A NOVEL INTRUSION DETECTION TECHNIQUE USING BOOSTING APPROACH ON BIG DATA

Mukesh Choudhary 1 Dr. Manish Shrimali 2
1 Research Scholar 2 Professor Department of CS & IT
[1&2] JRNRVU Udaipur ,Rajasthan, INDIA.

Abstract- The main objective of intrusion detection systems (IDS) is to discover the dynamic and malicious form of network traffic that simply changes according to the characteristics of the network. The IDS methodology represents a prominent developing area in the field of computer network technology and its security. A different form of IDS has been developed working on distinctive approaches. One such kind of approach where it is used is the machine learning mechanism. In the proposed methodology an experiment is applied to the data-set named as KDD-99 including its subclasses such as a denial of service (DOS), other types of attacks and the class without any form of attack. Depending upon the machine learning algorithms various distinct forms of IDS have been developed which further checks the optimization based potential features in connection with the neural network classifier for the various forms of IDS based attacks. This approach provides a comparative study between the ANN and the optimizer-based ANN technology. The experimental analysis shows the convolution neural network with SVM show effective analysis providing accurate forms of IDS thereby improving its detection based on individual class along with maintaining its results fundamentally.

Keywords- Intrusion Detection Systems, Denial of service

I. INTRODUCTION

In the present scenario, the use of the internet is growing at a large pace with is highly developed and emerging forms of ever-growing network and its connectivity but the use of internet poses a great threat to cybersecurity. In order to maintain a high level of security, there is an important need to overcome the cyber threats posing problems to various organizations, companies, and firms. One of the major challenges among the cyber-security is to maintain the integrity of the intrusion detection system (IDS) thereby protecting it from major forms of attacks and to conquer the various form of risks of the intruding system [3][16]. The main function of the IDS is to identify a more precise form of intrusion. The illegal hackers of the security have found a large number of ways to break the security of the system whether it is a cloud network or the wireless-based network [6]. Many types of research have been performed by the technologists to curb the security threats from distinct forms of intrusions done to the cloud computing systems and the wireless system.

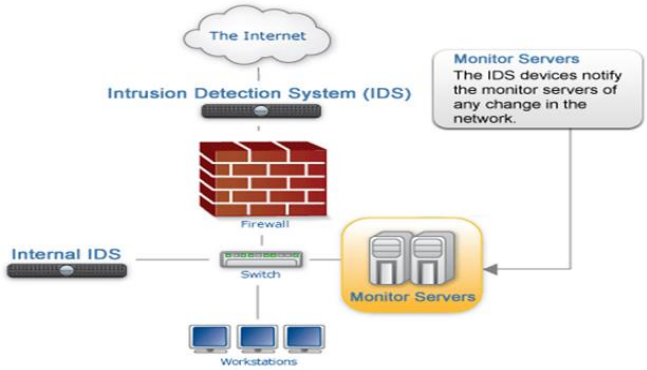

Figure 1: Sample IDS

So, the main objective of IDS is to protect the information whether it is governmental, public or private entity [10]. The use of IDS is mainly required in detecting the false and the poor detection rates. Whenever an attack is observed by the system or a harmful activity is done to the system, it automatically generates an alarm resulting in a false-positive alarm [3]. The research mainly focusses upon the enhanced capabilities of the intrusion detecting system and thereby reduces the occurrence of the false type alarms. The main requirement of the IDS is not only to encounter the intruders in the data path but also to supervise the intruders of the data. The most important security aspects of an intrusion detection system consist of maintaining the following conditions.

- *Confidentiality:* Only an authorized user can detect the system.
- *Availability:* Here, computer technology provides various forms of resources and access to the legal users of the system without disturbing the working operation of the system.
- *Integrity:* The information must be protected from any kind of malicious activity.

1.1 IDS: Architecture

The architecture of IDS comprises of its unique core element i.e. sensor popularly known as the analysing engine to pinpoint the intrusions occurring in the system. The sensor consists of a mechanism that helps in detecting the intrusions. In the following figure.1, the sensor gets the data (raw) from the given sources as shown which consists of the audit trails, knowledge-based data and, syslog. The 'syslog' includes the authority to the particular system or the system file configuration [1][14].

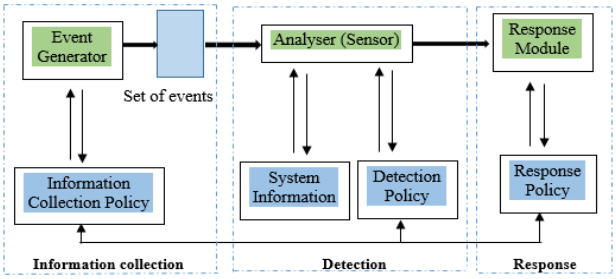

Figure 2: IDS Components

The sensor consists of a component known as an event generator which performs the data collection shown in figure above. It detects the way of collecting the data. The event generator consists of network, operating system, and the network applications where it generates a set of events including audit (log) of the system or the packets of the network. This form of set events also involves the policy of information collection i.e. in or out of the system. Sometimes it is not necessary to store the data as it reaches simply to the analyser. So, basically, the key role of the sensor is to extract or filter the data and remove the unwanted form of the data that is achieved from the event data set system [6, 7]. Additionally, the database holds the configurational parameters of IDS that includes its mode of communication methods based on the response module. The sensor itself contains its own data observing all the historical multiplex forms of intrusions [8].

1.2 Types of IDS

There are three types of IDS

1. Host-based IDSs: Host-based intrusion detection system (HIDS) also supervise the information flow and the attacks over the system is identified on the basis of network events. In host-based intrusion detection functions system events are supervised. In a host-based system, the Intrusion Detection System (IDS) suggest on each individual computer or host at the activity (Marcus A. Maloof, 2006).In this perspective on Individual systems, IDS is applied.

