Genetic algorithm for intrusion detection system in computer network

Hamizan Suhaimi[1], Saiful Izwan Suliman[2], Afdallyna Fathiyah Harun[3], Roslina Mohamad[4], Yuslinda Wati Mohamad Yusof[5], Murizah Kassim[6]
[1,2,4,5,6]Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia
[3]Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, Malaysia

Article Info

Article history:

Received Jan 3, 2020
Revised Mar 15, 2020
Accepted Apr 7, 2020

Keywords:

Computer network systems
Genetic algorithm
Intrusion detection system,
KDD’99

ABSTRACT

Internet connection nowadays has become one of the essential requirements to execute our daily activities effectively. Among the major applications of wide Internet connections is local area network (LAN) which connects all internet-enabled devices in a small-scale area such as office building, computer lab etc. This connection will allow legit user to access the resources of the network anywhere as long as authorization is acquired. However, this might be seen as opportunities for some people to illegally access the network. Hence, the occurrence of network hacking and privacy breach. Therefore, it is very vital for a computer network administrator to install a very protective and effective method to detect any network intrusion and, secondly to protect the network from illegal access that can compromise the security of the resources in the network. These resources include sensitive and confidential information that could jeopardise someone’s life or sovereignty of a country if manipulated by wrong hands. In Network Intrusion Detection System (NIDS) framework, apart from detecting unauthorized access, it is equally important to recognize the type of intrusions in order for the necessary precautions and preventive measures to take place. This paper presents the application of Genetic Algorithm (GA) and its steps in performing intrusion detection process. Standard benchmark dataset known as KDD’99 cup was utilized with forty-one distinctive features representing the identity of network connections. Results presented demonstrate the effectiveness of the proposed method and warrant good research focus as it promises exciting discovery in solving similar patent of problems.

Corresponding Author:

Saiful Izwan Suliman,
Faculty of Electrical Engineering,
Universiti Teknologi MARA, Shah Alam, Malaysia.
Email: nazwi81@gmail.com

1. INTRODUCTION

In the past decades, Internet applications are among the major booms in the society with the advancement of communication technology. This allows fast development of hardware, software and equipment to facilitate and ease daily work. However, this technology growth has not only contributed to the good sides of our activities, but has also led to many problems when fallen to the wrong hands. For computer network that connects all hardware in a local area network, unauthorised access to its resources is the main problem due to the borderless concept [1, 2]. Anyone can illegally access our computer network if they have the skill to defeat the security system installed. This issue highlights the need for Intrusion Detection System (IDS) in order to protect a computer network from illegal access. This system includes software and hardware that will monitor network traffic to detect malicious attacks such as unusual traffic connection or connection that breaks security and certain predetermined policies. There are two categories of IDS which are signature-based IDS and anomaly-based IDS [3-5]. These two systems vary in terms of configuration, detection and cost. A computer network intruder may modify traffic connection to escape signature-based IDS detection system installed [6]. However, this may not be the case for anomaly-based IDS as it can detect slight change in network pattern to further analyse the situation. The drawback of signature-based IDS is that only unwanted traffic from a set of known traffic can be detected. This type of IDS will normally encounter problem if a new attack connection intrudes a network as it does not store the pattern before [7]. Because of this limitation, signature-based IDS is less effective compared to other IDS as it is dependent on past encounters of attack connections [8, 9]. Additionally, the main database which stores the known attack connections can sometimes contain error traffics due to bad detections. This can lead to the increase of false positive rate [10]. This study was conducted to overcome the afore-mentioned issue by applying a different approach. This is by increasing the possibility of getting accurate detection, extra features in each traffic connection were identified and added into the “chromosomes” or candidate solutions. Additionally, parameters tuning process which determines the optimal value for each variable was also given extra focused as it will determine the final outcomes of the experiments.

