xen, are both controlled by NC. As the thesis's chosen hypervisor, KVM was a natural choice.

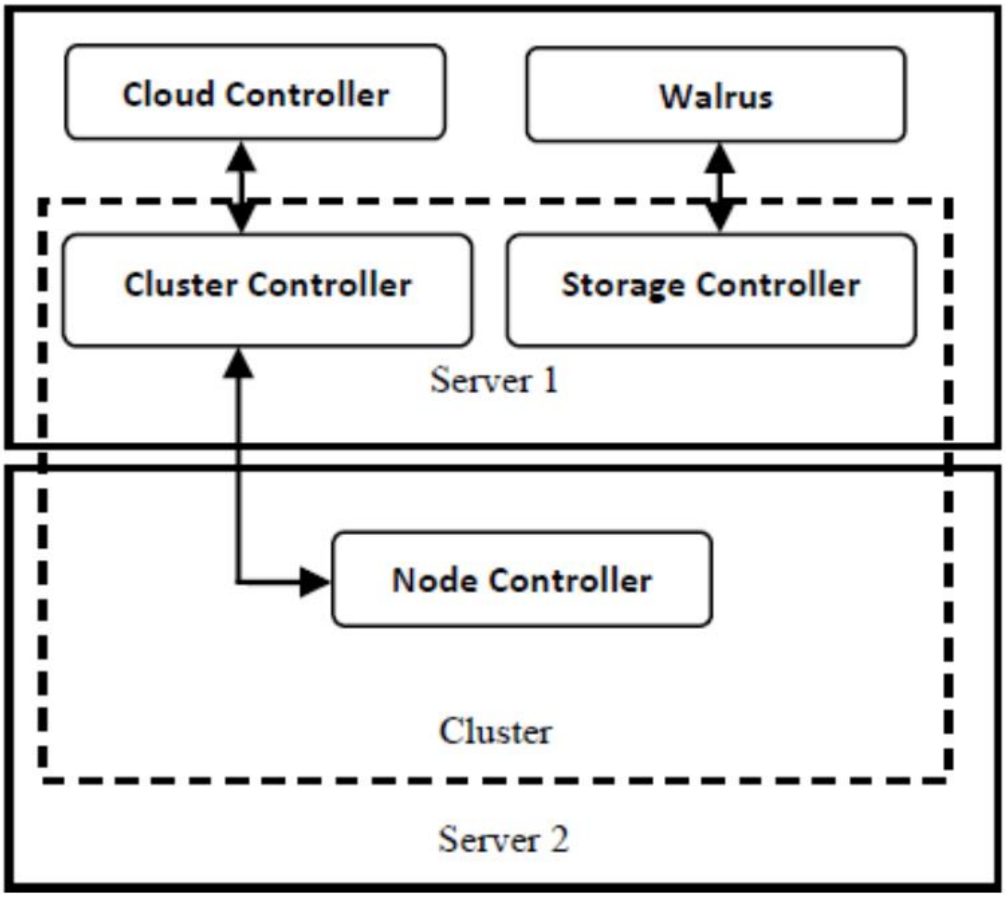


Figure.3. Architecture of the Eucalyptus Cloud

3.3.3 Description of CICIDS2017 dataset

As much as 98 percent+ accuracy and less than 1 per cent false alarms are already claimed by researchers in the field of intrusion detection. Researchers and manufacturers were compelledto devote money and effort to the development of useful goods because of this high rate of accuracy. In reality, only a few models have been recognized by the industry to design an IDS. By analysing contemporary IDS models and training and testing datasets, we are able to identify the root cause of this issue. The Canadian Institute of Cybersecurity's CICIDS2017 collection provides the most up-to-date attack scenarios. This cutting-edge dataset not only includes the most recent network assaults, but it also meets all of the criteria for attacks that really occur in the real world. We noticed just a few flaws in this dataset whenwe investigated its properties. An obvious flaw is a large dataset, which was compiled from five days' worth of Canadian Institute of Cybersecurity traffic data spread over eight files. AnIDS might be designed from a single dataset. There are a lot of redundant entries in the dataset, making it unsuitable for training any IDS. Even if the dataset comprises contemporary assault scenarios, we also discovered that the dataset has a substantial class imbalance. Class imbalance datasets can mislead the classifier, biassing it towards the majority class. " The research community was given a subset of the CICIDS2017 dataset to work with in developing and testing detection algorithms in an attempt to address these issues.

3.3.4 AVAILABLE DATASETS

In this section, we ran comparison research on 11 IDS datasets that have been publicly available since 1998, and we discussed the inadequacies of those datasets, which point to the necessity of a new dataset that is both comprehensive and dependable.

3.3.4.1 DARPA (Lincoln Laboratory 1998–99)

This dataset was developed for the purpose of conducting an examination of network security, and it shed light on the problems caused by fake injection assaults as well as benign traffic. This dataset includes information about the following types of activities: e-mail, surfing, FTP, telnet, IRC, and SNMP monitoring. This package protects against a variety of attacks, including denial of service, guess password, buffer overflow, remote file transfer protocol, system flood, Nmap, and Rootkit. It does not include genuine attack data records, it does not have false positives, and it does not represent the network traffic that occurs in the real world. These are the negative aspects of it. It is therefore not applicable to the evaluation of IDSs on contemporary networks, both in terms of the many forms of attacks and the infrastructure of the network.

3.3.4.2 KDD’99 (University of California, Irvine 1998–99)

The tcpdump data was processed in order to produce this revised version of DARPA98, which was then saved as a dataset. It includes a variety of attacks, including buffer overrun, Smurf DoS, Neptune DoS, and pod DoS. Within a simulated environment, both benign traffic and attack traffic are combined together. It has a significant number of duplicate records and is riddled with data corruptions, both of which contribute to inaccurate testing findings. NSL-KDD was developed by the utilisation of KDD [31] in order to address some of KDD's deficiencies.

3.3.4.3 DEFCON (The Shmoo Group, 2000–2002)

Attacks such as port scanning and buffer overflow are included in the DEFCON8 dataset, which was established in the year 2000. The DEFCON10 dataset, on the other hand, was developed in 2002 and covers threats such as port scanning and sweeps, malformed packets, administrator privilege, and FTP using telnet protocol. When compared to the traffic that would be seen on a network in the real world, the traffic that was generated during the capture the flag (CTF) competition that was included in this particular data set was very different. This is because the bulk of the CTF competition traffic was made up of intrusive traffic, as opposed to the typical background traffic. This dataset was utilised in the assessment of several alert correlation strategies.

3.3.4.4 CAIDA (Center of Applied Internet Data Analysis 2002–2016)

This organisation maintains not one, not two, but three distinct data sets. (a) The CAIDA OC48 database stores a wide range of information that was obtained from an OC48 link located in San Jose. (c) CAIDA Internet traces 2016, which are active traffic traces from CAIDA's Equinix-Chicago monitor on the high-speed internet backbone. (b) CAIDA DDOS, which is passive traffic traces from CAIDA's Equinix-Chicago monitor on the high-speed internet backbone. (d) CAIDA Internet traces 2015, which is passive traffic traces from CAIDA's Equinix-Chicago monitor on the high-speed internet backbone. Both of these files are split into 5-minute pcap files every 5 minutes. The vast majority of CAIDA's datasets are highly unique to a given event or attack, and while they are anonymised with regard to their payload, protocol details, and destination, they cover a wide range of information.

3.3.4.5 LBNL (Lawrence Berkeley National Laboratory and ICSI 2004–2005)

The dataset consists of the entire headers of the network traffic that was recorded at a location of medium size. It is severely anonymized, meaning that no information is deleted that may be used to identify a particular IP address, and it does not carry a payload.

3.3.4.6 ISOT (Intrusion Dataset 2008)

There are both harmful and non-malicious datasets included in the set. The benign portion was produced by combining two datasets: (a) a dataset from the traffic lab at Ericsson Research that includes various forms of benign traffic such as traffic for web browsing, gaming, and torrenting; and (b) a dataset from the Lawrence Berkeley National Lab (LBNL) that includes various forms of benign traffic such as traffic for web, email, and streaming media applications. Together, these two datasets were combined to produce the benign portion. The benign component was produced by combining both datasets in order to do so. The malevolent portion includes traffic that was generated by the Storm and Waledac botnets. In order to integrate the three datasets, the techniques that came with each of the datasets were used. It has one subnet dedicated to malicious activity and contains 23 subnets that handle normal traffic. Seven flow-based characteristics and four host-based features are included in each flow.

3.3.4.7 CDX (United States Military Academy 2009)

This dataset records the results of network warfare tournaments and can be applied to the creation of contemporary datasets with labels. All types of service traffic, including web traffic, email traffic, DNS lookups, and other types of service traffic, are included. Attack tools such as Nikto, Nessus, and Web Scarab were used by the attackers so that they could carry out automated reconnaissance and attacks. These tools were utilised by the attackers. It is feasible to use it to evaluate the warning criteria of an intrusion detection system, but it suffers from a lack of traffic diversity and volume, thus this evaluation cannot be done very accurately.

3.3.4.8 Kyoto (Kyoto University 2009)

This dataset was obtained from honeypots, thus there is no tagging or anonymization; nonetheless, it only provides a restricted perspective of network traffic because it is only feasible to see attacks that are aimed at the honeypots. Honeypots are used to collect information about network traffic. In addition to the preceding datasets that may be used for NIDS research and evaluation, it comes with eleven more capabilities, some of which are IDS detection, malware detection, and Ashula detection. For the purpose of the simulation of typical traffic, only DNS and email traffic statistics were used, which means that the simulation does not correctly reflect normal traffic in the real world. Therefore, there are no false positives, which is essential for reducing the total number of notifications.

3.3.4.9 Twente (University of Twente 2009)

OpenSSH, the Apache web server, and Proftp are the three distinct services that make up this dataset. When gathering information from a honey pot network, Authtident was used on port 113, and NetFlow was utilised. There is activity going on in the network at the same time, and some of it includes things like authident, ICMP, and IRC traffic. These acts are neither completely good nor completely evil in and of themselves. In addition to that, this dataset contains some unknown alerts traffic that is not associated with anything else in it. Although it is labelled and has a more realistic appearance, it is obvious that there is a lack of both volume and variation in the strikes.

3.3.4.10 UMASS (University of Massachusetts 2011)

There are some traces from wireless applications included in the dataset, in addition to trace files, which are analogous to network packets [UMASS 2011] [Nehinbe 2011]. It was created by employing a single TCP-based download request assault scenario, which was then utilised in its production. Because there is not enough variety in the dataset's traffic and attacks, it cannot be utilised for evaluating solutions for intrusion detection and prevention systems (IDS and IPS).

3.3.4.11 ISCX2012 (University of New Brunswick 2012)

This dataset contains two profiles: the alpha-profile, which carried out a variety of multi-stage attack scenarios; the beta-profile, which is the benign traffic generator; and the gamma-profile, which generates realistic network traffic with background noise. Both of these profiles were generated using the same data set. These two profiles are included in the CICIDS2017 5 database as separate entries. In addition to the full packet content, it supports communication via the protocols HTTP, SMTP, SSH, IMAP, POP3, and FTP. The distribution of the simulated attacks, on the other hand, does not accurately reflect reality because it does not include any traces of HTTPS traffic, despite the fact that HTTPS accounts for almost 70 percent of all network traffic nowadays. In addition to this, the distribution of the simulated attacks is not based on the statistics of the real world.

3.3.4.12 ADFA (University of New South Wales 2013)

In addition to the standard training and validation data, this dataset also includes ten assaults for each vector. It features an attack using brute force on FTP and SSH passwords, a Meterpreter written in Java, the ability to establish a new superuser, a Linux Meterpreter payload, and C100 Web shell attacks. In addition to the lack of assault diversity and variation, the behaviours of some attacks in this dataset are not adequately distinguishable from the regular behaviour. This is the case despite the fact that there is a lack of assault variety and variation. This presents a dilemma because there are not enough different kinds of attacks to pick from.

3.3.4.13 CTU-13 (CTU University 2013)

The CTU University in the Czech Republic is responsible for the creation of this dataset. The collection includes botnet traffic together with benign traffic and communication traffic that occurs in the background. This dataset makes use of records from bidirectional Netflow. They came up with 13 unique situations and collected malware traffic corresponding to each of those scenarios. As a guest operating system, they used Windows XP Service Pack 2, and the host operating system was Linux Debian. Each and every one of them had a working connection to the university's network. When it came to labelling, at first every type of traffic was classified as background traffic. It was determined that the traffic that originated through switches, proxies, and authentic computers was safe to proceed with. The term "botnet traffic" was applied to any and all traffic that originated from compromised workstations.

