Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 12, Number 1, Spring 2023, pp. 29-40
Received: 06/06/2020, Revised: 07/04/2021, Accepted: 08/31/2021

Heuristic intrusion detection technique based on nonlinear regression and sigmoid function

Shahriar Mohammadi[1]*, Mehdi Babagoli[2]
1- *Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran
2- Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran

[1]mohammadi@kntu.ac.ir, [2]Mehdi.babagoli@email.kntu.ac.ir

Corresponding author's address: Shahriar Mohammadi, Faculty of Industrial Engineering, KN Toosi University of Technology, Tehran, Iran.

Abstract- The expansion of Internet technologies during the last decades has led to the dependence of user's activities in cyberspace on services provided by computer networks. One of the most important services is Intrusion Detection System (IDS) which controls network traffic for detecting abnormal behavior as well as anomaly activities. The robustness of the IDS is considered as an essential issue in the networks. In this paper, a brand-new model based on meta-heuristic algorithms is projected to detect abnormal packets. In order to develop a high-performance strategy, a benchmark dataset (NSL-KDD), high-accuracy feature selection method and four meta-heuristic algorithms are employed. The dataset consists of 150490 normal and abnormal packets which are captured from a military network connection, and 16 most important features are extracted among 41 features using wrapper feature selection method. The mentioned feature selection method uses the naïve-bayesian approach to evaluate feature subsets. After the feature selection process, four meta-heuristic algorithms are utilized to detect the anomalies in network. The parameters of the cost function (a combination of non-linear regression and sigmoid) are optimized using meta-heuristic algorithms. The experimental results show that the imperialist competitive algorithm (ICA) outperforms other implemented meta-heuristic algorithms in terms of accuracy.
Keywords- network security, intrusion detection system, meta-heuristic, Naïve-Bayesian, nonlinear regression, sigmoid function

I. INTRODUCTION

Creating effective defense in cyberspace security issues has become a big problem in recent decades. Regarding that establishing computer network without security failure is impossible, intrusion detection and prevention systems are considered as an essential issue in cyberspace security scope. An Intrusion Detection System (IDS) monitors the system activity, analyzes the network traffics and reports on observation of any security violations. In comparison to other network security techniques, IDS are able to identify sources of attack in addition to detecting and reporting[1]. Basically, IDS can be categorized as signature-based and anomaly-based detection. Signature-based IDS generates noteworthy information for specified, well-known attacks along with alarm [2]. On the contrary, anomaly IDS might identifies the linkage stream which is considered as malicious and having a potential detection in unseen intrusion events is beneficial [3]. Figure 1 shows the generic architecture of IDS.

Correspondingly, categorizing the IDS can be done based on the structure of the protective system: Host-based IDS (HIDS) and Network-based IDS (NIDS).

HIDS only protects the endpoint and mainly related to the operation system (OS) information (like mobile, workstation and server), whereas NIDS analyzes the network related information and sits on the ingress points of network to monitor the traffic and detects the malicious activity (Fig. 2).