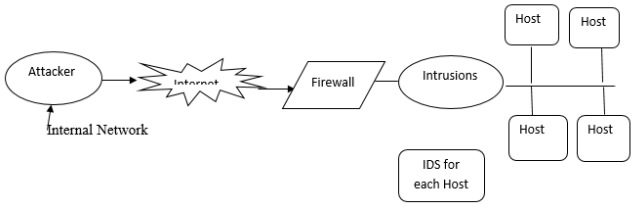

Figure 3: Host based IDS

2. Network-Based IDSs: Network-based intrusion detection system (NIDS) supervise is used for the detection of intrusions and the information flowing is being audited over the internet network. In a network-based system or NIDS, the single packets flowing through a network are investigated. In this technique, NIDS is applied first and then Firewall so that it can describe all the data packets flowing through the network. NIDS can supervise multiple structures at a time due to its advantage of supervising complete network in a single time [11].

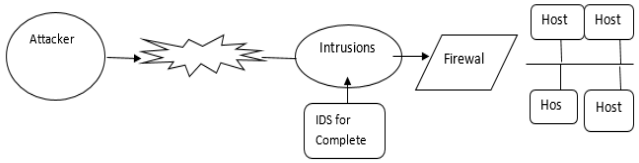

Figure 4: Network based IDS

3. Distributed IDSs

Gather audit data from multiple hosts and possibly the network that connects the hosts

Detect attacks involving multiple hosts

1.3 Intrusion Detection Approaches

The two main categories through which network can be analysed for the detection of intrusion are [9, 10]:

1. Misuse detection: Misuse detection is a perspective where the detection of intrusions on the basis of paradigm matching. Here the abnormal structure behaviour is defined at first by collecting the paradigm of attack, and any other behaviour is defined as normal behaviour by matching them opposing the already recorded attacks. In short, anything we don't know is normal. Using attack signatures in IDS is an example of this advancement. Signature-based systems can only detect and identified, prior been established. It stands opposing deviation apprehension access which employs the reverse perspective, defining normal system behaviour and defining any other behaviour as abnormal. The disadvantage of this perspective is that intrusion detection is accurate only for the known attacks. We needed to update the database of attacks to recognizing the new unseen attacks.

2. Anomaly detection: An Anomaly-Based Intrusion Detection System is a structure for ascertaining computer intrusions and misuse by supervising system activity and categorize as normal or anomalous. The categorization is based on rules, rather than paradigm or signatures, and will any type of misuse is identified that falls out of normal system function [13].

1.4 Types of IDS attacks

1. Passive attack: It is in nature of eavesdropping on or monitoring of transmission. Passive attacks include traffic analysis, monitoring of unprotected communications, decrypting weakly encrypted traffic, and capturing authentication information such as passwords. Passive interception of network operations enables adversaries to see upcoming actions. Passive attacks result in the disclosure of information or data files to an attacker without the consent or knowledge of the user.

2. Active attack: It involves some modification of the data Stream or creation of the false stream. The attacker tries to bypass or break into secured systems [4]. This can be done through stealth, viruses, worms, or Trojan horses [5]. Active attacks include attempts to circumvent or break protection features, to introduce malicious code, and to steal or modify information. These attacks are mounted against a network backbone, exploit information in transit, electronically penetrate an enclave, or attack an authorized remote user during an attempt to connect to an enclave. Active attacks subdivided into four categories; masquerade, replay, modification of message, and denial of service.

3. Distributed attack: It requires that the adversary introduce the code, such as a Trojan horse or back-door program, to a ―trusted‖ component or software that will later be distributed to many other companies and users Distribution attacks focus on the malicious modification of hardware or software at the

factory or during distribution [1, 4]. These attacks introduce malicious code such as a back door to a product to gain unauthorized access to information or to a system function at a later date.

4. Insider attack: It is among the most difficult to detect and prevent. It involves someone from the inside, such as a disloyal employee, attacking the network. Insider attacks can be malicious or no malicious. Malicious insiders intentionally eavesdrop, steal, or damage information; use information in a fraudulent manner; or deny access to other authorized users. No malicious attacks typically result from carelessness, lack of knowledge, or intentional circumvention of security for such reasons as performing a task.

5. Phishing attack: The hacker creates a fake web site that looks exactly like a popular site such as the SBI bank or PayPal. The phishing part of the attack is that the hacker then sends an e-mail message trying to trick the user into clicking a link that leads to the fake site. When the user attempts to log on with their account information, the hacker records the username and password and then tries that information on the real site.

6. Hijack attack*:* In a hijack attack, a hacker takes over a session between you and another individual and disconnects the other individual from the communication. You still believe that you are talking to the original party and may send private information to the hacker by accident.

7. Spoof attack*:* In a spoof attack, the hacker tries to access the network IP address. After gaining access to the network with a valid IP address, the attacker can modify, reroute, or delete your data.

8 Buffer overflow*:* A buffer overflow attack is when the attacker sends more data to the system that is expected. A buffer overflow attack usually results in the attacker gaining administrative access to the system in a command prompt or shell.