3.3.4.14 SSHCure (University of Twente 2014

SSH assaults carried out on a campus network are included in this dataset [38]. SSHCure stores NetFlow records that were taken from Cisco 6500 series routers and exported from those routers. It is composed of two parts, both of which were gathered on the UT campus over the course of a month. Each section depicts distinct circumstances. The first section is comprised of SSH traffic aimed towards honeypots, and the second section is comprised of SSH traffic originating from regular servers. There are 11348 recorded instances of attacks.

3.3.4.15 UGR '16 (University of Granada 2016)

Specifically for the goal of conducting research on cyclostationarity-based network IDSs, this dataset was generated by the University of Granada. The information for the dataset was compiled in an Internet service provider of tier 3 during the course of a period of four months. They modified the Netflow records so that the identifiers read differently. The dataset does not give a very wide range of different kinds of assaults. In addition to this, they blended controlled-environment botnet captures with background traffic, which decreases the overall quality of the dataset.

CHAPTER 4

PRAPOSED STUDY

4.1 Outline

In this chapter we will discuss the whole process which I have worked regarding intrusion detection system.

4.2 Selecting a Datasets

The evaluation was the primary factor in our decision to go with CICIDS2017. The only dataset that fulfilled all 11 evaluation criteria was this one. It is made up of two networks, one of which is known as the attack network, while the other is known as the victim network. The victim network has a highly secure infrastructure that comprises a firewall, router, switches, and the majority of the popular operating systems. This infrastructure protects the network from outside threats. In addition, each computer in the network is equipped with an agent that is accountable for the behaviour that is regarded as being harmless. The attack network is a completely autonomous infrastructure that is constructed using a router, a switch, and a collection of personal computers (PCs) that each have their own public IP address. This is done in order to carry out the various attack scenarios. The production of realistic background traffic is one of the most essential goals that needs to be accomplished for IDS and IPS databases, which is one of the many key tasks that need to be accomplished. A CIC-B-Profile system (Satpute et al., 2013) was utilised in order to fulfil the requirements of this dataset. This system is responsible for profiling the abstract behaviour of human interactions, and it also supplies natural and benign background traffic. This B-Profile is an abstraction of the behaviour of 25 users based on the protocols HTTP, HTTPS, FTP, SSH, and email. It was created using the data from the dataset. The CICIDS2017 dataset is one that I have worked on, and the explanation can be found in the figure below.

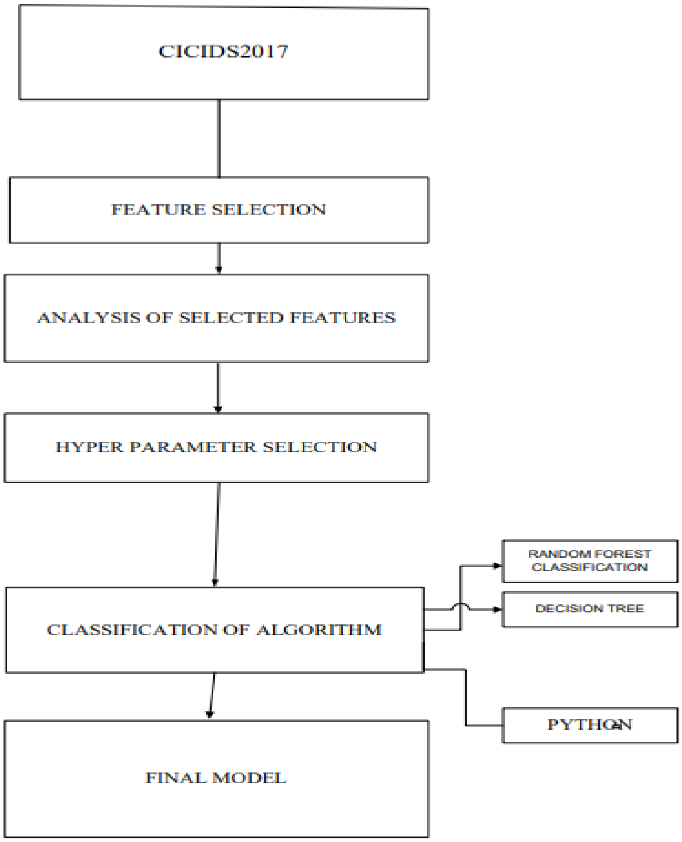

CICIDS2017
FEATURE SELECTION
ANALYSIS OF SELECTED FEATURES
HYPER PARAMETER SELECTION
CLASSIFICATION OF ALGORITHM
RANDOM FOREST CLASSIFICATION
DECISION TREE
PYTHON
FINAL MODEL

Figure.4. Work flow diagram of CICIDS

4.3 Classification of Algorithm

In this study, performance in terms of accuracy, learning capacity, scalability, and speed are the most important characteristics to take into account when choosing classifier algorithms. Five algorithms, including Random Forests, Bayesian Network, Random Trees, Naive Bayes, and J48 classifiers, have been researched and shown to support this hypothesis. Using the Information Gain feature selection, this paper shows that random forest trees are capable of learning and performing well when it comes to detecting assaults. The Bayesian Network surpasses other algorithms when it comes to categorizing assaults. Random Tree is a scalable and efficient method. Since Naive Bayes has a low model complexity, it is a better choice for classifying data than other algorithms.

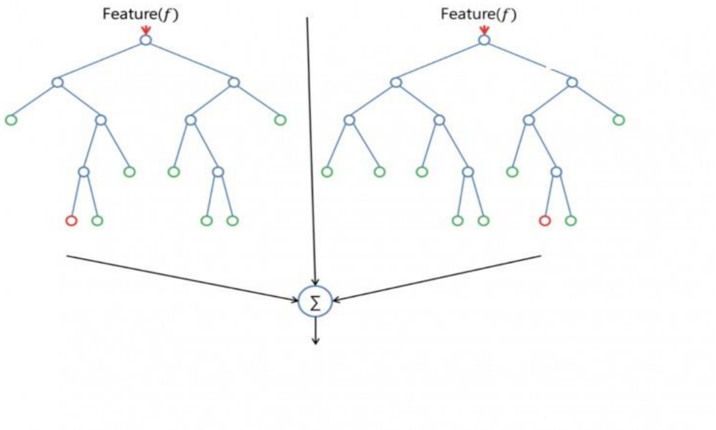


Figure.5. Random Forest classifier tree

1) Random Forest (RF)

Ensemble classifier approaches include Random Forest. A "forest" of classifiers is a decision tree classifier ensemble. To generate each decision tree, qualities are randomly selected at each node. In 2001, Breich introduced the random forest algorithm.

2) Bayes Network (BN)

Probabilistic connections between variables of interest are encoded using a Bayesian Network (BN) modelling technique. Assumptions about the model behaviour of the target system are used to determine how accurate this technique is. If the assumption is significantly altered, then detection accuracy is reduced.

3) Random Tree (RT)

Random Tree is a decision tree based on a set of random characteristics (random). A decision tree is made up of a number of nodes and branches that may be connected in various ways. A test attribute is represented by a node, and the results are represented by branches. In the form of class labels, decision leaves display the final choice made after the computation of all attributes.

4) Naive Bayes (NB)

According to the Bayesian categorization system, the likelihood of belonging to a given class may be predicted statistically. Based on the Bayes theorem, we may classify data in a Bayesian fashion. As the Nave Bayes classification, the Bayesian classification is better recognized by its more formal name. Ignoring other attribute values, Nave Bayes considers that attribute values have no effect on the class they belong to.

5) J48

Machine learning algorithm J48 or C4.5 is frequently used and is part of the decision tree algorithm. The entropy idea is used to form a decision tree in this technique.

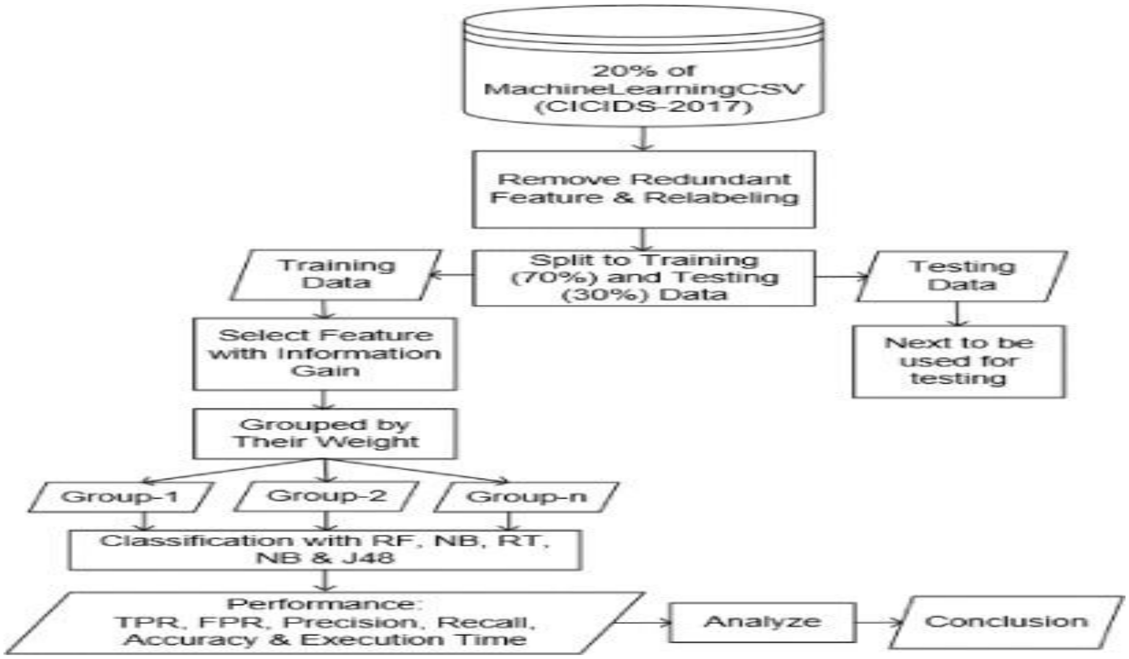

20% of MachineLearningCSV (CICIDS-2017)
Remove Redundant Feature & Relabeling
Split to Training (70%) and Testing (30%) Data
Training Data
Testing Data
Select Feature with Information Gain
Next to be used for testing
Grouped by Their Weight
Group-1
Group-2
Group-n
Classification with RF, NB, RT, NB & J48
Performance: TPR, FPR, Precision, Recall, Accuracy & Execution Time
Analyze
Conclusion

Figure. 6. Experimental Design

- To classify each feature group or feature subset, respectively, the Random Forest (RF), Bayes Net (BN), Random Tree (RT), Naive Bayes (NB), and J48 classifiers are utilised. In order to carry out the analysis, the following considerations are taken into account: the True Positive Rate, the False Positive Rate, the Precision Rate, the Recall Rate, the Accuracy Rate, the Proportion of Incorrectly Categorized Data, and the Execution Time of the Analysis. At this point in the procedure, a strategy known as 10-fold cross-validation is utilized.
- It is an absolute must to analyse and contrast the TPR, FPR, Precision and Recall, Accuracy, Percentage of Incorrect Categorization, and Execution Time of each classifier algorithm. At each and every level of the learning and testing process, a ten-fold cross-validation is conducted. It is essential that you arrive at some inferences or
- conclusions at this juncture.

