Performance Comparison of Host based and Network based Anomaly Detection using Fuzzy Genetic Approach (FGA)

[1]Harjinder Kaur, [2]Nivit Gill

[1]*M.E (CSE) Student, Punjabi University Regional Centre of IT & Management, Mohali, Punjab, India,*

[2]*Assistant Professor (CS), Punjabi University Regional Centre of IT & Management, Mohali, Punjab, India,*

Abstract: **Intrusion is a deliberate unauthorized access, attempt, misuse or damage to some valuable data. Intrusion Detection Systems (IDS) are used to detect and report the intrusions for the computer systems and for the computer networks. IDS analyses the data or traffic and classifies the behavior of the particular host and a network into the normal or the suspicious activity. This paper compares the performance of the host based and the network based intrusion detection systems implemented using the Fuzzy Genetic approach. System log files are used as the dataset for the host based intrusion detection (HIDS) and NSL-KDD dataset is used for the network based intrusion detection (NIDS). Simulation results reveal that HIDS detects the normal behavior as well as the anomalous behavior better than NIDS.**

Keywords: **Intrusion, host/network based intrusion detection, NSL-KDD dataset, fuzzy logic, genetic algorithms.**

I. INTRODUCTION

Due to the enormous growth of the usage of computers, computer networks and its applications, the security concern has become very crucial. Intrusion Detection Systems (IDS) are used to monitor the behavior of the system and the network and to report significant deviations from the normal activity, if any. Intrusion detection is used for identifying attacks in computer networks and malicious activities in computer systems [1]. Computer systems are usually referred to as hosts, and computer networks are referred to as networks. The systems that detect the anomalies in host's behaviour are named as Host based intrusion detection systems (HIDS) and the systems which detect the anomalies in network behaviour are called as Network based intrusion detection systems (NIDS). HIDS monitors the state and the behavior of the computer system using the log files and finds out its normal and the anomalous behavior. A network has different types of activities over the network and these activities could be traced by an intruder or attacker. NIDS monitors the activities of the network and classifies them as normal or an anomalous activity.

In this paper, a hybrid Fuzzy Genetic approach has been employed to implement HIDS as well as NIDS, and their performance is evaluated on the basis of best fitness. For HIDS, the system log files are used to analyze the user behavior and NSL-KDD dataset is used to classify the network traffic as the normal or the anomalous for the NIDS.

The subsequent parts of this paper are organized as follows: Section 2 reviews the work done by various authors in the field of intrusion detection and security, Section 3 sheds light on the several techniques of the intrusion detection systems. Then the Section 4 outlines the machine learning based techniques of anomaly detection and details the basic concepts of Fuzzy logic and Genetic algorithms that have been used for HIDS and for implementing NIDS. Section 5 explains the proposed work of the research. Next, Section 6 explains the datasets used: system log files used to extract user behavior of the host and the NSL-KDD dataset used to extract the behavior of the particular traffic of the network. In Section 7, the experimental results as well as their evaluation are presented. Finally, Section 8 concludes the paper along with its future scope.

II. RELATED WORK

Anomaly Detection techniques are proposed by various authors to detect and prevent the anomalies in the system and the network. The author K. Hanumantha Rao et al. [1] has proposed the machine learning algorithms to classify the normal and abnormal activities in the computer network by use of ID3 decision tree and K-means clustering learning methods. They have also used the supervised and unsupervised algorithms that have provided the best efficiency or the best learning. HIDS and NIDS fundamentals are provided by Jiankun Hu [2] along with the architectural framework. In this framework, multiple detection engines are used and its novelty lies in the existing feedback loops. It has also introduces Hidden

Markov Models (HMM) and HMM-based anomaly intrusion detection schemes. S. Mallissery et al. [4] and A. S. Ashoor et al. [5] have reviewed the various categories of intrusion detection techniques. The authors have categorized the intrusion detection techniques on the basis of data source and on the model of intrusions. V. Jyothsna et al. [6] and P. G. Teodoro et al. [7] have classified the anomaly detection techniques into statistical, cognition and machine learning based techniques. However with all these techniques, there is a need of hybrid intrusion detection which can detect static as well as dynamic attacks i.e. can detect attacks on host as well as on network. B. Shanmugam et al. [8] has proposed a hybrid IDS which is more accurate, low in false alarms, intelligent by using Fuzzy mechanisms and not easily deceived by small variation. But the system has crashed, as it could not withstand the traffic for more than three weeks without restarting.