1.5 Data set description

To enforce the intrusion detection the NSL-KDD data set is used, this data set was adopted in March 2009 and was the altered and enhanced version of the old data set. It is composed of chosen records from which it is a complete data set and has been the most severely used data set used for the study and assessment deviation on the basis of intrusion detection. KDD is the most approved data set for intrusion detection as a correlate to other possible data set as it is well labeled and consist various attack types and shows the multiple attack scenarios, whereas the other data set are limited. KDD'99 is consist of about 4 gigabytes of raw compressed data of 7 weeks of network traffic collected with the help of tcpdump which is packet investigator runs under the command line used to prevent network packets flowing under the network. This data was processed into around 5 million connection records, each with about 100 bytes. A dataset composed of around 4,900,000 data packets with each packet containing 41 features and one class indicating packet either normal or attack with classifying the attack type. The NSL-KDD data set that intrusion detection uses is a refined data set of original KDD'99 data set. KDD data set was having the issues of duplicate or redundant data that was removed in NSL-KDD (27) data set. Repeated data is having the drawback of inclination the learning algorithms that are withdrawn in NSL-KDD (27) data set, this makes data set more practical for attack detection. The data set consists of 41 features and 1 class labeled as Normal and Anomaly.NSL-KDD data set has two individual sets, one is trained data and other is test data set. Train data consist of 24 attack types and on the other hand, test data consist of 14 additional attacks involves prior 24 attacks, making the detection more realistic and accurate because the accuracy of learning algorithms are also checked for prior unseen attacks.

II. RELATED WORK

Dias GV et.al [1] conducted a study indicating an intrusion detection system based on SVM methodology that combines an algorithm (hierarchical clustering), feature selection method and the technique of SVM. The algorithm i.e. used helps in providing the support vector machine with maintaining an abstracted form of high level of trained examples obtained from the trained set-up of KDD Cup 1999. The study indicates high level performance of SVM based technology which further resulted in a reduced form of training-time. The method of feature-based selection was adopted to remove the un-necessary features of the training set in order to maintain the levels of accuracy. The dataset of KDD cup-1999 was used to analyze the proposed system. When the system was compared with the other forms of data set, the experimental analysis showed that the result based on the performance analysis was not so good as compared to KDD Cup-1999 dataset. So, the methodology based on this dataset showed better analysis in detection of probe and DoS based attacks, maintaining accuracy globally. Cannady et.al [2] proposed a study on the process of misuse detection which is defined as a process to recognize the instances of different types of attacks by measuring the unexpected activity and the activity that is going currently. Mostly, the present processes based on misuse detection uses a technology of rule-based systems with the aim to identify the provoked nature of the attacks known to us. But the above process was less reliable to guess the forms of distinct attacks done on the system. The use of ANN technology gave a potential to search and identify the activities of the network that rely on the incomplete, non-linear, and limited amount of sources. Kemmerer et.al [3] presented a study by framing a simple question of why there is a need for the intrusion detection system. Suppose, the owner of a house is out of town and he has locked his home with all the windows and doors closed. But, there is someone outside his home who wants to enter. Firstly, he rings the bell and checks the main door if it is locked or not then after some time he checks the windows of the house that too are locked which makes sure that the house is safe. So, the question is why an alarming bell is installed. This question particularly sticks to the IDS. Why there is a need to plant the detection systems if the security is tight and secure. The reason to install these detective systems is that the intrusions still exist because sometimes the people may forget to lock their doors or windows, the same case occurs with the computer-based networks which do not provide us 100% security of the system to work accurately. So, based on this study the researchers have tried to explain the techniques based on IDS to deal with these kind of intrusions present in the network. Steven T et.al [4] proposed a study on an application of STATL that represents a descriptive language based on a transition-based attacking system that is constructed to support the IDS. This form of descriptive language describes a process of penetration done to the computer network implemented by a hacker. These type of penetrations includes attacking activities performed by the hacker. The STATL description is used by the IDS to extract the stream events and the ongoing intrusions occurring in the system. As the IDS

works under distinct environments such as Windows NT, Linux etc. and the domains like the host or the network. So, this extensible form of language helps in dealing with different targets as required. This language basically describes both the host and the network attacks. Here, in this paper an IDS based tool-set i.e. based on the descriptive language has been executed. This tool-set depicts various favorable and desired results. There is a deep study of syntax based on the STATL language. Common real examples of both the network and the host are also described in the paper. Pi-Cheng et.al [5] conducted research based on two of its issues related to the IDS designs. The two issues include the selection based on the optimization of rule-based selection and the discovery in case of attack. This type of approach provides a connection between the junked packets. An algorithm is implemented for the attack identification and the rule-based selection. The study is performed on the threats and describes the relationship for an application based web-server and the gateway. The algorithm is implemented over a signature-based IDS for having a better form of results. Cavusoglu et.al [6] conducted research on security systems of IT. The information technology firms rely on various forms of technologies such as IDS and the firewalls to manage the risks of the organizations. There exist some most interesting facts related to security alerts in IT industries. This paper presented a study to demonstrate the values of IDS adopted in an IT company. The configuration of IT was represented by the true-positive and the false-positive rates which further consists of determining the negative or the positive rates of an organization. It was shown specifically that an organization or a firm experiences a positive-rate from an IDS based on one of the conditions that the rate of detection is more than the critical value. When a firm experiences a positive or a negative value, an IDS prevents the occurrence of hackers that means an IDS targets the hacker's activity whether the alarm is positive or negative as the rate of detection is same. The results so obtained showed that the positive rate detected by an IDS is the result of an increased amount of deterrence enabled by its improved detection. The use of an optimized form of IDS indicates that the firm experiences a value i.e. non-negative in nature. Chebrolu, Srilatha et.al [7] conducted research on IDS that examined all data features to detect intrusion or misuse patterns. Some of its features may be redundant or contribute small quantity to the detection process. The purpose of this study was to identify unique input features in building an IDS i.e. efficient and effective computationally. An investigated was done based on the performance of feature-selection algorithms. The first one was the Bayesian networks (BN) and the other was the classification and regression trees (CART) including an ensemble of both the BN and CART. The results showed that input feature-selection was mainly required to design an IDS i.e. light in weight, effective and, efficient for real scenario detection techniques. In the end, the researchers proposed an architecture i.e. hybrid in nature for joining the different feature-selection algorithms for current scenario intrusion detection.