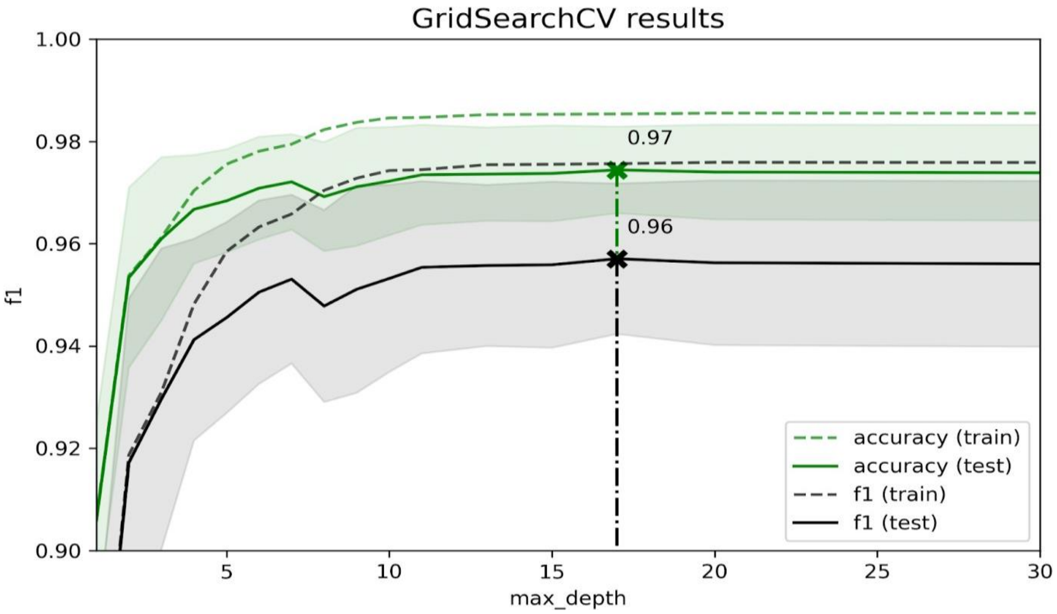


Figure. 7. Accuracy Graph

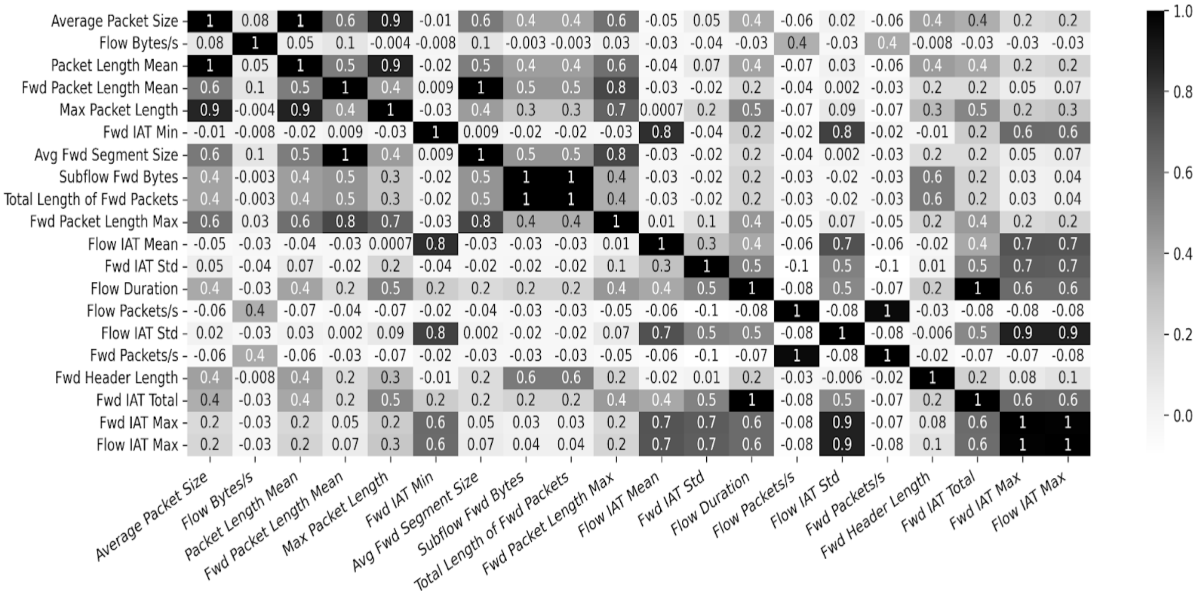


Figure. 8. Correlated Heat Map

4.3 Experimental Result

In addition to the five distinct classifier methods, the True Positive Rate (TPR), the False Positive Rate (FPR), Precision, Recall, Accuracy, Percentage of erroneously Classified, and Execution Time are the metrics that are used to evaluate the efficacy of Information Gain. These metrics are referred to collectively as the metric set. Throughout the course of the training, various different points in time are used to simulate and perfect the actual execution (the time measured from the classification process starts until the classification process stops). During this experiment, the RT, BN, RT, NB, and J48 classifiers are combined in a wide variety of different ways to classify each unique feature subset. RT stands for random tree; BN stands for random tree; RT, NB, and J48 stand for random tree; and RT, RT. A 10-fold cross-validation method was used during the course of this investigation in order to ascertain whether or not categorization algorithms are successful. The 10-fold cross-validation is used since it reduces the total amount of time spent calculating while also preserving the accuracy of the classification methods. As a direct and immediate result of this, 10 random folds of the input dataset with the same size will be constructed from it. During the process of cross-validation, nine of the ten-fold data sets will be employed for training, while just one of the ten-fold data sets will be utilised for testing. The ultimate outcome is a test fold, which is produced after ten repetitions of this procedure.

4.4 Overall Process

1) **Procedure** Process ()
2) Input: Fr = Feature Ranked data
3) The output includes the following: features subsets, TPR, FPR, accuracy, recall, and precision
4) Decrease the number of features from 77 to n depending on the feature weight.
5) For each and every function Fr in the data for Feature Ranked
6) Begin to choose features using the Feature Weight, and then save them on Feature Groups

Group1 is comprised of any characteristic that has a weight more than or equal to 0.6

Any characteristics that have a weight that is more than or equal to 0.5 are included in Group2. Group3 is comprised of all characteristics with weights more than or equal to 0.4 Group4 is comprised of any feature that has a weight more than or equal to 0.3 Group5 is comprised of all features that have a weight greater than or equal to 0.2. Group6 is comprised of all features that have a weight greater than or equal to 0.1. Group7 denotes the whole of the characteristics.

7) Regarding each of the Feature groupings
8) Using CICIDS-2017-20 percent, provide selected features to RF, BN, RT, and NB, as well as J48.
9) Apply Classifier Accuracy of the Random Forest model, denoted as C1 C2 equals the accuracy

of the Bayes Network model C3 equals the accuracy of the Random Tree model C4 = Naïve Bayes model accuracy C5 = J48 model correctness

10) Accuracy, recall, and precision of TPR and FPR calculations have to be determined**.**

11) Examine the Accuracy of C1, C2, C3, C4, and C5 and Compare Them.

Classifiers that use four characteristics chosen by Information Gain are mentioned below. Other classifiers are outperformed only by the RF and RT, which have an accuracy of 96.48 percent. RF, on the other hand, has a value of NaN. The term "NaN" stands for "Not a Number" or "undefined," respectively. NB has a higher TPR for identifying DoS/DDoS attacks than other classifiers, but a lower TPR for recognizing normal and infiltration traffic. Comparatively, BN has the lowest FPR (0.010) of all of the companies studied. Classifiers can only identify DoS/DDoS, PortScan, and Brute Force assaults using these four (4) characteristics. Only NB is affected by this in terms of normal traffic.

Table. 1. Performance Metrics using four features

Detection	RF	BN	RT	NB	J48
Normal	0.960	0.943	0.960	0.174	0.961
DoS/DDoS	0.992	0.996	0.992	0.999	0.991
Port Scan	0.995	0.992	0.995	0.983	0.995
Bot	0.438	0.642	0.430	0.687	0.381
Web attack	0.072	0.031	0.072	0.000	0.072
Infiltration	0.000	0.000	0.400	0.400	0.000
Brute Force	0.792	0.991	0.792	1.000	0.790
Recall	0.965	0.962	0.970	0.903	NaN
Precision	NaN	0.953	0.965	0.335	0.965
FPR	0.016	0.010	0.016	0.026	0.016

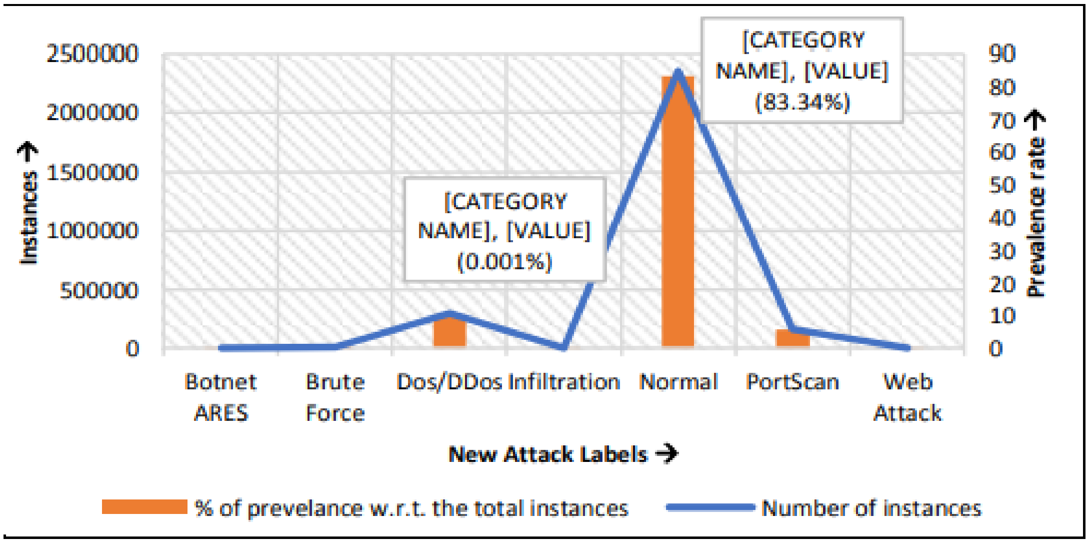


Figure. 9. Graphical representation of majority and minor class prevalence of CICIDS2017 dataset with respect to new attack labels

CHAPTER 5

RESULT & DISCUSSION

5.1 RESULT & DISCUSSION

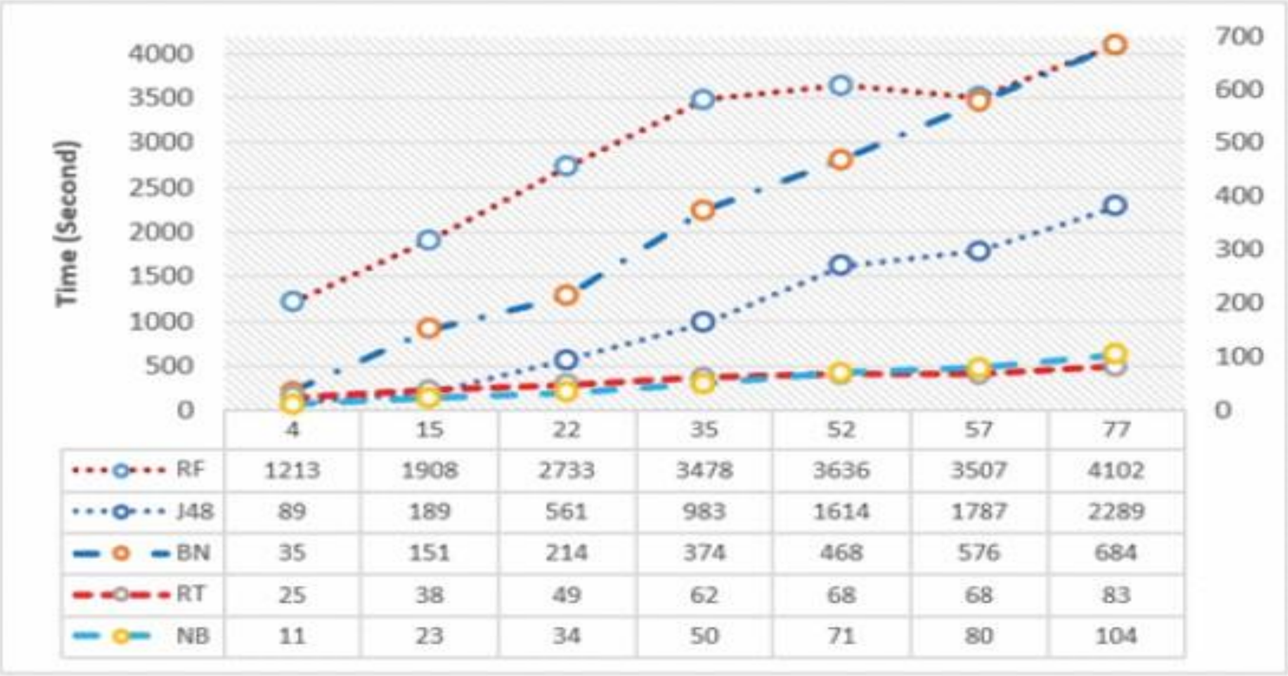
The below table shows the accuracy of CICIDS2017 DATA SET as 97.8% while using random classifier algorithm in Machine learning where base platform was written in Python library. In this research article we have considered a most recent dataset CICIDS2017 for detailed analysis keeping in view its increasing demand in the research community. Various shortcomings of the dataset have been studied and outlined. Solutions to counter such issues has also been provided. We tried to solve such issues through experiment. We also relabel the dataset with the labelling information provided by Canadian Institute of Cybersecurity.

Table. 02 - Comparative study of CICIDS DATASET with laboratory dataset

Accuracy =	**0.9784502521779** **CICIDS** **DATASET**	0.95665332 ISCX2012 DATA SET	0.941203753 ISOT DATASET
Precision =	0.9675425038639877	0.9123723343	0.897363532
Recall =	0.9601226993865031	0.9452721123	0.927352748
F1 =	0.9638183217859891	0.956633893	0.9463782936

.

For the particular feature method under consideration, this study additionally examines the impact of execution time. An overview of the execution time for each feature subset employing RF, J48, BN RT and NB can be seen in the image below. There is a major influence on the RF, J48, and BN of the pertinent features procedure. RT and NB have extremely short run times. As a rule of thumb, the more characteristics to evaluate, the more time it takes to complete.