Om Prakash Shukla et al. [9] has explained the optimization techniques and mainly focused on the Genetic algorithms which are inspired by natural evolution of generations. Each individual in the generations are characterized by a fitness function. The offsprings that are generated from the parents can provide optimal solution if they are well designed. R. Borgohain [10] has explained the Fuzzy logic and Genetic algorithms in detail and discussed various rules and parameters of these techniques. J. T. Yao et al. [12] has presented the intrusion detection techniques focusing mainly on the abnormal system events where fuzzy sets play an important role. Authors have also proposed a dynamic approach that discovers known or unknown intrusion patterns with the help of fuzzy sets and the support vector machines. The author M. Hassan [15] has reviewed some of the current and the past intrusion detection technologies, which includes fuzzy logic and genetic algorithms. Author has also proposed a work that implements the Genetic Algorithm (GA) and fuzzy logic using the KDD 99 dataset of network intrusion detection. GA efficiently identifies various types of network intrusions and fuzzy logic determines false alarm rate by which intrusive activities can be minimized.

This research work aims to compare the performance of the host and the network based anomaly detection using the fuzzy genetic approach. The hybrid of fuzzy logic and genetic algorithm has been earlier applied by the author M. Hassan for the network based intrusion detection using KDD 99 dataset [15].

III. INTRUSION DETECTION SYSTEMS

Although cryptography has provided a powerful tool for the computer and the network security, it focuses more on attack detection and prevention. Unfortunately prevention of all possible attacks is impossible. Therefore, a second line of defense is needed where the Intrusion Detection System (IDS) comes to play an important role. Intrusion refers to unauthorized activity including unauthorized access to data or a computing service [2]. IDS attempts to identify such intrusion activity that is occurring or has already occurred. There are several benefits of IDS; it can generate and trigger the alarms whether there is a manual or an automated response so that further damage can be prevented and it can also help to assess the damage that has been done and provide court evidence of intruders, which in turn provides deterrence to attackers [2].

There are several ways to classify IDS as shown in Figure 1. One way of classifying IDS into three categories, namely, Network based Intrusion Detection Systems (NIDS) and Host based Intrusion Detection Systems (HIDS), which focuses on what physical targets or the source the IDS tries to protect. Network-based intrusion detection systems (NIDSs) collect input data by monitoring network traffic (e.g., packets captured by network interfaces in promiscuous mode). Host-based intrusion detection systems (HIDSs) rely on events collected by the hosts they monitor [3]. Hybrid of HIDS and NIDS has been also used. Another method of IDS classification is to classify IDS into misuse detection IDS and anomaly detection IDS based on the model of intrusion. Misuse IDS inspects a suspicious event against predefined attack database to find a match. This mechanism is very effective for attacks whose characteristics are known a priori. Anomaly IDS inspect whether an event is abnormal or not. It is a promising mechanism for detecting attacks whose characteristics are not known a priori [4], [5].

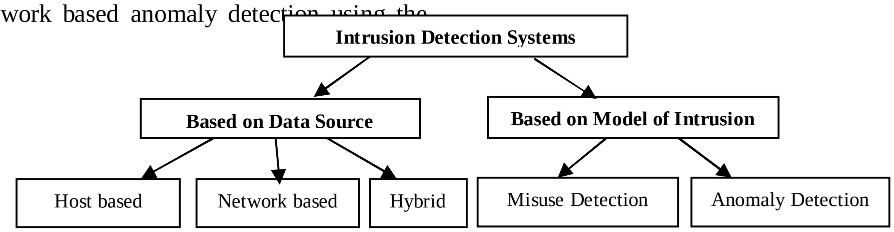


Figure 1: Types of Intrusion Detection Systems [5]