Kim, Dong Seong, et.al [8] proposed a method based on Genetic Algorithm to revamp SVM (Support Vector Machines) based IDS (Intrusion Detection System). The SVM denotes a novel-classification technique that has shown a high-class performance in various applications. The security researchers have proposed SVM based IDS. Here, they have used the fusion of SVM and GA to boost global performance. This type of inter-mixing resulted in an "optimal detection model" for SVM classifier where this method not only represented the "optimal-parameters" for SVM but also resulted in an "optimal-feature set" among the data-set. A demonstration was done to check the feasibility of the method by performing experiments on data-set named KDD 1999 for detection of intrusions in the system. Carl, Glenn, et.al [9] proposed a study based on detection using Denial-of-service (DoS) techniques that includes change-point detection, activity profiling, and signal analysis (wavelet-based) that further faced a major challenge to analyze the attacks on the network that generated from the sudden unexpected activities or flash-events. This survey of techniques and testing results provided a mechanism to identify DoS based flooding attacks. As the detectors used in the process are quite good but none of them has shown the complete accurate detection. The adjoining of various methodologies with smart and intelligent network handlers would definitely produce excellent results. Kim, Jungwon, et.al [10] conducted research on the use of artificial immune systems in IDS which is an interesting concept that relied on two main reasons. Firstly, the immune system of a human provides the best protection. Secondly, the present techniques used for maintaining computer security are less reliable and complex in nature. Here, the researchers have used various distinct algorithms for the development of the systems and the best possible outcomes. The analysis has been done based on the important developments within this area of research, in addition to forming suggestions for future research options. Panda, et.al [11] worked on the mining techniques if the data that are applied in designing the IDS in order to secure computational resources against access i.e. unwanted. This paper has shown the unique performance of well-defined data-mining classifier-algorithms such as ID3, J48 and Naïve Bayes that have been evaluated based upon the 10-fold-cross-validating test. The data that has been used is KDDCup'99 IDS which further shown that the Naïve Bayes method is the most effective algorithm of learning based process, and the mechanism adopted for decision trees is more interesting for the purpose of detection. P. Garcı´a-Teodoro, et.al [13] conducted a study on IDS i.e. an anomaly based network technique which consists of protecting the system target against all the harmful activities. This paper starts with a study and a method to review the anomaly based IDS. Further, the development of the system based on detection methods and various research projects are explained. The paper states the major challenges of anomaly-based intrusion detection system, dealing with special issues based on its applications. Muamer N., et.al [14] conducted a study on using a smart and intelligent form of data-mining approaches to observe the intrusion occurring in the local-networks. This paper suggested an improved strategy for Intrusion Detection System (IDS) that combines the expert systems, the processes of data mining as implemented in WEKA. The classification generally consists of the detection principle as well as some of the aspects of WEKA such as open-source data-mining processes. The combining methodology gives better performance of IDS based systems and helps to maintain the detection more effectively. The result was based on evaluating a new design produced a better form of detection based on efficiency. So, the study presented a good approach to analyze the experiments on behalf of intrusion detection. M. A. Jabbar, et.al [15] proposed the research based on the intrusion detection system to notify and identify the type of activities or normal users or the hackers

performing malicious operations. The IDS represents complicated and a linear problem dealing with traffic-data of the network. Many forms of IDS classes have been developed and proposed which further produced distinct levels of accuracy with the aim to maintain a robust and effective Intrusion detection system that is a necessary requirement. In this paper, a model has been designed for intrusion detection system (IDS) using a classifier based on the random forest where the Random Forest (RF) denoted an ensemble classifier and that performed very well as compared to the other classifiers that worked traditionally for effective classification of different forms of attacks. The experiments were conducted on a data-set named NSL-KDD in order to calculate and analyze the performance of the system and the empirical form of the result showed that the proposed model is more efficient for high rate detection and the detection of false alarm. Yu-lin He, at al [16] proposed the method of fuzziness based on the technique of instance selection for a huge amount of data sets in order to increase the supervised learning algorithm based efficiency. It did so by improving the design shortcomings of the intrusion detection system (IDS). The methodology proposed was dependent over a new type of single layer feed-forward neural network (SLFN) known as random weight neural network (RWNN).

III. THE PROPOSED METHOD

3.1 Proposed Methodology

In this paper we have proposed a hybrid model which consists of SVM i.e. Support Vector Machine combined different classification-algorithm to mitigate the rates of the false-positive alarms. To obtain the pre-thesis objective a methodology has been proposed which is further divided into three types of phases.

Phase 1: Collection and preprocessing

- Data-set collection
- Extraction of features through a data i.e."tcpdump"
- Converting the obtained features into binary representation
- Preparation of the input for its classification

Phase 2: Classification:

- To find the best classifier from the available classifier.
- To test and train the tool of classification by the dataset-partitioning process.

Phase 3: Result analysis

- To compare the obtained results with their existing work.

The proposed working methodology is designed as below in the figure.5. In order to start the proposed work. The first step is to study all the data-set obtained from the different sources, to eagerly check the data format of the data and further to analyze that which form of mining technique should be applied over the data. When the set of data gets collected then the process of feature extraction is carried out. Further, the process is will be classified into two classes. The first one is testing and the other one is the training of classification tools i.e. done with the help of classifiers. Then the results are further analyzed based on various forms of performance metrics.