	4	15	22	35	52	57	77
RF	1213	1908	2733	3478	3636	3507	4102
J48	89	189	561	983	1614	1787	2289
BN	35	151	214	374	468	576	684
RT	25	38	49	62	68	68	83
NB	11	23	34	50	71	80	104

FIGURE. 10. - EXECUTION TIME GRAPH

5.2 Summary

In order for intrusion detection systems (IDS) to acquire the ability to learn and evolve, which in turn makes them more accurate and efficient in the face of an enormous number of unpredictable attacks, advanced methods and techniques of soft computing and artificial intelligence are widely used in IDS. In this thesis, an intrusion detection system (IDS) based on a neural network is created on a cloud platform. It has been demonstrated that the implemented IDS has a good level of accuracy, and the amount of time spent is reasonable.

CHAPTER 6

Conclusion and future work

6.1 Conclusion and future work

In order to demonstrate the influence of feature selection on enhancing anomaly detection accuracy, this study conducted experiments. Because of its ability to accurately compute the weight of features' information, Information Gain has been recognized as the best information classifier for the trials employing feature sets 15, 22, and 35. J48, on the other hand, performs best with feature sets 52, 57, and 77. Although BN has a lower degree of precision than RF and J48, it is nevertheless able to detect all traffic utilizing feature subsets 52, 57, and 77, despite its lower level of accuracy. Experiments have also shown that the chosen traits reduce FPR, particularly for BN. Experimental results show that the number of features picked has an impact on the execution time of a programme. Ranking characteristics according to weight values is what the Information Gain proposes to do. It is, nevertheless, necessary for an expert to decide the minimum weight value, which influences the number of characteristics that will be picked. We intend to experiment with a variety of feature selection strategies in order to come up with the best possible mechanism. Each feature subset that impacts an assault will be analysed as part of future research.

References

1) A. M. ALEESA, MOHAMMED YOUNIS, AHMED A. MOHAMMED, NAN M. SAHAR, DEEP-INTRUSION DETECTION SYSTEM WITH ENHANCED UNSW-NB15 DATASET BASED ON DEEP LEARNING TECHNIQUES. Journal of Engineering Science and Technology Vol. 16, No. 1 (2021) 711 – 727

2) Amir Haider Malik, Muhammad Adnan Khan, Muhib Ur Rahman, Abdur Rehman Sakhawat (2020) A Real-Time Sequential Deep Extreme Learning Machine Cybersecurity Intrusion Detection System. Computers, Materials and Continua 66(2):1785-1798 DOI:10.32604/cmc.2020.013910

3) Axelsson, S. (1998). *Research in intrusion-detection systems: A survey*. Technical report 98–17. Department of Computer Engineering, Chalmers ….

4) Brown, C., Cowperthwaite, A., Hijazi, A., & Somayaji, A. (2009). Analysis of the 1999 darpa/lincoln laboratory ids evaluation data with netadhict. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–7.

5) Barida Baah, Chioma Lizzy Nwagbo. An Enhanced Network Intrusion Detection System Using Data Mining KEYWORD: Network Intrusion, false positive, false negative and data mining. June 2021

6) Caminero, G., Lopez-Martin, M., & Carro, B. (2019). Adversarial environment reinforcement learning algorithm for intrusion detection. *Computer Networks*, *159*, 96–109.

7) Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, *3*(5), 698–713.

8) Caswell, B., & Beale, J. (2004). *Snort 2.1 intrusion detection*. Elsevier.

9) Chebrolu, S., Abraham, A., & Thomas, J. P. (2005). Feature deduction and ensemble design of intrusion detection systems. *Computers & Security*, *24*(4), 295–307.

10) Chittur, A. (2001). Model generation for an intrusion detection system using genetic algorithms. *High School Honors Thesis, Ossining High School. In Cooperation with Columbia Univ*.

11) Chouhan, N., & Khan, A. (2019). Network anomaly detection using channel boosted and residual learning based deep convolutional neural network. *Applied Soft Computing*, *83*, 105612.

12) Crosbie, M., & Spafford, E. H. (1995). *Active defense of a computer system using autonomous agents*.

13) Daumé, H. (2012). *A course in machine learning. ciml. info*.

14) Idriss Idrissi, Mohammed Boukabous, Mostafa Azizi, Omar Moussaoui (2021). Toward a deep learning-based intrusion detection system for IoT against botnet attacks. IAES International Journal of Artificial Intelligence (IJ-AI) 10(1):110-120 DOI:10.11591/ijai.v10.i1.pp110-120

15) Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security, 28*(1–2), 18–28.

16) Kanimozhi, V., & Jacob, T. P. (2019). Artificial intelligence based network intrusion detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. *2019 International Conference on Communication and Signal Processing (ICCSP)*, 33–36.

17) KR, K., & Indra, A. (2010). Intrusion Detection Tools and techniques–a Survey'. *International Journal of Computer Theory and Engineering, 2*(6), 901.

18) MOHAMMAD NASRUL AZIZ, TOHARI AHMAD, CLUSTERING UNDER-SAMPLING DATA FOR IMPROVING THE PERFORMANCE OF INTRUSION DETECTION SYSTEM. Journal of Engineering Science and Technology Vol. 16, No. 2 (2021) 1342 – 1355.

19) Omar Almomani, A Hybrid Model Using Bio-Inspired Metaheuristic Algorithms for Network Intrusion Detection System. March 2021 Computers, Materials and Continua 68(1):409-429 DOI:10.32604/cmc.2021.016113.

20) Raiyn, J. (2014). A survey of cyber attack detection strategies. *International Journal of Security and Its Applications*, *8*(1), 247–256.

21) Satpute, K., Agrawal, S., Agrawal, J., & Sharma, S. (2013). A survey on anomaly detection in network intrusion detection system using particle swarm optimization based machine learning techniques. *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, 441–452.

22) SHAAR, F., & Ahmet, E. F. E. (2018). DDoS attacks and impacts on various cloud computing components. *International Journal of Information Security Science*, *7*(1), 26–48.

23) Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security, 31*(3), 357–374.

24) Yin Luo (2021). Research on Network Security Intrusion Detection System Based on Machine Learning. International Journal of Network Security, Vol.23, No.3, PP.490-495, May 2021 (DOI: 10.6633/IJNS.202105 23(3).14).

25) Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, Network intrusion detection system: A systematic study of machine learning and deep learning approaches.

January 2021 Transactions on Emerging Telecommunications Technologies 32(2) DOI:10.1002/ett.4150

PLAGIARISM CHECK REPORT

SRIjAN

ORIGINALITY REPORT

6%	4%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	docplayer.net Internet Source	4%
2	"Intelligent Information and Database Systems", Springer Science and Business Media LLC, 2012 PubliCation	2%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

PUBLICATIONS

International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211
Volume 10, Issue 6, June-2022, Impact Factor: 7.429, Available online at: www.ijaresm.com

An Analysis of Intrusion Detection Systems Based on Artificial Intelligence

Srijan Mishra[1], Dr Sashank Singh[2]

[1]Advance Computing & data Science Department, Integral University, Lucknow, India
[2]Assistant Professor, Advance Computing & data Science Department, Integral University, Lucknow, India

---*****************---

ABSTRACT

Internet-based communication and commercial services are provided by hundreds of millions of computers around the world using a variety of hardware and software platforms. Even though computers are connected, criminals can exploit this to launch Internet attacks.. An ever-growing number of Internet dangers necessitates the development of security-focused solutions that are both nimble and adaptive. One of the most critical tools for spotting online dangers is an intrusion detection system (IDS) IDS have been developed using a wide range of approaches from different fields of study. There are many advantages to using artificial intelligence (AI)-based techniques when it comes to developing IDS. The current status of AI-based solutions in tackling intrusion detection challenges has not been examined and appreciated comprehensively, however. IDS generation was the focus in this study, which looked at various AI-based methodologies. Data sources, processing criteria, techniques used and datasets were all examined. Classifier design and feature reduction approaches were also used in the comparison of experimental environments. Discussions about the advantages and disadvantages of AI-based approaches have already taken place. Readers will gain a better understanding of the research methods employed by IDS by reading this document. IDS and related fields will benefit greatly from this study's findings because they shed light on the existing literature. The publication also discusses potential future possibilities for this area of study's research.

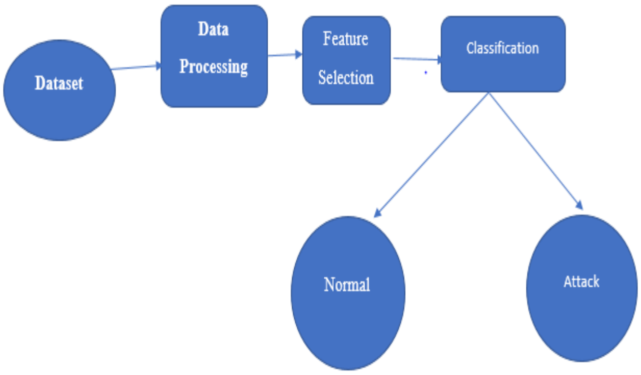
Keywords—IDS , optimization, technologies

I. INTRODUCTION

As a result of the rapid expansion and extension of the Industry Wide Web and network technologies, the computing world is confronted with huge changes and difficulties that must be met. Over the past several decades, intrusion detection systems have been the subject of extensive research and development. In part, this is due to an increase in computer and network dangers in recent years, and computerised evaluation has become an important aspect of information technology security [11] Viruses, self-propagating worms, and distributed denial-of-service attacks are all severe threats to the internet and the infrastructures that rely on it, and they are becoming more common. It is the process of analysing unauthorised access to computer systems on a network in order to determine who is acquiring unauthorised access (crackers) and who has legal access but is abusing it (hackers).

Three basic components of an intrusion detection system (IDS) are an information source that broadcasts a stream of event records, an analysis component that does statistical analysis, and a notification component that notifies the user of an incursion. Vol. 5, No. 1 (January 2013) 72. The International Publication of Network Security and Its Applications is a peer-reviewed journal that publishes original research on network security and its applications (IJNSA) A decision maker who applies a set of rules to the outputs of the analysis engine and makes decisions about what actions should be made as a result of the analysis engine's results is defined as follows: [13] Recent research has produced a large number of results that seek to combine data mining and machine learning approaches to intrusion detection systems in order to construct more intelligent intrusion detection models, according to a review of the literature published in 2014.

On the basis of their capabilities, intrusion detection systems can be classified into two categories: anomaly detection and abuse detection (often known as "abuse detection"). Host-based intrusion detection examines information obtained through network communications; network-based intrusion detection examines information obtained through network communications; and vulnerability assessment-based intrusion detection examines information obtained through network communications. It is possible to detect invasions by looking for patterns in behaviour that match the signatures of previous intrusions or vulnerabilities uncovered by an intrusion detection system that performs abuse detection. An IDS includes a pattern database, which contains every possible attack signature and is updated on a regular basis. When the IDS detects a match between the data and the attack pattern, it deems the attack to have occurred. [16]. Utilizing it is a simple process. It is not necessary for the IDS to "learn" network events in order for it to be put to use. One of the most difficult aspects of using IDS is the overhead, which can get uncomfortably large at times. When analysing system logs, the operating system must keep track of all information associated with the activities taken, which generates vast amounts of data and uses significant amounts of disc space and CPU resources. On the other hand, anomaly detection-based intrusion detection systems look for unexpected network traffic in order to detect intrusions [18]. Traditional network intrusion detection systems detect intrusions by looking for patterns of known attacks that have been previously saved. A human specialist's attack pattern is utilised to analyse network connection attributes in order to detect intrusions [19], and this information is used to detect intrusions. Network administrators may take preventative precautions if they used distributed monitoring to detect planned and coordinated attacks [20] and act quickly.



II. EASE OF USE

A. INTRUSION DETECTION SYSTEM

Intrusion Detection System-A resource's overall integrity and confidentiality can be compromised by intrusion detection. Detection of intruders is the goal of intrusion detection, which aims to locate those who are trying to break into systems and damage security safeguards. Modern IDS scan all data features to detect any intrusion or misuse tendencies, despite the fact that some parts are redundant and offer little to the detection process.

The field of intrusion detection systems has advanced significantly in the previous decade. All of IDS's tactics are based on the tried and true. An overview of recent research on IDS using a variety of approaches is presented herein.

A.M. Chandrashekhar[1] used neural networks and support vector machines to improve a hybrid intrusion detection system presented in a publication by the same name. Using fuzzy neural networks and radial support vector machines, a very effective intrusion detection system is shown in this paper. Initiating clustering, using fuzzy neural networks, creating SVM vectors, and using radial SVM classification are only a few of the methods presented.

Mohammad Sazzadul Hoque [2] provides an example of how to develop an intrusion detection system using evolutionary algorithms. This paper defines and implements an Intrusion Detection System that uses a genetic algorithm to identify a wide range of network intrusions. This benchmark dataset was used to create and test our system, and we were able to get an acceptable detection rate on it.