IV. ANOMALY

An anomaly is an event that is suspicious from the perspective of security. There are various anomaly detection techniques that can be classified on the basis of type of processing (related to the behavioural model) as Statistical based, Cognition based and Machine learning based [6] (Figure 2). Anomaly detection can be applied on both the host as well as the network. Statistical based detection systems capture the network traffic activity and generate a profile based on the stochastic behavior of the network. This profile is based on metrics such as the traffic rate, the number of packets for each protocol, the rate of connections, the number of different IP addresses, etc. [7]. Cognition-Based (also called knowledge-based or expert systems) detection systems work on a set of predefined rules, classes and attributes identified from training data, set of classification rules, parameters and procedures inferred. Machine learning based systems are used where no information is available except the ground truth and they used the small subset of data points to estimate the unknown attributes of test points [6]. There are various machine learning based techniques of anomaly detection among which Bayesian Networks are used to encode probabilistic relationships among the variables of interest and ability to incorporate prior knowledge and data, clustering techniques work by grouping the observed data into clusters, according to a given similarity or distance measure [7], Neural Networks have the ability to interpolate from noisy, incomplete and limited data and have the potential to recognize future unseen patterns [8]. Fuzzy Logic and Genetic Algorithms which are used to implement HIDS and NIDS are discussed in the following subsections.

A. Fuzzy Logic

Fuzzy logic was introduced by Dr. Lofti Zadeh of UC/Berkeley in the 1960’s as a means to model the uncertainty of natural language. Fuzzy logic is based on building a set of human language rules as specified by the user. The fuzzy systems convert these rules into their mathematical equivalents and thus simplifying the job of the computer and the system designer. The obtained results are much more accurate and it represents the way that systems behave in the real world. Fuzzy logic has also included the benefit of its simplicity and flexibility. Fuzzy logic can handle the problems with inaccurate and incomplete data and it can also model nonlinear functions of arbitrary complexity [8].

For several reasons, fuzzy logic is very appropriate for using on intrusion detection. One reason is that usually there is no clear boundary between normal and anomaly events. The use of fuzziness of fuzzy logic helps to smooth the abrupt separation of normality and abnormality. Another reason is that it can defined clearly when to raise an alarm that is fuzzy. At what degree of intrusion we should raise an alarm is often depends on different situations [12].

B. Genetic Algorithms

Genetic Algorithm is an evolutionary optimization technique which is inspired by natural evolution process where population of individuals gives feasible solution to the various problems. Many individual solutions are obtained from the initial population based on the size and the nature of the problem. Individuals are selected randomly from the population that allows the overall limits of possible solutions. The proportion of the existing population is chosen during each consecutive generation to breed a new generation [9].

Genetic Algorithms use the fitness functions in which individual solutions are selected, where fitter solutions are typically more likely to be selected. By such certain selection methods, we select the best solutions. The next way is to yield second generation population of solutions from the solutions that are selected through genetic operators. A pair of the "parent" solutions is selected for each new solution to be produced from the group selected previously. By using methods of crossover and mutation, a "child" solution is produced which typically includes many of his parent’s characteristics and the process continues till a new population of solutions is evolved [9]. These processes ultimately results in the next - generation of chromosomes that is dissimilar from the initial generation. In general, the average fitness is increased by this procedure for the population, but only the best population from the first generation is selected for the further breeding, along with a minor proportion of less fit solutions. Moreover the main genetic operators are Crossover and Mutation. This generational process is repeated till a termination condition has been reached [10] [11].