The idea of deploying the Leader Based Intrusion Detection System (LBIDS) into access network was proposed by Rajkumar and Vayanaperumal [11]. This system was implemented in order to detect and prevent DOS such as Sybil and Sinkhole. By implementing the simulation in NS2 software, the system was simulated based on three core security challenges such as authentication, preventing DOS attack and positive incentive provision. The outcomes show that they were able to fulfill the quality of service in the network by using the proposed method. IDS system that can minimize false positive detection by GA was proposed by Narsingyani and Kale [12]. In their research, KDD’99 cup dataset was used as main data in order to experiment the detection system. The main categories attacks focused in the experiment were duration, protocol, service, flag, source byte, destination byte and attack-name. Java language was implemented in the system where it was built on third party software package JGAP or GA/GP java toolkit. The results show that false positive rate can be reduced by increasing the number of rules in training data.

The study about anomaly-based IDS using the hybrid GA and K-Centroid Clustering was conducted by Chakrabarty et al. [1]. In the study, both KDD’99 Cup and NSL-KDD dataset were utilized to run the experiment for IDS. NSL-KDD dataset was investigated as it can be used to solve the problems in KDD’99 dataset. Compared to KDD’99 dataset, it was an effective dataset that can create clear comparison between normal data and intrusion connection in developing IDS. As for GA, it was performed on the specific clusters. Cluster was created to differentiate groups for a specific intrusive and normal data. When carrying out clustering, different clusters were also constructed for a similar type of connection. Based on the final results, approximately 86% and 96% accuracy rate for KDD’99 cup and NSL-KDD dataset were produced respectively. A research by Bhattacharjee P. S. et. al from Assam University, India proposes the variable fitness function in GA for IDS using NSL-KDD dataset [2]. Three different fitness functions were used which are vectorized-fitness function, weighted vectorized-fitness function and fuzzy weighted vectorized-fitness function. From the different fitness functions that were used, the following results were recorded:

Table 1 illustrates that Fuzzy Weighted Vectorized-GA (FV-GA) produces better fitness point which is the closest to 1. The number of attacks that can be detected were dominated by FV-GA where it represents the highest number of attacks detected from the simulation. As a conclusion, the proposed FV-GA performances was better than V-GA and WV-GA. In 2014, Sharma, Kumar and Kaur propose the hybrid of GA and Fuzzy Logic approaches for IDS [13]. NSL-KDD was used as the main dataset since KDD’99 dataset has redundancy issues, where the numbers of training and testing data in NSL-KDD are enough to produce good attack detection. In the study, the proposed method provides better performance than a single GA technique.

Table 1. Attack detection results using NSL-KDD datasets for VGA, WV-GA and FV-GA

Method	Fitness Point	Normal	Type of Attack			
			Dos	Probe	R21	U2R
VGA	0.9652 (f1)	928	143	329	8	0
VGA	0.9340 (f2)	639	98	380	8	0
WV-GA	0.9547 (f1)	849	115	372	13	0
WV-GA	0.9186 (f2)	493	91	512	7	0
FV-GA	0.9902 (f1)	1733	577	111	0	0
FV-GA	0.9918 (f2)	1733	577	111	0	0