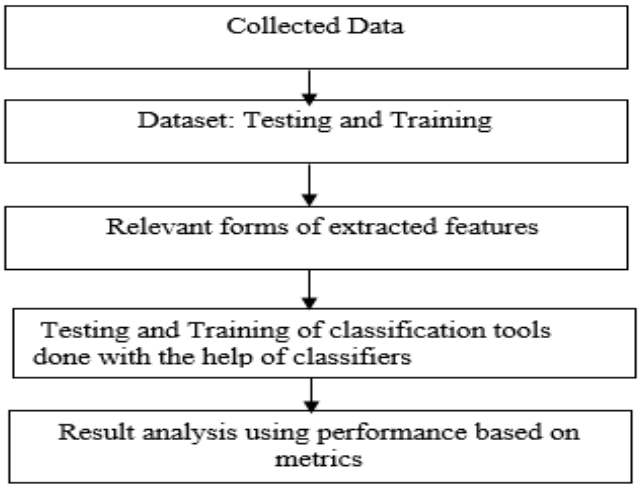

Figure 5: Proposed Methodology

3.2 Proposed methodology: Flowchart

The proposed steps of the flow chart are given below:

1. KDD-99 Data Set: This is a type of data-set used for the (Third International Knowledge Discovery and Data Mining Tools) Competition, held in conjunction with KDD-99 (The Fifth International Conference on Knowledge Discovery and Data Mining). The main task was to build a network based on intrusion detection and to predict a model i.e. capable of discriminating a good or a bad form of data-set. This form of data-set maintains a standard including a wide variety of network-based intrusions.

2. Label Features: A label helps in providing complete information regarding the set of data.

3. Input in PSO: Each of the particles has its velocity and position to search for a better solution. So, the velocity and position are the inputs used in PSO.

4. Initialize particles: The PSO-based technique is initialized with a population of random solution.

5. Update fitness function: It helps in judging the individual solutions based on how well they can handle the problem.

6. Optimize Objective Form: Here, the objective is optimized.

7. Initialize chromosomes: The process is initialized by building a population of chromosomes which is a set of possible solutions to the optimization problem.

8. Check the convergence: These type of methods helps in testing the conditional-convergence, absolute-convergence, interval of convergence or divergence of an infinite series.

9. Cross Over: A point or place of crossing from one side to the other.

10. Roulette Selection: It is a method used in genetic-algorithms for selection of potentially useful solutions for the purpose of recombination.

11. Optimize features: This type of method achieves the best designing technique.

12. Neural Networks: It represents a biologically inspired information processing system.

13. Test Model: It performs a system or software system.

14. Precision, Recall Accuracy: The precision is a good measure that determines the cost of False Positive is high.

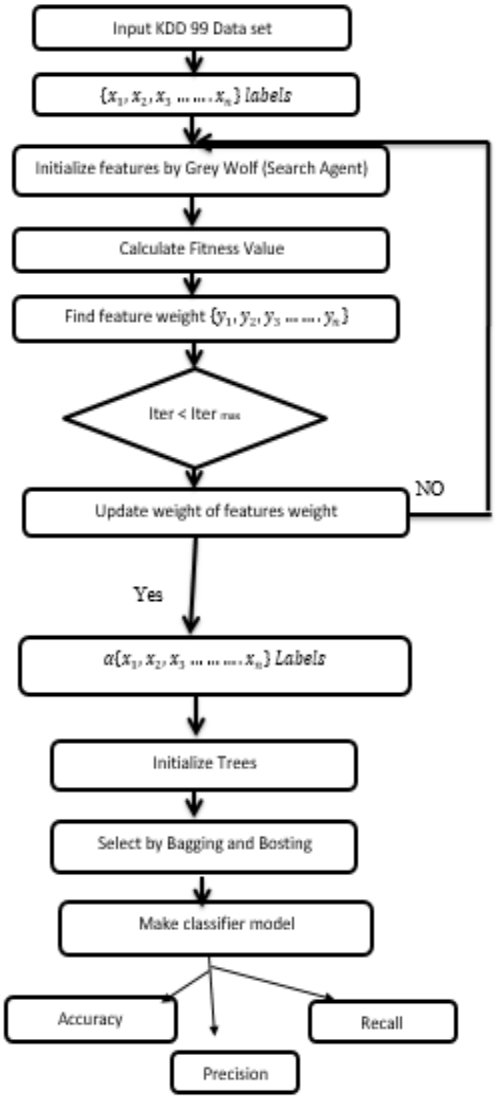

Figure 8: Proposed Flowchart

3.3 Algorithm Used

1. Grey Wolf Optimization Algorithm (GWO*):* Grey wolves have the ability to memorize the prey position and encircling them. The alpha as a leader performs in the hunt. For simulating the grey wolves hunting behavior in the mathematical model, assuming the alpha (α) is the best solution. The second optimal solution is beta (β) and the third optimal solution is delta (δ). Omega (ω) is assumed to be the candidate solutions. Alpha, beta and delta guide the hunting while position should be updated by the omega wolves by these three best solutions considerations.

Encircling prey

Prey encircled by the grey wolves during their hunt. Encircling behavior in the mathematical model, the below equations is utilized.

$$\vec{A}(T+1) = \overrightarrow{A_P}(T) - \vec{X}.\vec{Z}$$
$$\vec{Z} = \left|\vec{Y}.\overrightarrow{A_P}(T) - \vec{A}(T)\right|$$

Where

T←iterative number

$\vec{A}$←grey wolf position

$\overrightarrow{A_P}$←prey position

$$\vec{X} = 2x.\overrightarrow{r_1} - x$$
$$\vec{Y} = 2\overrightarrow{r_2}$$

Where

$\overrightarrow{r_1}$and $\overrightarrow{r_2}$←random vector range[0,1]

The x value decrease from 2 to 0 over the iteration course.

$\vec{Y}$←random value with range [0,1] and is used for providing random weights for defining prey attractiveness.