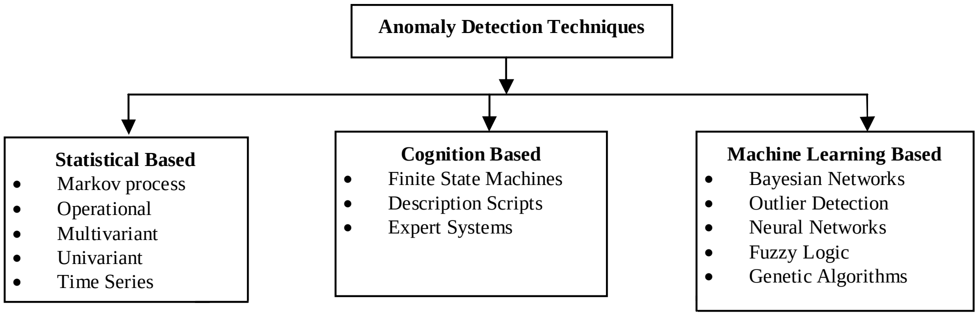


Figure 2: Types of Anomaly Detection Techniques [6]

V. PROPOSED WORK

The proposed work of the research includes three stages (Figure 3): Stage 1 is to design the host based intrusion detection (HIDS), Stage 2 is to design the network based intrusion detection (NIDS) and comparison of both the systems is to be done in Stage 3 based on the best fitness obtained in Stage 1 and Stage 2. Both Stage 1 and Stage 2 include two similar phases: Phase 1 and Phase 2. Phase 1 applies the Fuzzy approach on the input dataset i.e. system log files in case of HIDS and NSL-KDD dataset in NIDS. In Phase 2, Genetic Algorithms are applied on the Fuzzy output of Phase 1 for both HIDS and NIDS so as to find the optimized output by use of the fitness function. In Stage 3, the performance comparison of HIDS and NIDS is done with the help of the best fitness value obtained in Phase 2 of both Stage 1 and Stage 2.

A. *Stage 1*

In this stage, the host based intrusion detection system is implemented with system log files as the input dataset for the Phase 1. Fuzzy approach applied in Phase 1 includes following components: Input, Output, Inference Engine and fuzzy rules. Input contains different set of events as identified from the system log files. Output classifies the data into Normal class or Anomaly class. There are if-then type fuzzy rules that map the input to the output and fuzzy inference engine helps to convert the input to output with the help of such fuzzy rules.

Phase 2 uses the genetic approach for the optimization, in which a fitness function is designed on the fuzzy output obtained and the optimized result is found using the best fitness of the algorithm.

B. *Stage 2*

This stage is contrived for the network based intrusion detection system in which NSL-KDD is used as the dataset for first phase. In Phase 1, fuzzy logic is applied on the input dataset with following components. Input contains different set of attributes obtained by extracting the behaviour of network traffic. Output categorizes the attributes that define the class of network traffic as Normal or Anomaly. Then if-then type fuzzy rules relate the input and the output. Fuzzy Inference Engine maps the input to output using fuzzy rules.

In Phase 2, genetic approach is used for the optimization and a fitness function is applied on the output generated by Phase 1 and then a best fitness algorithm generates the optimized result.

C. *Stage 3*

This stage includes the comparison of the Stage 1 and Stage 2 i.e. the comparison of performance of HIDS and NIDS, based on the value of the best fitness obtained.

VI. DATASETS

In this paper, the performance of Fuzzy Genetic based HIDS and NIDS is evaluated and compared. Fuzzy Genetic HIDS is tested for the analysis of the user behavior of the system (host) and Fuzzy Genetic NIDS observes the behavior of the network.

A. *Dataset for HIDS*

Generally system log files are used as dataset for the host based intrusion detection Systems [16], in order to extract the user behavior of the system. For the fuzzy genetic based HIDS, seven day’s log files have been analyzed and it has been found that the behaviour is among one of the classes i.e either normal or anomaly.