For intrusion detection, Rung-Ching Chen [3] used a primitive set of support vector machines, which he developed himself. This work proposes an intrusion detection technique that makes use of an SVM-based system on a RST in order to limit the number of features included in the study. In addition, the performance of the SVM was compared to that of a complete feature and Entropy. The output of our framework RST-SVM technique is more accurate than the output of either the complete feature or the entropy approaches separately. According to the results of the experiment, RST-SVM has a greater accuracy.

Adaptive intrusion detection systems were developed by Dewan Md.Farid and Nouria Harbi[4] using Navie Bayes and decision trees in conjunction with each other. One of the objectives of this research was to determine ways to increase the performance of both the nav Bayesian classifier and the ID3 approach. This hybrid technique was successfully evaluated on the five classes of the KDD99 benchmark dataset, resulting in a reduction in false positives and an increase in balance detection rates. The assaults on the KDD99 dataset were identified with 99 percent accuracy using the technique that was given in this paper. Future research will focus on minimising the number of false positives in R2L assaults as well as extending this detection model to real-world intrusion detection systems (IDS).

Ansam Khraisat's Iqbal Gondal [6] investigated intrusion detection technologies. An overview of the approaches, kinds, and technologies of intrusion detection systems, as well as the advantages and disadvantages of each, is provided in this study. For the objective of detecting zero-day vulnerabilities, we investigate a variety of machine learning techniques. False alarms or low accuracy may be the result of such systems due to difficulties in gathering and updating information on fresh assaults. In order to devise a strategy for resolving IDS difficulties, we studied prior study findings and ongoing AIDS performance improvement efforts.

According to Rohitha Gunathilake [7], statistical quality control techniques can be used in a number of settings to identify network intrusions. Assaults, detection and prevention tactics and intrusion detection systems are also examined in this essay, with a focus on the quality control measures utilised in the construction of each. There are three further phases to the process: moving average and exponentially weighted controls, and cumulative sum control charts.
P. Natesan and P. Balasubramanian [8] describe a multi-stage filter for network intrusion detection that makes use of enhanced adaboost technology. They created the Enhanced Adaboost method for detecting network infiltration, as well as rough set theory for extracting crucial properties from network infiltration. In addition to the features that were picked, the experiment is carried out with all 41 of them. By utilising the attributes offered, the rate of attack detection can be dramatically increased while the computational cost can be greatly lowered. Many intrusion detection system challenges, such as the rate at which attacks are detected, the rate at which false alarms are generated, and the processing time required are addressed in order to build a durable, scalable, and reliable intrusion detection system. It is vital to have a high detection rate and to detect attacks as quickly as possible.

For wireless adhoc sensor networks, Mohammad Saiful Islam Mamun [9] provides a hierarchical design-based intrusion detection approach based on hierarchical design. Hacking and other security concerns are a source of concern for WSN. As part of this research, we describe a novel IDS architecture for ad hoc sensor networks that is based on a hierarchical overlay design. In addition, we offer a reaction mechanism that is based on design principles. Our intrusion detection system (IDS) exceeds prior comparable solutions in terms of how it spreads the entire effort of detecting infiltration. This structure, which we propose, divides the entire responsibility of intrusion detection into four levels, which results in a structure that saves a large amount of energy. Because each monitor only needs to maintain track of a small number of nodes within its range, it consumes a little amount of energy. The detection system has been thoughtfully built.

Hybrid intrusion detection for clustered wireless sensor networks has been created, say Hicham Sedijelmaci and Mohammed Fehan [10], and it is both effective and efficient. For clustered wireless sensor networks, we have developed a distributed hybrid intrusion detection approach (HIDSs). The suggested distributed learning approach for SVM training in WSN allows for the accurate detection of both normal and aberrant behaviour. (around 98 percent accuracy rate). Signature-Based Detection (SBD) with the SVM classifier achieve high detection rates while preserving low false positive rates, as seen in the accompanying figure.

III. OPTIMIZATION

Optimization-The term "optimization" refers to the process of solving NP-hard problems. It is impossible for deterministic techniques to tackle NP-hard problems in a certain amount of time. The objective function, which can either be maximised or minimised depending on the situation, is used in optimization procedures to determine the best solution. Because we are using metaheuristic optimization, this optimization process is also referred to as metaheuristic optimization or metaheuristic optimization. Some metaheuristics or high-level approaches are used in this case. Optimization techniques have an objective function that they strive to achieve. It is possible to utilise functions that are either single-objective or multi-objective in nature. All points are pointing in the same direction. When dealing with a single-objective function, there is only one point, and that single point is the best possible answer to the question. When It is possible that two or more optimal solutions exist if particles converge at two or more points in space. The best of the best is selected in the case of a multi-objective function, on the other hand. Optimization algorithms make use of a number of different types of local search.

A. *Types of optimization*

Particle Swarm Optimization- Swarm intelligence is at the heart of the particle swarm optimization technique. Particle Swarm Optimization (PSO) is based on the coordination dynamics of groups of animals (PSO). It is true that flocks are the consequence of the application of well-chosen local regulations, which are universal and time- and place-independent. Research by Reynolds (Flocks, 1987) reveals that the movement of the entire flock is a function of individual bird behaviour that follows only three simple principles.

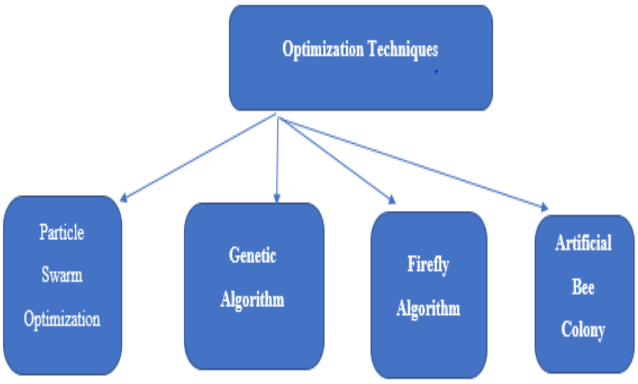
- Separation is a psychological concept that requires people to modify their physical positions in order to avoid interacting with their neighbours;
- Birds must line up with agents who are close by in order to be aligned;
- Individuals must shift toward the average position of their local flock mates in order to maintain cohesion.

Local awareness is the ability to know what's going on in the immediate region, but it cannot see beyond that area. Each bird has its own position and velocity vector. According to the PSO method, the swarm of particles (the PSO swarm), each of which represents a potential solution to an optimization problem, is tracked by its position in a multidimensional search space (the issue space) Initially, the particles are scattered around the search space in quest of the minimal value of the target function (or maximum). Experiments and estimates of well-known good search sites accumulate over time.

Genetic Algorithm- It's a problem-solving tool that's modelled after biological evolution. It uses Darwin's principle of evolution and the survival of the fittest to optimise a population of candidate solutions toward a current fitness level. Evolutionary algorithms (GAs) are built around a chromosome-like data structure that can be used to evolve genomes. An initial population of chromosomes is often generated at random and represents all possible solutions to a problem that are considered candidates for solution. Bits, characters, and numbers are used to represent different regions of each chromosome. These locations could be described using genetics. The quality of each chromosome in regard to the desired solution is calculated using an evaluation function; this function is used to determine the goodness of each chromosome in relation to the desired solution.

Firefly Algorithm- Dr. Xin's name is Firefly algorithm (FA) is a new optimization tool that she proposed. Flies are drawn to one other in the same way as humans are attracted to light. Firefly flash patterns are one-of-a-kind and vary depending on the type of insect. Insects are drawn to fireflies by their mating patterns and their prey. Female fireflies respond to the male by flashing in diverse patterns that are unique to each other throughout the mating process. Fireflies' light output is inversely proportional to their separation. This shows that the attractiveness of fireflies is based exclusively on their luminosity. As the distance between two fireflies grows, so does the intensity of the light they emit. For example, the brightness of the fireflies is correlated with the performance of a fitness function in a firefly algorithm.

Detecting intrusions in a secure information system requires an intrusion detection system (IDS). Resources that are not authorised to be accessed by intruders in the network are being sought by hackers. You must maintain track and analyse both the user's and the system's behaviour. Changing the parameters of a system can cause it to behave in a nonsensical manner. As a result, the system must have the ability to monitor both normal and anomalous activity patterns on a regular basis. There are two forms of IDS, both of which use real-time deployment and detection techniques to identify and block threats. Host-based IDS (HIDS) and network-based IDS (NIDS) are the two forms of IDS based on deployment (NIDS). To keep an eye on a computer's internal activity, HIDS does a real-time analysis of network traffic data, employing appropriate detection techniques. The three forms of IDS that are based on detection methods include misuse detection, anomaly detection, and hybrid detection systems. Misuse detection applies a set of recognised criteria or indications to detect known dangers. Algorithms that identify anomalous attacks establish an abnormal activity profile by looking for deviations from this profile. It is possible to detect both known and unexpected threats with a hybrid IDS system.



Artificial Bee Colony Algorithms- Swarm intelligence is an artificially intelligent self-organizing system for resolving and optimising complicated problems. It imitates the behaviour of ant, bee, and wasp swarms and colonies, as well as fish schools and flocks. This current optimization algorithm is based on the intelligent foraging technique of bee colonies, which Karaboga describes as a modern artificial bee colony. A food source, foragers who are employed, and foragers who are unemployed make up ABC. First, the food source is represented by a variety of factors, including how close the colony is, how good the nectar is, and how much labour it takes to collect it. As a second point, employed foragers are well-versed in the local food supplies. It's a good thing to know. Predetermined probabilities are used to disseminate the information. Forager bee Scout and spectator bees are two types of unemployed foragers who are always looking for new food sources to feed on. In contrast to observation bees, which remain within the colony and rely on information provided by the colony's paid foragers to find new food sources, scout bees forage in the area surrounding the colony.

CONCLUSION

In this paper, we present and create an Intrusion Detection System that use different optimization algorithms to swiftly identify various types of network intrusions. The standard dataset served as the basis for developing and testing our system, and we were able to achieve a decent detection rate. In order to determine a chromosome's fitness, a standard deviation equation involving distance was used. To improve the detection rate and procedure, as well as the rate of false positives, we feel that a better equation or heuristic should be used in this detection process. The use of more statistical analysis and maybe more complicated equations is something we want to do in the near future to further improve our intrusion detection system.

REFERENCES

[1] A. M. Chandrashekhar1 and K. Raghuveer2 FORTIFICATION OF HYBRID INTRUSION DETECTION SYSTEM USING VARIANTS OF NEURAL NETWORKS AND SUPPORT VECTOR MACHINES International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.1, January 2013.

[2] Mohammad Sazzadul Hoque1 , Md. Abdul Mukit2 and Md. Abu Naser Bikas AN IMPLEMENTATION OF INTRUSION DETECTION SYSTEM USING GENETIC ALGORITHM International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

[3] Rung-Ching Chen 1, Kai-Fan Cheng 2 and Chia-Fen Hsieh USING ROUGH SET AND SUPPORT VECTOR MACHINE FOR NETWORK INTRUSION DETECTION International Journal of Network Security & Its Applications (IJNSA),Vol 1, No 1, April 2009.

[4] Dewan Md. Farid1 , Nouria Harbi1 , and Mohammad Zahidur Rahman COMBINING NAIVE BAYES AND DECISION TREE FOR ADAPTIVE INTRUSION DETECTION International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.

[5] Intelligent Intrusion Detection System Using Clustered Self Organized Map Member, IEEE, Alia Abu Ghazleh+, Member, IEEE, Amer Al-Rahayfeh† , Member, IEEE, Abdul Razaque‡, Member, IEEE 2018 Fifth International Conference on Software Defined Systems (SDS).

[6] Ansam Khraisat*, Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman Survey of intrusion detection systems: techniques, datasets and challenges.

[7] Rohitha Goonatilake1 , Rafic Bachnak1 , and Susantha Herath STATISTICAL QUALITY CONTROL APPROACHES TO NETWORK INTRUSION DETECTION International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011

[8] P. Natesan1 , P.BalasubramanieMulti Stage Filter Using Enhanced Adaboost for Network Intrusion Detection International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.3, May 2012

[9] Mohammad Saiful Islam Mamun HIERARCHICAL DESIGN BASED INTRUSION DETECTION SYSTEM FOR WIRELESS AD HOC SENSOR NETWORK International Journal of Network Security & Its Applications (IJNSA), Vol.2, No.3, July 2010.

[10] Hicham Sedjelmaci1 and Mohamed Feham1NOVEL HYBRID INTRUSION DETECTION SYSTEM FOR CLUSTERED WIRELESS SENSOR NETWORK International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.4, July 2011

[11] Ghanshyam Prasad Dubey, Prof. Neetesh Gupta and Rakesh K Bhujade, “A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM”, International Journal of Soft Computing and Engineering (IJSCE), vol.1, no.1, pp.14-18, 2011.