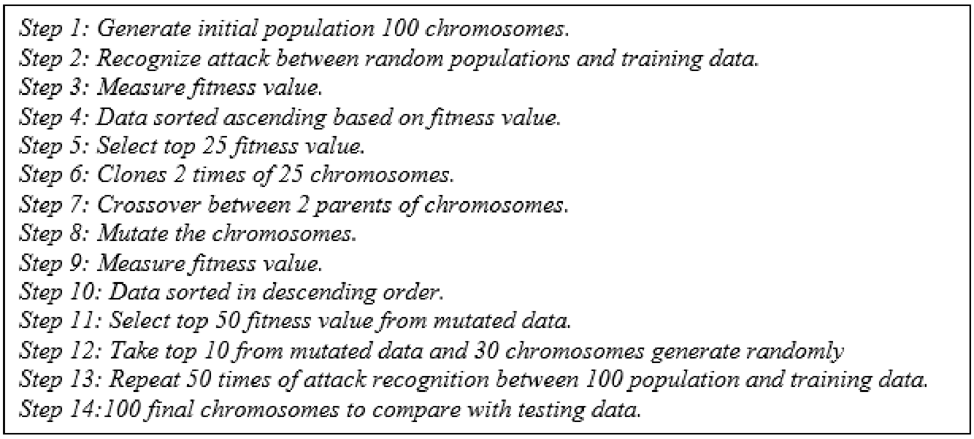
A new feature selection IDS using hybrid GA-SVM was conducted by Gharaee and Hosseinvand [14]. A new fitness function was introduced along with Least Squares Support Vector Machine (LSSVM). The experiment was tested on KDD’99 Cup and UNSW-NB15 datasets. The proposed method involves three main steps: feature selection based on GA, training and classification. Training of the dataset is the first step in the algorithm. The traffic data then classified into normal or anomaly class. In the paper, a new fitness function was introduced based on three parameters: False Positive Rate (FPR), True Positive Rate (TPR) and number of selected features to calculate each subset of features. In generating results, MMIFS model was applied using KDD'99 Cup into LSSVM model in order to compare with other methods such as GA-Fuzzy SVM, MMIF, SVM and C4.5. The final results were produced based on their proposed method using UNSW-NB15 dataset in terms of accuracy detection rate, TPR and FPR. From the whole results, it shows that the proposed method performs better selection for the training data and indirectly provides the improvement in detection intrusion.

The study about developing a network security for cloud computing using IDS was done by Singh and Hazela [15]. Cloud computing also have the similar risk such as risk of unauthorized modification of data, risk of theft and risk of loss. In the study, a distributed multilayer architecture in Cloud Computing environment was introduced for providing intrusion detection and to perform complex event correlation analysis by using distributed security components. A further study about GA and Neural Networks was made in order to design the architecture based on the knowledge. Feature selection technique based on GA was used to select suitable subset of features. As addition, two types of chromosome were proposed based on different criteria which are bandwidth of the resources and created based on the job length. The main purposes in this study is to allocate the most suitable capacities for the job length and resources based on the bandwidth. The final results show that their cloud IDS were able to detect cloud attacks by 57% of random sets of cloud attacks.

Signature matching algorithm is one of the methods that can be used to identify attacks especially from the internal system. Desai and Gaikwad proposed a hybrid method to identify attacks through the internal and external of network system [16]. FGA was implemented to detect external attacks in the network. SQL injection became the main focus in the internal system where it consists of static and dynamic injection detections. In static part, whole query is compared with the stored signature and classified as malicious if it matches. Meanwhile, for dynamic detection, calculation of similarity index from comparison was made and considered as malicious if it exceeds the threshold value.

2. RESEARCH METHOD

Intrusion detection is the process of recognizing the attacks in the traffic between two or more resources where the entire network becomes vulnerable to third parties. The attacks can happen when the security breach towards network has occurred [17-19]. Techniques that involved in proposed GA system is discussed below:



Step 1: Generate initial population 100 chromosomes.
Step 2: Recognize attack between random populations and training data.
Step 3: Measure fitness value.
Step 4: Data sorted ascending based on fitness value.
Step 5: Select top 25 fitness value.
Step 6: Clones 2 times of 25 chromosomes.
Step 7: Crossover between 2 parents of chromosomes.
Step 8: Mutate the chromosomes.
Step 9: Measure fitness value.
Step 10: Data sorted in descending order.
Step 11: Select top 50 fitness value from mutated data.
Step 12: Take top 10 from mutated data and 30 chromosomes generate randomly
Step 13: Repeat 50 times of attack recognition between 100 population and training data.
Step 14:100 final chromosomes to compare with testing data.