Stage

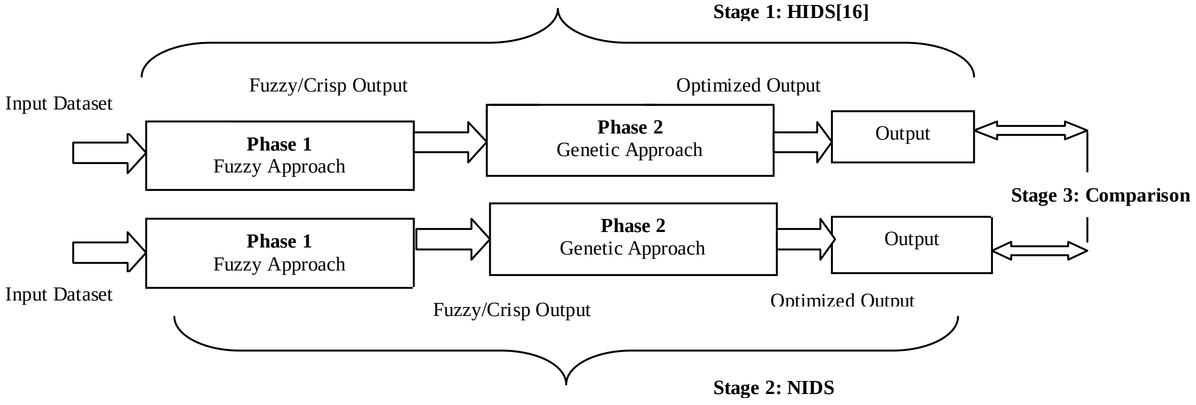


Figure 3: Proposed work of the research

B. *Dataset for NIDS*

NSL-KDD dataset is used for the network based intrusion detection systems. This dataset is the new version of KDDCUP 99 dataset. Since 1999, KDD’99 has been the most wildly used data set for the evaluation of anomaly detection systems. This data set is based on the data captured in DARPA’98 IDS evaluation program. But there are some critiques of attack taxonomies and performance measures [13].

KDD'99 features can be classified into three groups. First group includes the basic features that encapsulate all the attributes that are extracted from a TCP/IP connection. Second group incorporates the traffic features that are computed with respect to a window interval and is further divided into two sub groups:

- Time-based traffic features that give statistics about the connections in the past two seconds and
- Host-based traffic features that are fabricated to the same host using a window of 100 connections.

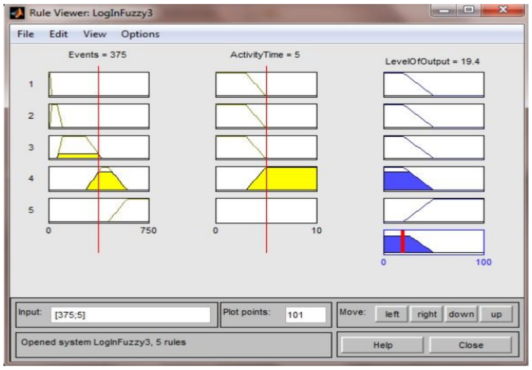
Third group contains the content features that use the domain knowledge and are derived from the suspicious behavior of the data portions, for e.g. the number of failed login attempts [14]. There are 42 attributes in NSL-KDD dataset, among which 41 attributes define the various properties of the network connection and the last attribute shows whether the connection produced is a normal traffic or an anomalous traffic.

VII. RESULTS AND DISCUSSION

The proposed work has been implemented using MATLAB as it provides toolboxes for both Fuzzy logic and genetic algorithms [11]. System log files have been studied and analyzed to get the set of events and timing information as an input to the HIDS and the fuzzy phase generates the output, as either information or a warning.

NSL-KDD dataset is used as the input for the NIDS [13]. Protocol type and the flags are the two inputs in NIDS using mamdani fuzzy inference model, and the fuzzy output can either be a normal class or an anomaly. Memberships of the input and the output of HIDS and NIDS are shown in Figure 3 and Figure 4.

For both the systems, the output of fuzzy phase is supplied to genetic optimization phase. Genetic algorithm measure the performance in terms of the best fitness value. The fitness value of an individual is the value of the fitness function for that individual. Since the MATLAB genetic algorithm toolbox find out the minimum of the fitness function and the best fitness value for the population is the smallest fitness value for any individual [11]. The results obtained for the optimization phase for the ten simulations are shown in Table 1 and Table 2 for HIDS. Table 1 presents the best fitness values of ten runs for information (Normal) class as detected by Fuzzy Genetic based HIDS, whereas the best fitness values for warning (Anomaly) class detected by Fuzzy Genetic based HIDS for ten simulations are given in Table 2. Best fitness for the Normal and Anomaly class of HIDS are 0.453 and 15.89 respectively.