[12] Iftikhar Ahmad, Azween Abdullah and Abdullah Alghamdi, (2010) “Towards the selection of best neural network system for intrusion detection”, International Journal of the Physical Sciences, vol.5, no.2, pp.1830-1839.

[13] J.T. Yao, S.L. Zhao, L. V. Saxton (2005), “A study on fuzzy intrusion detection”, Proc. of Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security SPIE. 5812, pp. 23-30.

[14] Hansung Lee, Jiyoung Song, and Daihee Park (2005), “Intrusion Detection System Based on Multi-class SVM”, Dept. of computer & Information Science, Korea Univ., Korea, pp. 511–519.

[15] David Wagner and Paolo Soto (2002), “Mimicry Attacks on Host Based Intrusion Detection Systems” Proceedings of the 9th ACM conference on Computer and communications security, pp. 255 – 264.

[16] Rung-Ching Chen and Su-Ping Chen (2008), “Intrusion detection using a hybrid support vector machine based on entropy and tf-idf”, International Journal of Innovative Computing, Information and Control, vol. 4, no. 2, pp. 41301-424.

[17] Srinivas Mockamole, Andrew H. Sung, Ajith Abraham and Vitorino Ramos (2004), “Intrusion detection systems using adaptive regression splines”, In Proceedings of the 6th International Conference on Enterprise Information Systems, ICEIS, vol.3, pp.26-33.

[18] Snehal A. Mulay, P.R. Devale and G.v. Garje (2010), “Intrusion Detection System using Support Vector Machine and Decision Tree”, International Journal of Computer Applications, vol 3, no 3, pp. 40-43.

[19] Deepak Tinguriya and Binod Kumar (2010), “A Effect Approach for Intrusion Detection System using Incremental SVM”, BLB-International Journal of Science & Technology, vol.1, no.2, pp.127-134.

[20] Ajith Abraham, Ravi Jainb, Johnson Thomas and Sang Yong Han (2007), “D-SCIDS: Distributed soft computing intrusion detection system”, Journal of Network and Computer Applications, vol. 30, pp. 81–98.

ARTIFICIAL INTELLIGENCE-BASED INTRUSION DETECTION SYSTEM FOR CLOUD COMPUTING

Srijan Mishra, Dr.Shashank Singh(Assistant Professor)
Advanced Computing and Data Science Department
Integral University, Lucknow
srijanankit@gmail.com , *shashank@iul.ac.in*

Abstract. – It is possible to communicate with others and do business across this global network, which is comprised of hundreds of millions of computers running a variety of hardware and software configurations. This makes it simpler for hackers to abuse resources and conduct Internet attacks since computers are linked to one another. There are significant roadblocks to the development of a security-oriented approach that may be flexible and adaptive in light of the expanding number of Internet assaults. An intrusion detection system is necessary for the identification of online threats (IDS). Maintaining the security of a network requires the use of an intrusion detection system or IDS. Because the cloud platform is continuously expanding and becoming more prevalent in our everyday lives, it is imperative that we develop an effective IDS for it as well. On the other hand, typical intrusion detection systems may encounter difficulties when used in the cloud. A cloud segment may get overburdened by the pre-determined IDS architecture because of the added detection overhead. In the context of a networked system with an adaptable architecture. Using a neural network-based IDS, we show how to make full utilize available resources while not putting undue strain on any single cloud server. The proposed IDS uses a neural network machine learning to identify new threats even more effectively.

Keywords: – Artificial Intelligence, Intrusion Detection System, Machine learning Algorithm.

I . Introduction

Today, the Internet is an essential part of our everyday lives, and it&#39; utilized in everything from commerce to entertainment to education. It has become increasingly common for businesses to use the Internet to get access to information. A computer system may be hacked in a variety of ways because of the wide availability of information on the Internet. Internet assaults and incursions are becoming increasingly commonplace. An incursion or assault may be described as “any combination of actions that seek to breach the security objectives”. Some of the most important security objectives are availability, integrity, confidentiality, accountability, and assurance. Assaults may be divided into four categories: Probing, Denial of Service, User to Root, and Remote User. Several anti-intrusion technologies have been developed to stop a huge proportion of Internet assaults. Six anti-intrusion systems have been described by Halma and Bauer (1995), and they include IDS of them. The other five include detection and countermeasures. The most critical of these components is the ability to identify an incursion perfectly.

2. Related Work –

Using machine algorithms in a big data environment, Saud Mohammed Othman and Fadl Mutaher Ba-Kiwi [1] presented their work on an intrusion detection model. According to this study, Big Data&#39;s ever-increasing volume has shifted the significance placed on data security and analytic technologies. An IDS monitors and analyses data in order to discover any system or network intrusions that may have occurred. Traditional strategies for detecting network assaults have become increasingly complex due to the volume, diversity, and speed of data created in the network. IDS uses Big Data approaches to analyze Big Data in an accurate and effective manner.

By utilizing the Spark-Chi-SVM architecture, they were able to construct an intrusion detection model that is capable of managing enormous amounts of data. The Spark Big Data platform was utilized in the recommended approach in order to facilitate the handling and analysis of data in a timely manner. The large dimensionality of big data adds to the difficulty and length of the categorization process. A Survey of Intrusion Detection Systems utilizing Machine Learning Techniques was provided by Sharmila Wagh and Vinod K. Pachghare [3]. Computers and network-based technologies are becoming more and more commonplace in today&#39;s environment, according to the authors. The importance of network security cannot be overstated in today&#39; s computer age. Detection of system assaults and classification of system activity into normal and abnormal forms are the goals of an Intrusion Detection System (IDS). IDS systems that use machine learning to identify intrusions have become increasingly commonplace.

This paper [4] describes a Cloud-based distributed machine learning-based intrusion detection system. Edge network components from Cloud providers are to accompany the proposed system in the Cloud. Incoming network communication may be intercepted by the edge network routers on the physical layer as a result. Before being transmitted to a module that employs the NaiveBayesclassifierr to identify anomalies, the network data collected by each Cloud router is preprocessed using a time-based sliding window technique. When there is a buildup of network congestion, server nodes that are powered by Hadoop and MapReduce are made accessible to each anomaly detection module. A server for synchronizing anomalous network traffic data is assigned to each time frame. This server collects anomalous network traffic data from each router in the system. Following this, Random Forest classifiers are applied to each attack in order to establish the kind of assault that was committed.

Hassan Musafer and Ali Alessa&#39; s study, titled &quot; Machine Learning-Based Network Intrusion Detection Detection: Dimensionality Reduction Approaches, &quot; was recently published. [5]. They claim that all parties, including consumers, businesses, and governments, are worried about the safety of computer networks. As it becomes increasingly difficult to safeguard networked systems against assaults, the strategies that attackers use to carry out such assaults also evolve. A portion of the solution is in making the existing intrusion detection systems more effective. The use of machine learning to create intrusion detection systems is an approach that is becoming more popular because of its efficiency (IDS). When improvements are made to IDS qualities like as discrimination and representation, there is a considerable increase in the system&#39; s overall performance. Through the use of Deep Learning&#39; s Auto-Encoder (AE) and Principal Component

Using Principal Components Analysis (PCA), the dimensionality of the features were reduced for the purpose of this inquiry (PCA). It is possible that the combination of these two approaches will result in the acquisition of low-dimensional attributes that can be utilised in the building of classifiers such as Random Forest (RF), Bayesian Network, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).

Intrusion Detection System-

One of the fundamental components of security infrastructure is an effective security system that can detect, prevent, and maybe respond to computer threats. FoSecurityervices, it uses a variety of methods, including audit and network traffic data in computer or network systems, to monitor target sources of activity. All threats must be swiftly and correctly identified by a intrusion detection system (IDS). IDS may help network managers find security flaws objectively. Intruders from the outside may attempt to get access to the network illegally. security infrastructure or make resources inaccessible to insiders who abuse their system resources, all of which are breaches of security objectives.

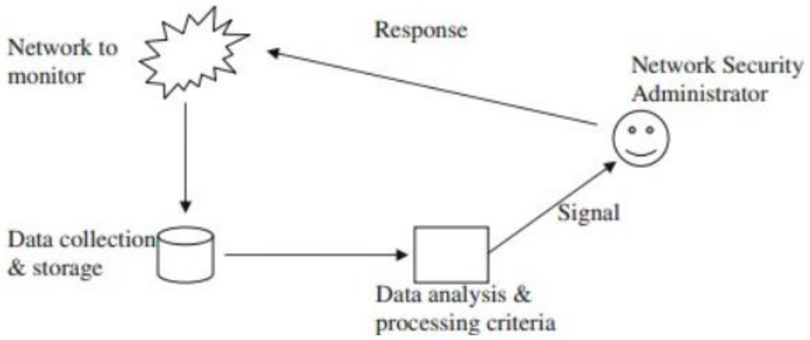

Network to monitor
Response
Network Security Administrator
Signal
Data collection & storage
Data analysis & processing criteria

Fig.1. Architecture of Intrusion Detection System

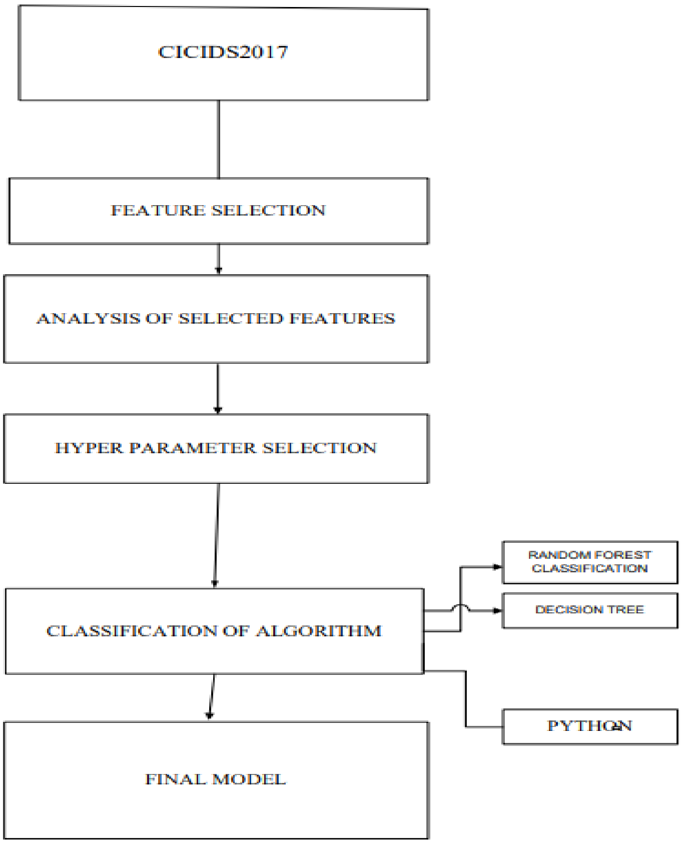
As the number of computer attacks has risen, many IDS designs have been proposed. Axelson (1999) suggested a common design for IDS in Figure 1. The following are the most common IDS components, according to Axelson (1999): The network to be monitored must be identified in order to keep tabs on any unauthorized intrusions. Alternatively, the whole network might be utilized for this. A disc-based event data collection and storage device is in charge of collecting and storing data from various sources. The IDS&#39;s data processing and analysis unit is its brain. All the functions essential for detecting aberrant traffic patterns are included in this device. A signal is generated when an attack is detected. When an IDS detects a problem that needs to be addressed, the system may either take action on its own or provide a signal that has to be addressed to the network administrator. This program&#39;s activities might result in an automated response to an intrusion or a notice of malicious behavior for a network security administrator. Modules of an IDS may be categorized in a variety of ways. Based on how data is gathered and stored, IDS may be divided into two categories:

Host-based IDSs are those that gather data from a host. Other sources of data include system calls and operating system logs such as the NT events and CPU utilization logs as well as application logs. Buffer overflow attacks are easily detected by host-based

IDS since they are not reliant on the operating system. A switched network and encrypted data render these methods ineffectual. It is called a network-based IDS if the IDS collects data in the form of packets from the network. These IDS are cross-platform compatible and may be set up on almost any system.