Figure 1. Genetic Algorithm in NIDS

Figure 1 shows the steps involved in the implementation of Network Intrusion Detection System (NIDS) using GA for both training and testing process.

A. Separate main data into training and testing data set

Before the start of the training process, the raw dataset which consists of 311,030 connection data was divided into two parts. This selected data was used either for training or testing based on predetermined probability of selection. The ratio for the separation of these 2 processes was conducted based on 90:10 portion, in which the larger bulk utilized in the training process.

B. Generate Initial Population

In the first steps, initial population was generated randomly which consists a set of chromosomes (candidate solution). Each chromosome stores genes that is associated with the problem domain. In this study, it will be the properties of each traffic connection. This population will be mainly used during the training process. The size of the population was heuristically determined in the initial stage of the study to identify the suitable value that gives optimal solutions.

C. Fitness Function

In this paper, the proposed fitness function is given by the formula (1):

$$F = \frac{a}{A} - \frac{b}{B} \tag{1}$$

Where a represents the number of attacks detected by the candidate solutions while A is the total number of attacks in the training dataset. Variable b is the number of normal connections that were detected by the proposed method in an iteration, while B represents the total normal connections in the dataset. A candidate solution (chromosome) is considered of high quality if the fitness value closes to 1 [20-22].

D. Cloning

From the overall population, a set of chromosomes with high fitness value will be selected for the cloning process. During this process, the selected chromosomes will be duplicated a number of times. This is done to increase the probability of getting better solutions [23].

E. Crossover

Figure 2 illustrates the process of crossover between 2 chromosomes that were selected randomly from two parents (cloned chromosomes). Crossover will take place based on the crossover rate. This rate which is generated randomly will determine the number of features that will crossover between these chromosomes [24].

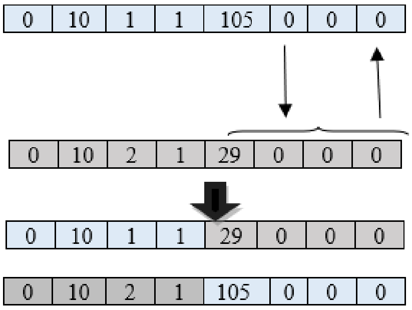


Figure 1. Process of crossover

F. Mutation

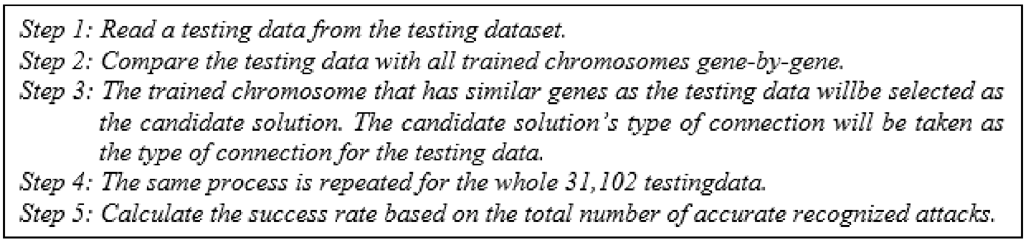
Mutation takes places between the two crossover chromosomes. This process will only be executed based on predetermined probability rate. This rate cannot be too large as it could alter the whole genes and eventually produce a new offspring [25]. During this process, only a slight change to the genes is necessary so that it will still carry most of the properties of the parents added with the enhanced gene(s) through the mutation process.

G. Selection

The mutation process conducted has altered the genes of a chromosome. Therefore, as shown in (1) will be utilized again to calculate the new fitness value of the chromosome. The mutated chromosomes were then sorted based on the fitness value. Top ten mutated chromosomes were selected as a part of the new population for the next iteration. These chromosomes will be combined with fifty chromosomes from the initial population and thirty chromosomes which were generated randomly. This training process was repeated for fifty iterations and the final population of chromosomes produced will undergo the next phase which is testing process.