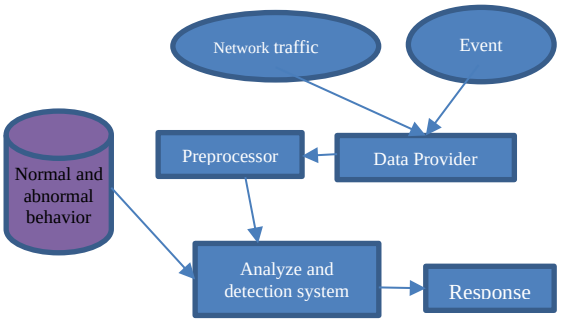

Fig. 1. IDS architecture

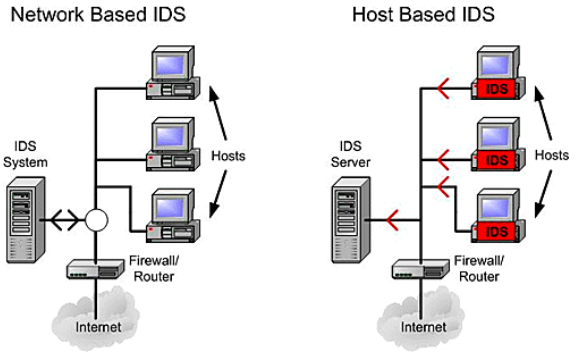

Fig. 2. NIDS vs. HIDS

Polymorphic mechanisms are used by intruders to dissimulate the attack payload and scape from the detection systems. In order to boost the efficiency of IDS, a lot of supervised and unsupervised learning approaches from the field of machine learning and pattern recognition have been used [4]. Current IDSs analyze all data characteristics to detect intrusion or misused patterns. These features are used as the basic knowledge of machine learning algorithms. Some of the features contain duration of the connection, protocol type, outbound commands in FTP session and so on [5]. In this paper, a model based on meta-heuristic algorithms is proposed for detecting intrusion detection system.

Optimization algorithms are divided into exact and approximate algorithms. Exact algorithms can find the best solution but they are not efficient because the complexity of algorithms increases exponentially when the size of input data grows [6]. On the other hand, the approximate algorithms never guarantee to find the best solution. Indeed, the aim of this method is to investigate search spaces, using exploration and exploitation to obtain near-optimal solution. Approximate algorithms are time efficient because the required time to find the best solution is polynomial [7]. As shown in Fig. 3 there are three types of approximate algorithms:

1) Hyper- Heuristic [8, 9]
2) Meta-Heuristic [10, 11]
3) Heuristic [12]

Hyper-heuristic research contains several hierarchical structures. Although, they all emphasis on the prevalent aim of automating the design and adapting of heuristic methods to solve the complicated computational search issues [8]. Less computational effort in optimization algorithms, iterative methods and simple heuristic is dispelled with the solutions which can be founded by Meta-heuristics. Meta-heuristic algorithms are designed to solve multi-objective optimization problems since they can discover several optimal solutions in a single run [13]. The differences between heuristic and meta-heuristic algorithm are listed as below:

- Heuristics methods are often problem-dependent, indeed, it can be defined for a specific problem. Meta-heuristics are problem-independent techniques that can be applied to a various range of problems.
- Meta-heuristics are type of heuristics method, but a more robust one, since a mechanism to avoid trapping in a local minimum is present in any meta-heuristic.
- In NP-hard problems, meta-heuristic algorithms are more applicable than heuristic algorithms.

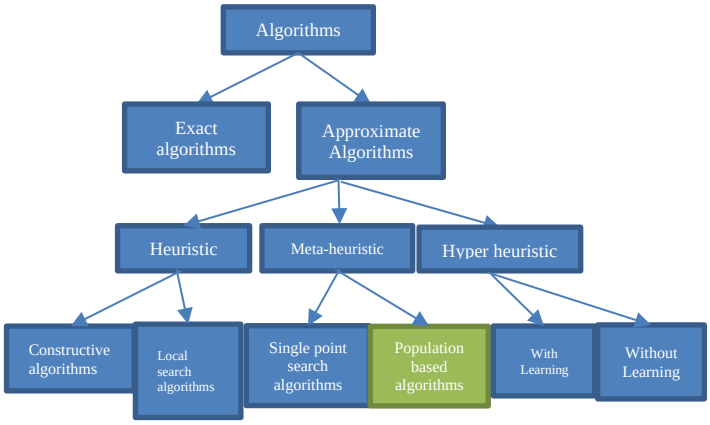

Fig. 3. Clear classification of algorithms

The mentioned algorithms can use the dataset to create a model for optimization problems. In this study, the NSL-KDD dataset [14] is used. The dataset comprised of 41 features, 125973 train instances and 22544 test instances. Each record in train dataset labeled as abnormal or normal traffic. We utilized feature selection methods to extract most important features from dataset and delete the irrelevant attributes. After feature selection procedure, a model based on combination of nonlinear regression (NR) and sigmoid function is used to detect abnormal traffic and the parameters of this model are optimized using imperialist competitive algorithm (ICA), ant colony optimization (ACO), genetic algorithm (GA) and particle swarm optimization (PSO).

Our work is organized as follows. The literature review is discussed in Section 2. In Section 3, overview of the proposed model and its prototype with the detail of its phases such as normalization, feature selection and detail of implemented meta-heuristic algorithms are presented. Section 4 discusses the experimental result, and finally, section 5 ends this paper with conclusions and future work directions.