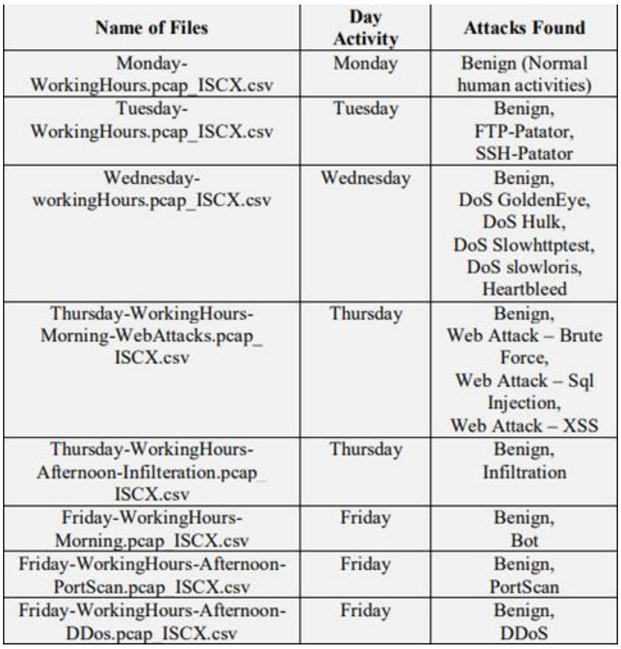
3. Descriptions of CICIDS2017 dataset –

As much as 98 percent+ accuracy and less than 1 percent false alarms are already claimed by researchers in the field of intrusion detection. Researchers and manufacturers were compelled to devote money and effort to the development of useful goods because of this high rrate of accuracyIn reality, only a few models have been recognized the industry to design an IDS. By analanalyzingtemporary IDS models and training and testing datasets, the dataset not only includes the most recent network assaults, but it also meets all of the criteria for attacks that really occur in the real world. We noticed just a few flaws in this dataset when we investigated its properties. An obvious flaw is a large dataset, which was compiled from five days' worth of Canadian Institute of Cybersecurity traffic data spread over eight files. An IDS might be designed from a single dataset. There are a lot of redundant entries in the dataset, making it unsuitable for training any IDS. Even if the dataset comprises contemporary assault scenarios, we also discovered that the dataset has a substantial class imbalance. Class imbalance datasets can mislead the classifier, biassing it towards the majority class. " The research community was given a subset of the CICIDS2017 dataset to work with in developing and testing detection algorithms in an attempt to address these issues. Fig below shows the description of the CICIDS2017 dataset on which I have worked. we are able to identify the root cause of this issue. The Canadian Institute of Cybersecurity&#39;s CICIDS2017[7] collection provides the most up-to-date attack scenarios.


CICIDS2017
FEATURE SELECTION
ANALYSIS OF SELECTED FEATURES
HYPER PARAMETER SELECTION
CLASSIFICATION OF ALGORITHM
RANDOM FOREST CLASSIFICATION
DECISION TREE
PYTHON
FINAL MODEL

3. Shortcomings of the CICIDS2017- CICIDS2017 dataset has several flaws, as we previously noted, and the purpose of this work is to remedy those weaknesses so that futureresearchers may better understand them.

Scattered Presence- We can see in table 1 that there are now eight files containing the CICIDS2017 dataset's data. It's time-consuming to deal with individual files. As a result, we created a single file that had 3119345 instances of each of the files in question.



Name of Files	Day Activity	Attacks Found
Monday-WorkingHours.pcap_ISCX.csv	Monday	Benign (Normal human activities)
Tuesday-WorkingHours.pcap_ISCX.csv	Tuesday	Benign, FTP-Patator, SSH-Patator
Wednesday-workingHours.pcap_ISCX.csv	Wednesday	Benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Thursday-WorkingHours-Morning-WebAttacks.pcap_ ISCX.csv	Thursday	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ ISCX.csv	Thursday	Benign, Infiltration
Friday-WorkingHours-Morning.pcap_ISCX.csv	Friday	Benign, Bot
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Friday	Benign, PortScan
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Friday	Benign, DDoS

Table 1. Description of the file
containing the CICIDS2017 dataset

3.1Huge Volume of Data - All of the potential recent assault labels may be found in one location after integrating all of the data files. However, the combined dataset grows enormously in size. The sheer amount of information available becomes a problem in and of itself. It has a drawback in that it takes more time to load and analyse data.

3.2 Missing Values - The combined CICIDS2017 dataset has 288602 cases with a missing class label and 203 instances with missing metadata, as well. We found this to be a problem. To create a dataset with 2830540 unique occurrences, all but a few of the original data points were eliminated**.**

Classification of Algorithm

According to the findings of this investigation, the performance of a classifier algorithm in terms of accuracy, learning capacity, scalability, and speed are the most essential factors to take into consideration when making a selection. Research and findings have been proven to support this concept using a total of five different classification algorithms, including Random Forests, Bayesian Network, Random Trees, Naive Bayes, and J48 classifiers. This study demonstrates, through the application of the Information Gain feature selection, that random forest trees are capable of learning and performing admirably when it comes to the detection of assaults. The Bayesian Network surpasses other algorithms when it comes to categorizing assaults. Random Tree is a scalable and efficient method. Since Naive Bayes has a low model complexity, it is a better choice for classifying data than other algorithms.

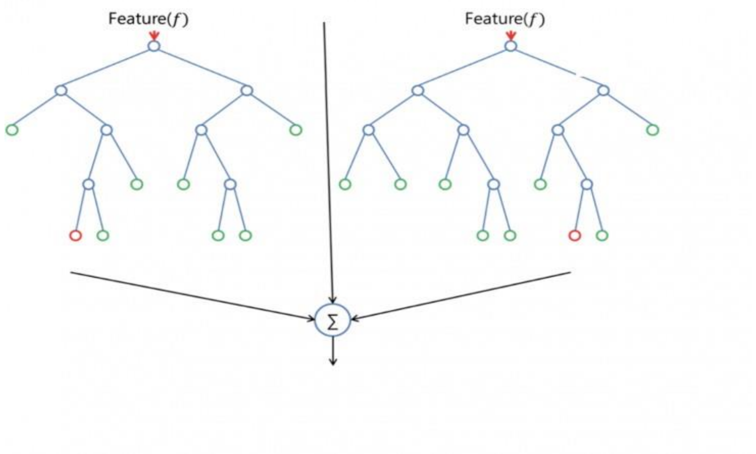


Fig:- Random Forest classifiers tree.

1) Random Forest (RF)

Ensemble classifier approaches include Random Forest. A "forest" of classifiers is a decision tree classifier ensemble. To generate each decision tree, qualities are randomly selected at each node. In 2001, Breich introduced the random forest algorithm.

2) Bayes Network (BN)

Probabilistic connections between variables of interest are encoded using a Bayesian Network (BN) modeling technique. Assumptions about the model behavior of the target system are used to determine how accurate this technique is. If the assumption is significantly altered, then detection accuracy is reduced.

3) Random Tree (RT)

The term "random tree" refers to a decision tree that is built using a random set of attributes (random). There are many nodes and branches that can be linked together in a variety of ways in a decision tree.A node is used to represent an attribute being tested, while branches are used to indicate the findings. In the form of class albethey e, decision leaves display the final choice made after the computation of all attributes.

4) Naive Bayes (NB)

According to the Bayesian categorization system, the likelihood of belonging to a given class may be predicted statistically. Based on the Bayes theorem, we may classify data in a Bayesian fashion. Like the Nave Bayes classification, the Bayesian classification is better recognized by its more formal name. Ignoring other attribute values, Nave Bayes considers that attribute values have no effect on the class they belong to

5) J48

Machine learning algorithm J48 or C4.5 is frequently used and is part of the decision tree algorithm. The entropy idea is used to form a decision tree in this technique.

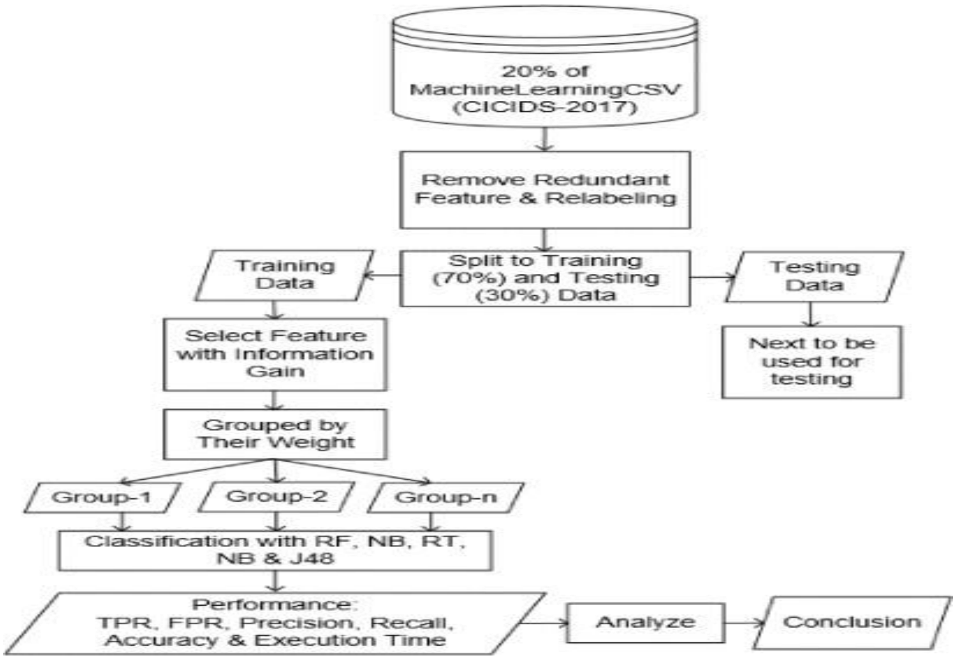

20% of
MachineLearningCSV
(CICIDS-2017)
Remove Redundant Feature & Relabeling
Training Data
Split to Training (70%) and Testing (30%) Data
Testing Data
Select Feature with Information Gain
Next to be used for testing
Grouped by Their Weight
Group-1
Group-2
Group-n
Classification with RF, NB, RT, NB & J48
Performance: TPR, FPR, Precision, Recall, Accuracy & Execution Time
Analyze
Conclusion

Fig:- Experimental design

- For the purpose of classifying each feature group or feature subset, respectively, the Random Forest (RF), Bayes Net (BN), Random Tree (RT), Naive Bayes (NB), and J48 classifiers are utilised. In order to carry out the analysis, the following considerations are taken into account: the True Positive Rate, the False Positive Rate, the Precision Rate, the Recall Rate, the Accuracy Rate, the Proportion of Incorrectly Categorized Data, and the Execution Time of the Analysis. In addition, the True Positive Rate is compared to the False Positive Rate. In addition, the True Positive Rate and the False Positive Rate are contrasted with one another. At this stage of the process, a methodology known as 10-fold cross-validation is applied.
- Analysing and comparing the TPR, FPR, Precision and Recall, Accuracy, Percentage of Incorrect Categorization, and Execution Time of each classifier algorithm is an imperative requirement.At each and every level of the learning and testing process, a ten-fold cross-validation is conducted. It is essential that you arrive at some inferences or conclusions at this juncture.