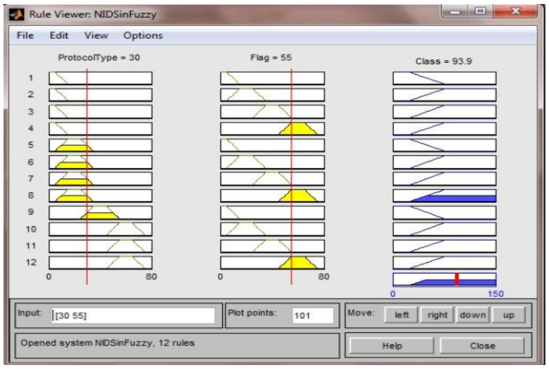


Input: [375;5]

Plot points: 101

Opened system LogInFuzzy3, 5 rules

Figure 3: Rule Viewer shows the membership of the inputs and the outputs of HIDS in MATLAB



Input: [30 55]

Plot points: 101

Opened system NIDSinFuzzy, 12 rules

Figure 4: Rule Viewer shows the membership of the inputs and the output of NIDS in MATLAB

Table 1: Result of normal class of HIDS for ten simulations in Genetic Algorithm

S.no.	Current Generation	Best Fitness
1	100	0.494271
2	98	0.486631
3	100	0.516763
4	99	0.527048
5	100	0.491253
6	98	0.453247
7	100	0.488476
8	99	0.508741
9	100	0.516352
10	100	0.514952

Table 2: Result of anomaly class of HIDS for ten simulations in Genetic Algorithm

S.no.	Current Generation	Best Fitness
1	63	15.9241
2	52	15.9209
3	54	15.9442
4	56	15.9296
5	73	15.8964
6	72	15.9294
7	63	15.9688
8	72	15.9075
9	57	15.9275
10	70	15.8924

Table 3: Result of Normal class of NIDS for ten simulations in Genetic Algorithm

S.no.	Current Generation	Best Fitness
1	100	6.35758
2	100	6.02847
3	100	7.22437
4	100	7.64742
5	100	6.96042
6	100	6.27029
7	100	6.43669
8	100	6.98617
9	100	7.00398
10	100	4.96558

Table 4: Result of Anomaly class of NIDS for ten simulations in Genetic Algorithm

S.no.	Current Generation	Best Fitness
1	99	41.6248
2	100	41.3976
3	99	41.7089
4	100	41.7071
5	100	42.5959
6	100	41.8324
7	99	43.0958
8	100	42.0743
9	100	41.4164
10	100	42.3338

Table 3 and Table 4 contain the output generated by the Phase 2 in form of best fitness values for ten runs of the NIDS implementation. Table 3 presents the best fitness values for Normal class as detected by Fuzzy Genetic based NIDS and Table 4 lists the best fitness values for Anomaly class in NIDS implementation. Best fitness for the Normal and Anomaly class of the Fuzzy Genetic NIDS are 4.96 and 41.62 respectively.

From the result tables, it has been observed that the performance of the HIDS and NIDS can be differentiated on the basis of their effectiveness in the detection of normal and anomaly class. On comparing the results of detection of normal class (Table 1 & Table 3) and anomaly class (Table 2 & Table 4) of HIDS and NIDS, it has been found that the value of best fitness of HIDS is smaller than the best fitness value of NIDS. It has been also observed that the convergence of HIDS takes less number of generations than NIDS. This implies that in the detection of the normal and anomaly class, Fuzzy Genetic implementation of HIDS shows better results than NIDS.