II. LITERATURE REVIEW

In this section an overview of IDS architecture and some applicable methods to detect and prevent the different type of intrusions are discussed. Generally, intrusion detection methods are classified in three categories: anomaly-based models, Signature based models and hybrid models. The details of these methods are described below.

A. Anomaly-based models

Statistical-based system (SBID) is the most important technique of the anomaly-based models. In anomaly based systems, the "normal" network activity is determined in the SBID system and then all traffic that falls outside the scope of the normal activity is marked as anomalous (not normal) [15]. This method is good for predicting both user and system activities. Anomaly-based techniques create a profile of users and their trustworthy activities to train a system and then using this profile to continuously monitor the network activities for suspicious activity. The mentioned model can be efficient for previously unknown attacks because it can consider all the suspicious behavior as an abnormal activity and raise an alarm. One of the main disadvantages of this model is false alarm rate (false positive), for instance, previously unknown legitimate traffic is detected as a malicious activity. Additionally, the extraction of an attack-free dataset for training a detector (except simulated data) is impractical [16]. Typically, network traffics contain a large number of port scans, denial-of-service attacks and backscatter, and worm activity. Incorrect training can lead to considering this activity as a part of the normal state for an anomaly detector. Brief overview of the recent state of the art research works in the field of intrusion detection are investigated bellows.

According to [17], a novel scheme is proposed to detect anomaly traffic based on neighbor outlier factor (NOF). Due to big volume of data in utilized dataset, distributed storage environment is used in this research for enhancing the conduct of intrusion detection systems. As it is shown in their results, the proposed method is incredibly better than other available machine learning approaches and it is capable of detecting almost all oddity data in acceptable consummation-time. According to [18] a new statistical-based intrusion detection system is conducted to evaluate distributed network systems. The proposed intrusion detector uses the Gaussian Markov random field distribution based on the hypothesis testing and employing the log-likelihood ratio criterion. The performance of proposed method has been measured with both the *Bhattacharyya distance* and *graph Laplacian matrices* of the consecutive time instants. According to research [19], a subspace for given data samples has been derived to distinguish the data with normal and abnormal profiles by using projection. However, the performance of this method highly depends on the choice of the employed subspace. Statistical anomaly detection schemes also have some drawbacks. Dexterous attackers can train a statistical anomaly detection to accept abnormal behavior as normal.

As shown in research [20], different hybrid meta-heuristic algorithms are used for feature selection. Among different implemented algorithms the multiverse optimizer- Bat algorithm (MVO-BAT) outperforms with better feature subsets, lower time consumption, and improve accuracy. Three classifiers are implemented to evaluate feature subsets and detect anomalies. As shown in their results, J.48 can reach better accuracy compared to SVM and Random Forest (RF). The required time for building the model is calculated for each feature selector and classifier. The final model proposed for IDS is evaluated on UNSW dataset and consists of MVO-BAT and J.48 algorithms. The accuracy, F-measure and time to build detection model obtained 92.80%, 94.34% and 1.24s, respectively.

According to [21], combination of SVM and modified GA is used for feature selection. In the detection phase, the artificial neural network (ANN) is used to detect intrusions. In order to improve ANN performance, two meta-heuristic algorithms are combined in the training phase. The proposed method is evaluated on the NSL-KDD dataset and obtained a 99.3% detection rate with the lowest training and testing time. One of the interesting achievements in this study is the number of features that they used for intrusion detection. As shown in their result, they used 4 features among 41 features and detect normal traffic, DoS, Probe, U2R and R2L traffic with the following rate:99.3%, 99.8, 98.6%, 98.9% and 99.10%.

B. Signature based model

Signature detection is investigating network traffic for a series of bytes or packet sequences known to be malicious. The signature-based network intrusion detection system (SNIDS) analyzes the information that is gathered and compares them to a large database of attacked signatures [22]. The advantage of this method is the precision in detecting intrusions which their patterns have been documented. These types of systems are in need of frequent rule-base updates and signature updates, and they cannot detect unknown attacks (called zero-day attacks). Another limitation of these methods is that the system engine performance decreases when the signatures keep on increasing. Accordingly, many intrusion detection engines

are deployed on systems with multi processors and multi Gigabit network cards [23]. Snort program is one of the most popular SNIDS that typically configured with a set of rules to detect popular attack patterns. A summary of influential research works in the field of SNIDS are described below.