Hunting

For grey wolves hunting behavior simulation, assuming α, β, and δ have better knowledge about possible prey location. The three best solutions firstly and ω (other search agents) are forced for their position update in accordance with their best search agents position. Updating the wolves' positions as follows:

$$\vec{A}(T+1) = \frac{\overrightarrow{A_1}+\overrightarrow{A_2}+\overrightarrow{A_3}}{3}$$

(1)

Where $\overrightarrow{A_1}$, $\overrightarrow{A_2}, and$ $\overrightarrow{A_3}$ are determined,

$$\overrightarrow{A_1} = \left|\overrightarrow{A_\alpha} - \overrightarrow{X_1}.Z_\alpha\right|$$
$$\overrightarrow{A_2} = \left|\overrightarrow{A_\beta} - \overrightarrow{X_2}.Z_\beta\right|$$
$$\overrightarrow{A_3} = \left|\overrightarrow{A_\delta} - \overrightarrow{X_3}.Z_\delta\right|$$

Where $\overrightarrow{A_\alpha}$, $\overrightarrow{A_\beta}, and$ $\overrightarrow{A_\delta}$← first three best solutions at a given iterative T

Z_α, Z_β, and Z_ω are determined,

$$\overrightarrow{Z_\alpha} \leftarrow \left|\overrightarrow{Y_1}.\overrightarrow{A_\alpha} - \vec{A}\right|$$
$$\overrightarrow{Z_\beta} \leftarrow \left|\overrightarrow{Y_2}.\overrightarrow{A_\beta} - \vec{A}\right|$$
$$\overrightarrow{Z_\delta} \leftarrow \left|\overrightarrow{Y_3}.\overrightarrow{A_\delta} - \vec{A}\right|$$

The parameter x updating is the final process. The parameter x exploitation and exploration is updated linearly for ranging [2, 0] in every iteration.

$$x = 2 - t\frac{2}{maxI}$$

Where

T←iterative number

MaxI←total number of iteration

2. Random Forest*:* Random forest is a learning method for classification, regression and generating the multitude of decision trees. It generates the multitude at the time of training and output of the class. It provides high accuracy and learning is very fast in it. It works very effectively on the large size database. It easily handles the large size input variables without variable deletion [12, 15].

IV. RESULT ANALYSIS

4.1 Description of dataset

As discussed above experiments are executed by using KDD-99 which having 41 feature sets. These features are used for optimization and then learning and now they are used to analyse in terms of attack. In this work, we use to evaluate the accuracy rate in an intrusion detection system. In the analysis, we take data on the basis of a number of intrusions. Attacks are generally fall into four categories 1) Dos, 2) Probe, 3) R2L 4) U2R. In our analysis we use three categories 1) Other attack which consists of the probe, R2L and U2R 2) DoS-attack 3) Normal attacks (non-attacks). In this work, we evaluate the accuracy, precision, recall, and F-measure in various cases.

Table.1: Attack type from KDD CUP 99 dataset

	ANN	ANN with GA	ANN with PSO	GWRF
Accuracy	89.32	93	90	96.23
Precision	88.45	91	90.38	97.33
Recall	87.12	92	92	98.33
F-measure	85.89	94	87	93.13

Table.2: Static data to analyse the efficiency of the approaches for the above discussed attack

Normal	Dos	R2L	U2R	Probe
	Smurf	PHF	Root-kit	Portsweep
	Process table	Xlock	Eject	Satan
	Pod	Send-mail	Perl	Saint
	Land	Guess_password	Buffer overflow	M-scan

Table.3: Algorithm Types vs. Types of attack in terms of accuracy, precision, recall, and F-measure

Algorithm type	**Types of attack**	**Accuracy**	**Precision**	**Recall**	**F-measure**
ANN	Other attack	87	84	87	83
	Dos Attack	88	87	89	84
	Normal Attack	90	87	86	86
ANN with GA	Other attack	94	88	85	87.23
	Dos Attack	89	91	86	86.23
	Normal Attack	90	90	83	89.13
ANN with PSO	Other attack	92	90	85	84.23
	Dos Attack	95	89	87	83.23
	Normal Attack	91	89	90	87.34
GWRF	Other attack	98.62	96.23	97.33	95.23
	Dos Attack	92.23	90.23	96.33	95.13
	Normal Attack	97.23	96.13	99.56	92.23

4.2 Result Analysis

In this section, we analyse the statistical data by simulation. Graphical result of table.1 and 3 is given by the simulation process.

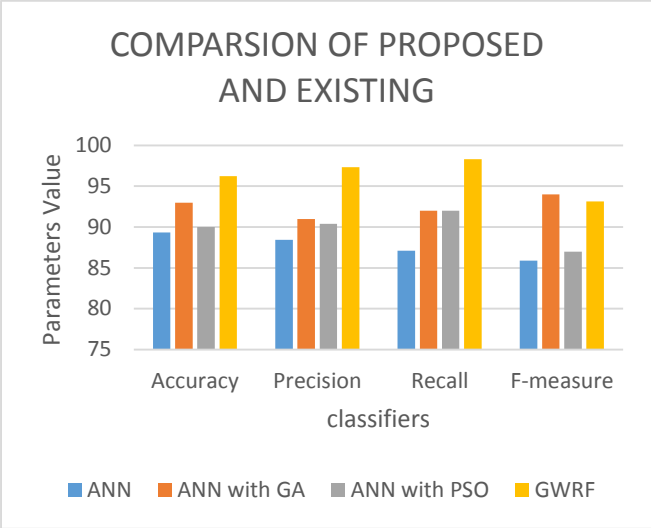

Figure 9: Simulated graph of table.1

Figure 9 shows the simulated analysis of table.1 in terms of accuracy, precision, recall, and f-measure. In this figure analysis on efficiency is demonstrated from all the four algorithms that are ANN represented by the green line, ANN with PSO represented by the purple line, ANN with GA represented by the red line and ANN with both PSO and GA represented by the blue line. The analysis demonstrates that ANN with both SPO and GA giver better result in terms of all the four parameters (accuracy, precision, recall, F-measure).