H. Testing

Figure 3 shows the process of testing which involves 100 chromosomes from the final population of the training stage. These chromosomes were used to compare with each of the 31102 testing data to predict the type of connection whether it is an attack or normal traffic. Success rate was determined based on the total number of testing data that is correctly predicted by the proposed method.



Step 1: Read a testing data from the testing dataset.

Step 2: Compare the testing data with all trained chromosomes gene-by-gene.

Step 3: The trained chromosome that has similar genes as the testing data willbe selected as the candidate solution. The candidate solution's type of connection will be taken as the type of connection for the testing data.

Step 4: The same process is repeated for the whole 31,102 testingdata.

Step 5: Calculate the success rate based on the total number of accurate recognized attacks.

Figure 2. Steps in testing process of IDS

3. RESULTS AND DISCUSSION

Two important variables in Genetic Algorithm that influence the final result in this study are the mutation and crossover rates [26]. In order to get the optimal values, many experiments were conducted utilizing different values of these two variables. The results of this experiments were further analyzed by calculating the True Positive Rate (TPR) value. This value determines the success rate of the prediction of the proposed algorithm. Perfect prediction of intrusion detection will give 100% TPR value.

Table 2 shows the results for intrusion detection based on three crossover rate investigated in this study. From total of 31102 testing data, experiment with crossover rate of 10 produces the best solution with 25942 testing data accurately predicted. This is followed by systems with crossover rate of 15 and 5 with 25883 and 21003 data correctly predicted respectively. This result is further illustrated in Figure 4. As shown in the figure, crossover rate of 10 gives the best result with 83.41% of TPR. This is followed by 15 and 10 crossover rates with 83.22% and 67.53% TPR respectively.

Table 2. Intrusion detection result based on crossover rate

Crossover Rate	Number of Attack Accurately Predicted	Number of Attack Inaccurately Predicted
5	21003	10099
10	25942	5160
15	25883	5219

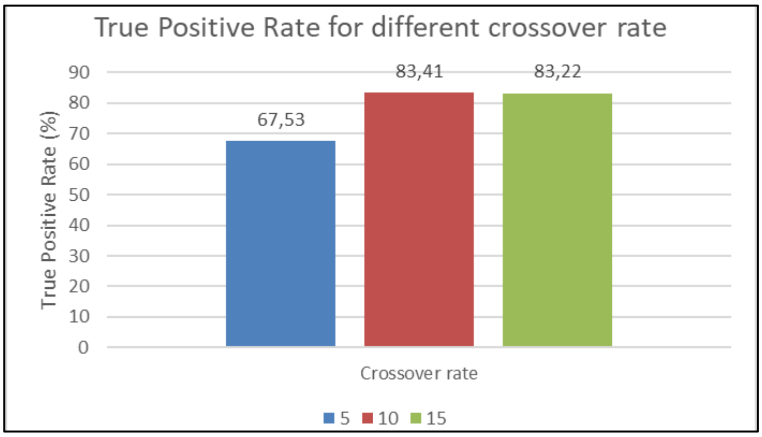


Figure 4. TPR based on different crossover rate

For the probability of mutation, three values were investigated which are 0.1 (10%), 0.15 (15%) and 0.20 (20%). These probability rates were utilized in order to determine whether the process of altering the gene(s) in each chromosome will be performed or not. This process was performed to enhance the fitness value of each chromosome even though it is not always the case. It could also lead to a worse candidate solution. The results for the three different probability of mutations are shown in Figure 5. As can be seen, the best result was obtained when 0.1 was utilized as the probability of mutation with 83.41% TPR value. This is followed by 0.2 and 0.15 with 83.27% and 83.2% TPR respectively. This outcome coincides with findings from previous works by other researchers in which the best solution was found when the mutation rate is relatively small which is between 0.07 – 0.1.