VIII. CONCLUSION AND FUTURE WORK

In this paper, anomaly detection systems HIDS and NIDS have been implemented using Fuzzy Genetic approach and their performance has been relatively compared with respect to the best fitness values obtained. By using hybrid Fuzzy Genetic Approach, optimized results of anomaly detection are obtained over the user behaviour of the single machine/host and on the network. It has been observed that the normal and the anomalous behavior are better detected in HIDS as compared to NIDS. For the future direction, the Normal and Anomaly class can be better detected by the combination of both the network and the host by using the various machine learning algorithms.

References

[1] K. Hanumantha Rao, G. Srinivas, Ankam Damodhar and M. Vikas Krishna, "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms," International Journal of Computer Science and Telecommunications, vol. 2, no. 3, 2011.

[2] Jiankun Hu, "Host-Based Anomaly Intrusion Detection," Handbook of Information and Communication Security, Springer Berlin Heidelberg, pp. 235-255, 2010.

[3] Giovanni Vigna, Christopher Kruegel, " Host Based Intrusion Detection," Handbook of Information Security, H. Bigdoli , December 2005.

[4] Sanoop Mallissery, Jeewan Prabhu, and Raghavendra Ganiga, "Survey on Intrusion Detection Methods," in proc. of 3rd Int. Conf. on Advances in Recent Technologies in Communication and Computing, Bangalore, 2011, pp. 224-228.

[5] Asmaa Shaker Ashoor and Sharad Gore, "Intrusion Detection System: Case study," in International Conference on Advanced Materials Engineering, vol. 15, Singapore, 2011, pp. 6-9.

[6] V Jyothsna, V V Ramaprasad, and K Munivara Prasad, "A Review of Anomaly based Intrusion Detection Systemss," International Journal of Computer Applications, vol. 28, no. 7, pp. 26-35, August 2011.

[7] P Garcia Teodoro, J Diaz Verdejo, G Macia Fernandez, and E Vazquez, "Anomaly based network intrusion detection: Techniques, Systems and Challenges," International Jouirnal of Computers and Security, vol. 28, no. 1, pp. 18-28, February-March 2009.

[8] Bharanidharan Shanmugam and Norbrik Bashah Idris, "Hybrid Intrusion Detection Systems(HIDS) using Fuzzy Logic," in Intrusion Detection Systems, Dr. Pawel Skrobanek, Ed. Croatia, Europe: InTech, 2011, ch. 8, pp. 135-155.

[9] Om Prakash Shukla, Amit Bahekar, Jaya Vijayvergiya, "Effective Fault Diagnosis and Maintenance Optimization by Genetic Algorithm," Research Expo International Multidisciplinary Research Journal, vol.2, no.2, pp. 20-25, 2012.

[10] Rajdeep Borgohain, "FuGeIDS: Fuzzy Genetic paradigms in Intrusion Detection Systems," International Journal of Advanced Networking and Applications, vol. 3, no. 6, pp. 1409-1415, 2012.

[11] Mathworks, Genetic Algorithm and Direct search toolbox For use with Matlab. Natick, United States: Mathworks Inc., version 1, 2004.

[12] J. T. Yao, S.L. Zhao, L.V. Saxton, "A study on fuzzy intrusion detection," Proc. of SPIE Data Mining, Intrusion Detection, Information Assurance, And Data Networks Security, Orlando, Florida, USA , vol. 5812, pp. 23-30, 2005.

[13] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," proc. of IEEE sympo. On computational Intelligence in Security and Defence Applications, 2009.

[14] Radhika Goel, Anjali Sardana, and Ramesh C. Joshi, "Parallel Misuse and Anomaly Detection Model," International Journal of Network Security, vol.14, no.4, pp. 211-222, July 2012.

[15] Mostaque Md. Morshedur Hassan, "Current Studies on Intrusion Detection System, Genetic Algorithm and Fuzzy Logic," International Journal of Distributed and Parallel Systems (IJDPS), vol. 4, no. 2, pp. 35-47, March 2013.

[16] Harjinder Kaur and Nivit Gill, "Host based Anomaly Detection using Fuzzy Genetic Approach (FGA)," International Journal of Computer Applications (IJCA), vol. 74, no. 20, pp. 5-9, July 2013.