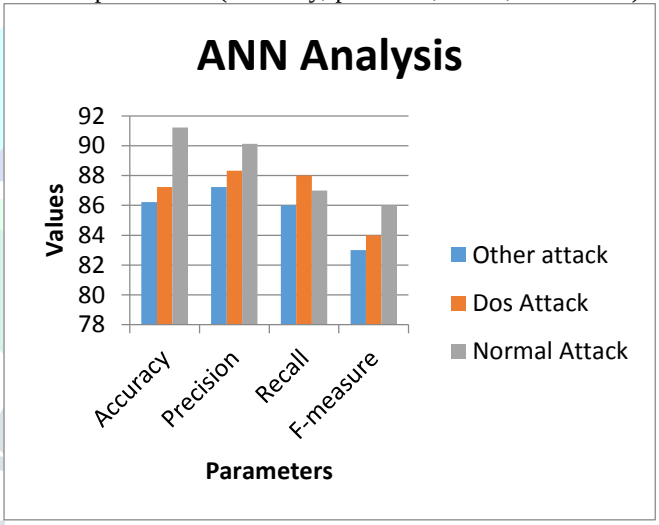

Figure 10: Analysis of ANN

Figure 10 represents the parameters investigation of various classifier and proposed approach. In investigation parameters like exactness, review, precision and f measure fluctuate as indicated by classifier yet one examination clear about the proposed approach (PSO with GA in the neural system) demonstrate huge enhance all parameters. In the event that examination just proposed a methodology, review demonstrate huge enhancement then different parameters so it will clear sign of decreasing false negative rate so assaults distinguishing proof is successful in proposed approach in view of enhancing weight given by PSO_GA approach.

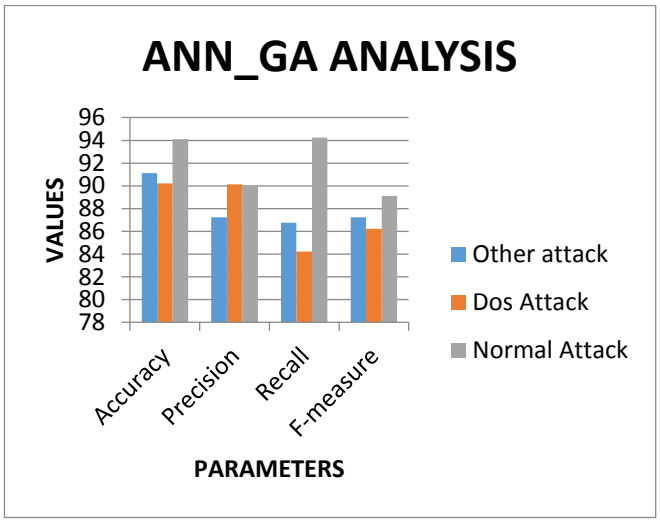

Figure 11: Analysis with ANN and ANN _GA

Figure 10 and 11 parameters investigation of various classifier and the proposed approach. In investigation parameters like exactness, review, precision and f measure fluctuate as indicated by classifier yet one examination clear about the proposed approach (PSO with GA in the neural system) demonstrate huge enhance all parameters. In the event that examines just the proposed a methodology, review demonstrate huge enhancement then different parameters so it will clear sign of decreasing false negative rate so assaults distinguishing proof is successful in proposed approach in view of enhancing weight given by PSO_GA approach GWRF.

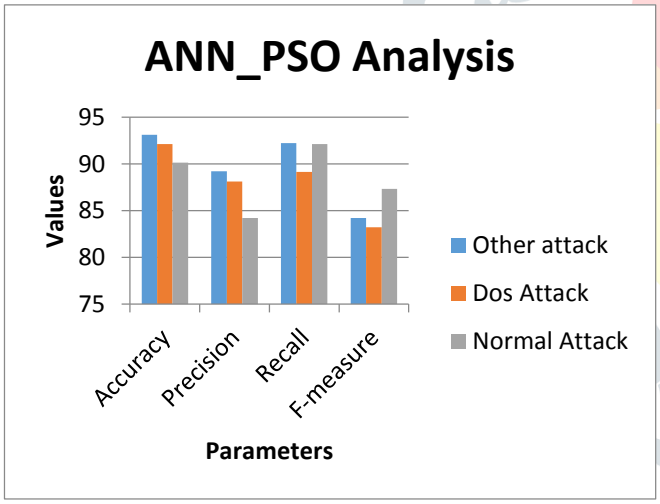

Figure 12: Analysis of ANN _PSO

Figure 12 performs profundity investigation of every one of the three classes in ANN and ANN_GA. In this examination, we endeavour to indicate what the criticalness of our methodology is. This exchange we proceed in perception (3) moreover. So first point which examination by typical class n which no assault working and in the two cases ANN and ANN with GA perform well contrast with other parameters like exactness, review, and f-measure yet ANN_GA still preferable precision over ANN so include weighted by enhancement by one way or another perform in view of decreasing covering data learning. On the off chance that examination through DOS assault it additionally indicates higher exactness in ANN with GA.so we can close Feature advance weight is better methodology so by what means can enhance improvement, these perceptions talk about in next standard.