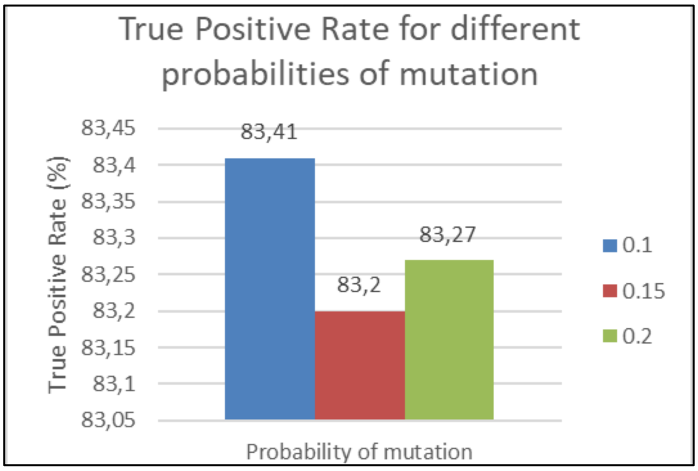


Figure 5. TPR for 0.1, 0.15 and 0.20 probabilities of mutation

4. CONCLUSION

Genetic Algorithm (GA) is among the powerful tool to solve many complex problems such as optimization, learning, prediction and identification. This algorithm depends on several important steps and variables such as crossover and mutation that will strongly influence the final results. These variables are known as genetic operators and extensively investigated in this study. In general, Genetic Algorithm needs mutation and crossover process in order to modify and produce new offspring that carry most genes of the original chromosomes. Based on the presented results, combination of crossover rate of 10 genes and 0.1 probability of mutation produced the best result. From total of 31102 testing data, it has managed to accurately predict 25942 attack connections which represents 83.41% True Positive Rate (TPR). Further study can be carried out in order to increase this rate to the acceptable level of at least 95%. This can be implemented by hybridizing GA with other method such as local search to further explore the search space to get a better prediction.

ACKNOWLEDGEMENTS

The authors would like to express the gratitude to the Ministry of Education, Malaysia and Universiti Teknologi MARA, Shah Alam, Selangor for the financial support given for this project (FGRS Grant) [File No: 600-IRMI/FRGS 5/3 (0007/2016)].

REFERENCES

[1] B. Chakrabarty, O. Chanda, M. Saiful. “Anomaly based Intrusion Detection System using Genetic Algorithm and K-Centroid Clustering”, *Int J Comput Appl*., vol. 163, no. 11, pp. 1-13, 2017.

[2] P. Bhattacharjee, A.K.M. Fujail, S.A. Begum, S. A. “Intrusion detection system for NSL-KDD data set using vectorised fitness function in genetic algorithm”, *Adv. Comput. Sci. Technol*, vol. 10, no.2, pp. 235-246, 2017.

[3] S. Mabu, C. Chen, L. Nannan, S. Kaoru, and H. Kotaro, "An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 1, pp. 130-139, 2010.

[4] M.S.A. Khan, "Rule based network intrusion detection using genetic algorithm." *International Journal of Computer Applications*, vol.18, no. 8, pp. 26-29, 2011.

[5] S.T. Powers, and H. Jun, "A hybrid artificial immune system and Self Organising Map for network intrusion detection." *Information Sciences*, vol. 178, no. 15, pp. 3024-3042, 2008.

[6] M.M.M. Hassan, "Network intrusion detection system using genetic algorithm and fuzzy logic." *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, no. 7, 2013.

[7] S.T. Powers, and H. Jun, "A hybrid artificial immune system and Self Organising Map for network intrusion detection." *Information Sciences*, vol. 178, no. 15, pp. 3024-3042, 2008.

[8] H. Jiang, and R. Junhu, "The application of genetic neural network in network intrusion detection." *JCP*, vol. 4, no. 12, pp. 1223-1230, 2009.