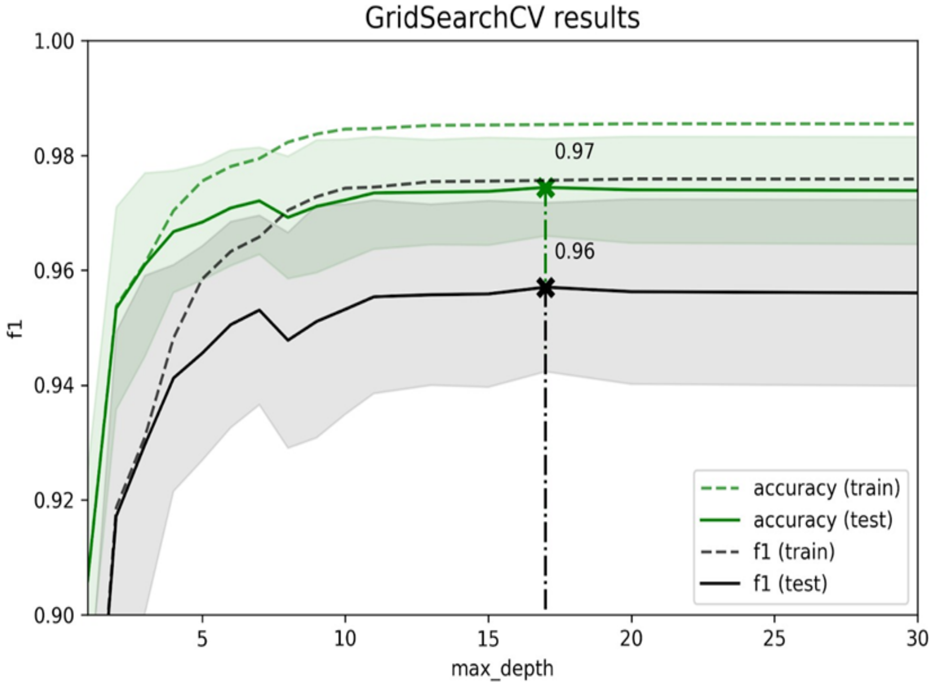


Fig:- Accuracy graph

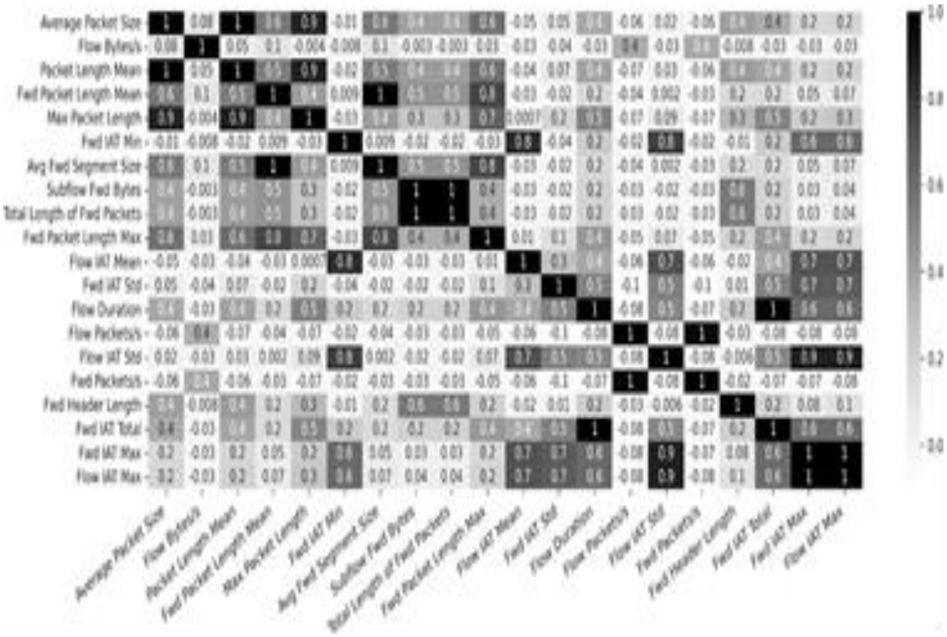


Fig:- Correlated Heat Map

. **Experimental Result**

In addition to the five distinct classifier approaches, the True Positive Rate (TPR), the False Positive Rate (FPR), the Precision, the Recall, the Accuracy, the Percentage of incorrectly Classified, and the Execution Time are the metrics that are utilised to evaluate the efficacy of Information Gain. These measurements make up what is known as the metric set when considered as a whole. The actual execution is simulated at a number of different periods in time throughout the course of the training, with the goal of further refining and improving upon it. During this experiment, the RT, BN, RT, NB, and J48 classifiers are combined in a wide variety of different ways to classify each unique feature subset. RT stands for the random tree; BN stands for the random tree; RT, NB, and J48 stand for the random tree; and RT. A 10-fold cross-validation method was used during the course of this investigation in order to ascertain whether or not categorization algorithms are successful. The 10-fold cross-validation is utilized since it cuts down on the overall amount of time spent calculating while also maintaining the reliability of the classification strategies. As a direct and immediate result of this, 10 random folds of the input dataset of the same size will be constructed from it. During the process of cross-validation, nine of the ten-fold data sets will be employed for training, while just one of the ten-fold data sets will be utilized for testing. The ultimate outcome is a test fold, which is produced after ten repetitions of this procedure.

Overall Process

1) Step-by-step procedure ()
2) Feature Ranked data are accepted as input.

3) The output includes the following: features subsets, TPR, FPR, accuracy, recall, and precision
4) Decrease the number of features from 77 to n depending on the feature weight.
5) For each and every function Fr in the data for Feature Ranked
6) 6) Begin to choose features using the Feature Weight, and then save them on Feature Groups

Group1 is comprised of any characteristic that has a weight more than or equal to 0.6 Any characteristics that have a weight that is more than or equal to 0.5 are included in Group2. Group3 is comprised of all characteristics with weights more than or equal to 0.4 Group4 is comprised of any feature that has a weight more than or equal to 0.3 Group5 is comprised of all features that have a weight greater than or equal to 0.2. Group6 is comprised of all features that have a weight greater than or equal to 0.1. Group7 denotes the whole of the characteristics.

7) Regarding each of the Feature groupings

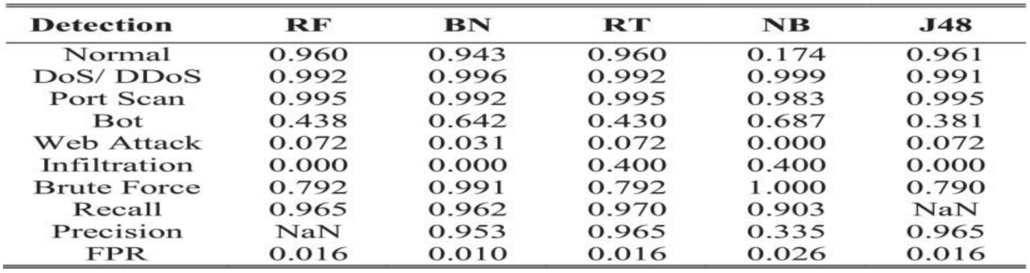
8) Using CICIDS-2017-20 percent, provide selected features to RF, BN, RT, and NB, as well as J48.

9) Apply Classifier Accuracy of the Random Forest model, denoted as C1 C2 equals the accuracy of the Bayes Network model C3 equals the accuracy of the Random Tree model C4 = Naïve Bayes model accuracy C5 = J48 model correctness

10) Accuracy, recall, and precision of TPR and FPR calculations have to be determined**.**

11) Examine the Accuracy of C1, C2, C3, C4, and C5 and Compare Them.

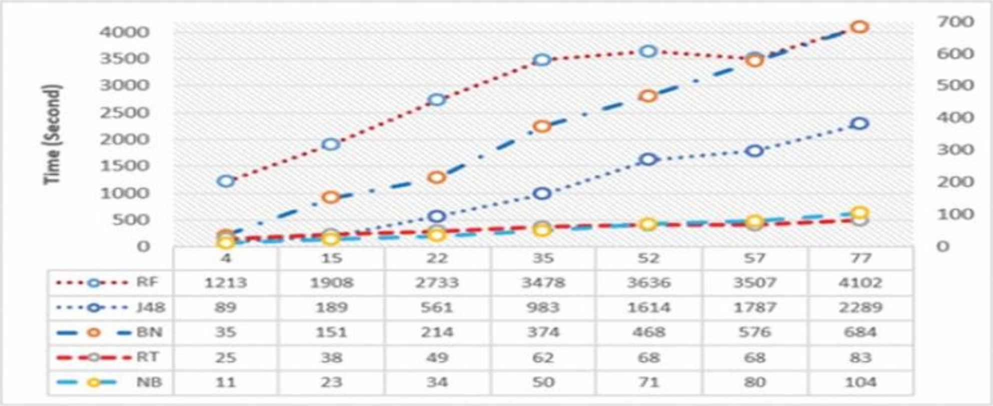
Classifiers that use four characteristics chosen by Information Gain are mentioned below. Other classifiers are outperformed only by the RF and RT, which have an accuracy of 96.48 percent. RF, on the other hand, has a value of NaN. The term "NaN" stands for "Not a Number" or "undefined," respectively. NB has a higher TPR for identifying DoS/DDoS attacks than other classifiers, but a lower TPR for recognizing normal and infiltration traffic. Comparatively, BN has the lowest FPR (0.010) of all of the companies studied. Classifiers can only identify DoS/DDoS, PortScan, and Brute Force assaults using these four (4) characteristics. Only NB is affected by this in terms of normal traffic.



Detection	RF	BN	RT	NB	J48
Normal	0.960	0.943	0.960	0.174	0.961
DoS/ DDoS	0.992	0.996	0.992	0.999	0.991
Port Scan	0.995	0.992	0.995	0.983	0.995
Bot	0.438	0.642	0.430	0.687	0.381
Web Attack	0.072	0.031	0.072	0.000	0.072
Infiltration	0.000	0.000	0.400	0.400	0.000
Brute Force	0.792	0.991	0.792	1.000	0.790
Recall	0.965	0.962	0.970	0.903	NaN
Precision	NaN	0.953	0.965	0.335	0.965
FPR	0.016	0.010	0.016	0.026	0.016

Fig: - Performance Metric Using Four Features

For the particular feature method under consideration, this study additionally examines the impact of execution time. An overview of the execution time for each feature subset employing RF, J48, BN RT and NB can be seen in the image below. There is a major influence on the RF, J48, and BN of the pertinent features procedure. RT and NB have extremely short run times. As a rule of thumb, the more characteristics to evaluate, the more time it takes to complete.



	4	15	22	35	52	57	77
RF	1213	1908	2733	3478	3636	3507	4102
J48	89	189	561	983	1614	1787	2289
BN	35	151	214	374	468	576	684
RT	25	38	49	62	68	68	83
NB	11	23	34	50	71	80	104

Fig: - Execution time

Conclusions & Future Work

Experiments were undertaken to demonstrate the impact of feature selection on improving anomaly detection accuracy. Feature sets 15, 22, and 35 were tested, and Information Gain was shown to be the top information classifier due to its accuracy in determining how much data is contained in each feature set. Feature sets 52, 57, and 77, on the other hand, are the best for J48. It is possible to detect all communication using feature sets 52, 57, and 77.5, even though BN's precision is lower than RF and J48, despite its lower level of accuracy. Experiments have also shown that the chosen traits reduce FPR, particularly for BN. Experimental results show that the number of features picked has an impact on the execution time of a program. Ranking characteristics according to weight values is what Information Gain proposes to do. It is, nevertheless, necessary for an expert to decide the minimum weight value, which influences the number of characteristics that will be picked. We intend to experiment with a variety of feature selection strategies in order to come up with the best possible mechanism. Each feature subset that impacts an assault will be analyzed as part of future research.

References

[1.] Othman, Suad Mohammed, et al. "Intrusion detection model using machine learning algorithm on Big Data environment." Journal of Big Data 5.1 (2018): 34.

[2.] Salo, Fadi, Ali Bou Nassif, and Aleksander Essex. "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection." Computer Networks 148 (2019): 164-175.

[**3**.] Wagh, SharmilaKishor, Vinod K. Pachghare, and Satish R. Kolhe. "Survey on intrusion detection system using machine learning techniques." International Journal of Computer Applications 78.16 (2013). Chiba, Z., Abghour, N., Moussaid, K., El Omri, A., &Rida, M. (2019, June). An Efficient Network IDS for Cloud Environments Based on a combination of Deep Learning and an Optimized Self-adaptive Heuristic Search Algorithm. In International Conference on Networked Systems (pp. 235-249). Springer, Cham.

[4.] Idhammad, Mohamed, Karim Adel, and Mustapha Belouch. "Distributed intrusion detection system for cloud environments based on data mining techniques." Procedia Computer Science 127 (2018): 35-41.

[5.] Abdulhammed, Razan, et al. "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection." Electronics 8.3 (2019): 322.

[6.] Basaveswara Rao B &Swathi K (2016) Variance-Index Based Feature Selection Algorithm for Network Intrusion Detection, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278- 8727, Volume 18, Issue 4, Ver. V (Jul.-Aug. 2016), PP 01-11.

[7.] Basaveswara Rao B &Swathi, K. (2017). Fast kNN Classifiers for Network Intrusion Detection System. Indian Journal of Science and Technology, 10(14).

[8.] Swathi, Kailas am, and BobbaBasaveswara Rao. "Impact of PDS Based kNN Classifiers on Kyoto Dataset." International Journal of Rough Sets and Data Analysis (IJRSDA) 6.2 (2019): 61-72.

[9.] Ali, Mohammed Hasan, and Mohamad FadliZolkipli. "IntrusionDetection System Based on Fast Learning Network in Cloud Computing." Advanced Science Letters 24.10 (2018): 7360-7363.

[10.] Ahmed, H. A. S., Ali, M. H., Kadhum, L. M., Zolkipli, M. F., &Alsariera, Y. A. (2017). A review of challenges and security risks of cloud computing. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(1-2), 87-91.

[11.] Ali, Mohammed Hasan, et al. "A hybrid Particle swarm optimizationExtreme Learning Machine approach for Intrusion Detection System." 2018 IEEE Student Conference on Research and Development (SCOReD). IEEE, 2018.

[12.] Umer, Muhammad Fahad, Muhammad Sher, and Yaxin Bi. "Flow-based intrusion detection: Techniques and challenges." Computers & Security 70 (2017): 238-254.

[13.] “Nsl-kdd data set for network-based intrusion detection systems.” Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March 2009.

[14.] Kyoto 2006+ dataset is available on: http://www.takakura.com/Kyoto_data/

[15.] https://www.unb.ca/cic/datasets/ids-2017.html

[16.] Basaveswara Rao B, et al., “A Fast KNN Based Intrusion Detection System for Cloud Environment”, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, Issue 7, 2018 PP 1509 -1515.

[17.] Ring, Markus, Sarah Wunderlich, DenizScheuring, Dieter Landes, and Andreas Hotho. "A Survey of Network-based Intrusion Detection Data Sets." Computers & Security (2019). 18. Ahmim, A., Maglaras, L., Ferrag, M