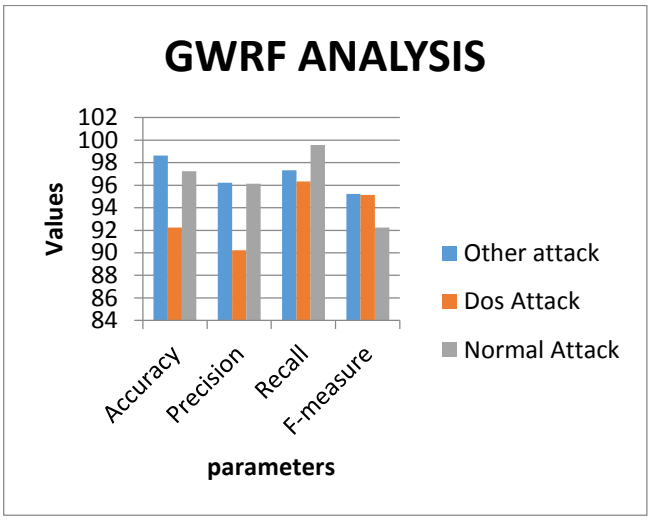

Figure 13: Analysis with ANN _PSO and GWRF

At last from the whole analysis, it can be concluded that algorithm GWRF gives a better result for all the attacks we examined in our work. In figure 13 examination proceeds from perception (2) and attempt to discovering the centrality of streamlining enhancement impact on various classes' recognition by characterization. On the off chance that investigation both chart demonstrates the compelling review however for ordinary class so decrease the false positive rate this enhancement occurring with all classes like DOS assault and different assaults yet the viable outcome appear in other assault which increment fundamentally in the proposed approach. So PSO streamlining is great however PSO with GA more enhance in other assault and ordinary class.

V CONCLUSION

Intrusion can be characterized in terms of confidentiality, integrity, and availability. An event or action causes a breach of confidentiality if it allows to access resources, residing in a computer in an unauthorized manner. An event or action causes a breach of integrity if it allows to change the states of resources, residing in a computer in an unauthorized manner. Similarly, an event or action causes a breach of availability if it prohibits legitimate users to access resources or services, residing in a computer. Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions. An intrusion detection system is a software or hardware that automates the process of monitoring and analyzing of events. The present scenario experiences various forms of developments and huge growth in advanced processing technologies consisting of connectivity among different networks but the methodology is vulnerable by the activities of the intruders or the attackers of the system. These specifically smart attackers interrupt the operation with new and fascinating methods of data-breaching among large networks. Though there are various forms of available intrusion of intrusion detection systems that can detect the intrusions occurring in the network i.e. based on the false positive detection rate and the alert rates but with the detection rate of intrusions, they also have a high false-positive rate resulting in an adequate system comprising of low accuracy level of the system and are generally more prone to different kinds of attack. This usually helps the intruder to enter into the system and perform a pre-planned attack. So, this pre-thesis will propose a hybrid approach to reduce false positive alarms. The experimental analysis

consists of a specified particular form of data-set and the process of feature-based selection will be done to improve the analysis. These features obtained will be used for the classification-tool training and testing the performance of the system. Finally, the result obtained will be compared with the results that already exist.

VI REFERENCES

[1] Snapp, Steven R., James Brentano, Gihan V. Dias, Terrance L. Goan, L. Todd Heberlein, Che-Lin Ho, Karl N. Levitt et al. "DIDS (distributed intrusion detection system)-motivation, architecture, and an early prototype." In *Proceedings of the 14th national computer security conference*, vol. 1, pp. 167-176. 1991.

[2] Cannady, James. "Artificial neural networks for misuse detection." In *National information systems security conference*, vol. 26. 1998.

[3] Kemmerer, Richard A., and Giovanni Vigna. "Intrusion detection: a brief history and overview." *Computer* 35, no. 4 (2002): supl27-supl30.

[4] Eckmann, Steven T., Giovanni Vigna, and Richard A. Kemmerer. "STATL: An attack language for state-based intrusion detection." *Journal of computer security* 10, no. 1-2 (2002): 71-103.

[5] Hsiu, Pi-Cheng, Chin-Fu Kuo, Tei-Wei Kuo, and Eric YT Juan. "Scenario based threat detection and attack analysis." In *Security Technology, 2005. CCST'05. 39th Annual 2005 International Carnahan Conference on*, pp. 279-282. IEEE, 2005.

[6] Cavusoglu, Huseyin, Birendra Mishra, and Srinivasan Raghunathan. "The value of intrusion detection systems in information technology security architecture." *Information Systems Research* 16, no. 1 (2005): 28-46.

[7] Chebrolu, Srilatha, Ajith Abraham, and Johnson P. Thomas. "Feature deduction and ensemble design of intrusion detection systems." *Computers & security* 24, no. 4 (2005): 295-307.

[8] Kim, Dong Seong, Ha-Nam Nguyen, and Jong Sou Park. "Genetic algorithm to improve SVM based network intrusion detection system." In *Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on*, vol. 2, pp. 155-158. IEEE, 2005.

[9] Carl, Glenn, George Kesidis, Richard R. Brooks, and Suresh Rai. "Denial-of-service attack-detection techniques." *IEEE Internet computing* 10, no. 1 (2006): 82-89.

[10] Kim, Jungwon, Peter J. Bentley, Uwe Aickelin, Julie Greensmith, Gianni Tedesco, and Jamie Twycross. "Immune system approaches to intrusion detection–a review." *Natural computing* 6, no. 4 (2007): 413-466.

[11] Panda, Mrutyunjaya, and Manas Ranjan Patra. "Network intrusion detection using naive bayes." *International journal of computer science and network security* 7, no. 12 (2007): 258-263.

[12] Zhang, Jiong, Mohammad Zulkernine, and Anwar Haque. "Random-forests-based network intrusion detection systems." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, no. 5 (2008): 649-659.

[13] Garcia-Teodoro, Pedro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. "Anomaly-based network intrusion detection: Techniques, systems and challenges." *computers & security* 28, no. 1-2 (2009): 18-28.

[14] Mohammed, Muamer N., and Norrozila Sulaiman. "Intrusion detection system based on SVM for WLAN." *Procedia Technology* 1 (2012): 313-317.

[15] Farnaaz, Nabila, and M. A. Jabbar. "Random forest modelling for network intrusion detection system." *Procedia Computer Science* 89 (2016): 213-217.

[16] Ashfaq, Rana Aamir Raza, Yu-lin He, and De-gang Chen. "Toward an efficient fuzziness based instance selection methodology for intrusion detection system." *International Journal of Machine Learning and Cybernetics* 8, no. 6 (2017): 1767-1776.