[9] A. Zainal, M.A. Maarof, and S.M. Shamsuddin, "Ensemble classifiers for network intrusion detection system." *Journal of Information Assurance and Security*, vol. 4, no. 3, pp. 217-225, 2009.

[10] W.A.N.G. Xiaoqiang, "Study on genetic algorithm optimization for support vector machine in network intrusion detection." *Advances in Information Sciences and Service Sciences*, vol. 4, no. 2, pp. 282-288, 2012.

[11] D.U. Rajkumar & R. Vayanaperumal, “A leader based intrusion detection system for preventing intruder in heterogeneous Wireless sensor network”, *In Bombay Section Symposium (IBSS)*, IEEE, pp. 1-6, 2015.

[12] D. Narsingyani & O. Kale, “Optimizing false positive in anomaly based intrusion detection using Genetic algorithm”, In *Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference*, pp. 72-77, 2015.

[13] S. Sharma, S. Kumar, M. Kaur, “Recent trend in Intrusion detection using Fuzzy-Genetic algorithm”, *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 5, 2014.

[14] H. Gharaee, H. Hosseinvand, “A new feature selection IDS based on genetic algorithm and SVM”, In *Telecommunications (IST), 2016 8th International Symposium on*, pp. 139-144, 2016.

[15] P. Singh, B. Hazela, “Design & Development of a new hybrid system to Prevent Intrusion at cloud using genetic algorithm”, *International Journal of Advance Research in Computer Science and Management Studies*, vol. 4, no. 6, 2016.

[16] A.S. Desai & D.P. Gaikwad, “Real time hybrid intrusion detection system using signature matching algorithm and fuzzy-GA”, *In Advances in Electronics, Communication and Computer Technology (ICAECCT)*, 2016 IEEE International Conference on, pp. 291-294, 2016.

[17] S.N. Pawar, and R.S. Bichkar, "Genetic algorithm with variable length chromosomes for network intrusion detection." *International Journal of Automation and Computing*, vol. 12, no. 3, pp. 337-342, 2015.

[18] K. Shafi, and H.A. Abbass, "Evaluation of an adaptive genetic-based signature extraction system for network intrusion detection." *Pattern Analysis and Applications*, vol. 16, no. 4, pp. 549-566, 2013.

[19] N. Naidu, and R. V. Dharaskar, "An effective approach to network intrusion detection system using genetic algorithm." *International journal of computer applications* vol. 1, no. 3, pp. 26-32, 2010.

[20] P. Tao, Z. Sun, and Z. Sun, "An improved intrusion detection algorithm based on GA and SVM." *IEEE Access*, vol. 6, pp. 13624-13631, 2018.

[21] J. Ghosh, D. Kumar, and R. Tripathi, "Features Extraction for Network Intrusion Detection Using Genetic Algorithm (GA)." In *Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough*, pp. 13-25. Springer, Cham, 2020.

[22] H. Suhaimi, S.I. Suliman, I. Musirin, A. Harun, R. Mohamad, M. Kassim, and S. Shahbudin, "Network intrusion detection system using immune-genetic algorithm (IGA)." *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 2, pp. 1059-1065, 2020.

[23] Farhaoui, Yousef. "Intrusion prevention system inspired immune systems." *Indones. J. Electr. Eng. Comput. Sci*, vol. 2, no. 1, pp. 168-179, 2016.

[24] S.I. Suliman, M.S.A. Shukor, M. Kassim, R. Mohamad, and S. Shahbudin. "Network Intrusion Detection System Using Artificial Immune System (AIS)." In *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pp. 178-182. IEEE, 2018.

[25] H. Suhaimi, S.I. Suliman, I. Musirin, A. Harun, R. Mohamad, "Network intrusion detection system by using genetic algorithm." *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1593-1599, 2019.

[26] N.A. Azeez, and A.B. Babatope, "AANtID: an alternative approach to network intrusion detection." *Journal of Computer Science and Its Application*, vol. 23, no. 1, pp. 129-143, 2016.