Authors in [24] proposed an automatically generator of signature using honeycomb. Honeycomb is a host-based intrusion detection system which is using honey-pot to capture malicious traffic targeting dark space. This project applied Longest Common Substring (LCS) in its signature generation. According to their experiment, 649 TCP connections and 123 UDP connections of attack were made and by conducting the LCS algorithm, approximately 30 signatures were created. The created signature was passed to Snort IDS for evaluation and a Slammer worm was detected using the created signatures, which is an advantage over the original Snort signature. As mentioned above, the false alarm is the major limitation in the SNIDS. Authors in [25] investigated alarm processing techniques. They have described techniques like alarm mining and various correlation algorithms as a post processing method to decrease the false alarm. In [26] an alarm ranking technique known as M-Correlator is proposed. This method considers three types of information, namely 1) alarms from different security manners such as firewalls, 2. Network configuration (like port numbers, application/OS in execution), 3) some user defined parameters like criticality of applications, amount of interest in a specific type of attack and so on. The rank of an alarm can be generated as a result of correlation of information. Also, by utilizing the cluster algorithm to generate a consolidated list of alarm message, grouping the related alarms is totally possible.

In order to increase the efficiency of intrusion detection system, statistical anomaly detection engines can be added to the signature-based systems. This procedure leads to automatically detection of the unknown attacks and possible generation of a signature.

C. Hybrid model

Considering the limitations of both Signature-based and anomaly-based detection methods, hybrid models combined both methods to detect the malicious activity in networks. The main goal of this method is to increase the detection precision and decrease the false positive rate [27]. Hybrid models can possess different privileges of both algorithms and detect the breaches in networks, but utilizing different methods in single model can be a very challenging task [28]. Some major researches in case of hybrid intrusion detection system (HIDS) research has been developed in recent years[29]. proposed a novel HIDS method based on combination of neural network detection component and basic pattern matching engine are used to detect anomalies in the network traffic. In this research network traffic was monitored from Nepty (network traffic analyze in python) tool [30] and the results has proved that the proposed HIDS method is better than two individual methods. Authors in [28] used machine learning methods to detect real-time intrusions in network. The suggested model trained with KDDCUP999 dataset and then the monitored incoming traffics were analyzed in real-time. As it is shown in their results, the proposed method could reduce the false negative errors. According to [31], the virtual jamming attack on IEEE 802.11 network was investigated and the performance of the proposed method has been evaluated with multiple real scenario. More recently, according to importance of HIDS, intrusion detection in IoT (internet of things) domain have been investigated frequently [32, 33]. Considering the security of IoT communication, most researches modeled IDS as a hybrid system. The performance results of these researches proved that the efficiency of mentioned model is much better than anomaly-based or signature based methods. Summary of these categories are demonstrated in Table 1.

TABLE 1. BRIEF DESCRIPTION OF IDS CATEGORIES

	Description	**Detection model**
Anomaly based	Process of detecting harmful activities whenever the behavior of the system deviates from the normal behavior	Just detect normal behavior
Signature based	Process of detecting harmful activities based on known patterns of previous attacks	Detects "known" attacks that were used during training
Hybrid models	If it has rule for an attack, mark it as that attack OR if it does not have rule, according to its normal-learning engine, considered as normal.	Detect the new attacks in addition to known attacks

However, some of the main restrictions of the above-mentioned research are as follows: disability to detect the intrusions with constant precision, lack of adaptability to new attacks and disability to block or prevent the attacks. Due to IDS methods limitation new and more robust detection mechanisms need to be developed.

III. PROPOSED IDS METHOD

Intrusions are any threat or malicious activity that have direct impacts on availability, integrity and confidentiality. Network traffic consist meaningful information, which can be extracted from packets. According to dealing IDSs with plenty of information, one of the principal tasks of IDSs is to accumulate the best quality of features and remove the irrelevant ones [34]. The dataset used in this experiment is NSL_KDD and the features of dataset evaluated have been using Naïve-Bayesian method. The selected features are

considered as an input of the proposed model. Finally, the parameters of model are optimized by utilizing different meta-heuristic algorithms.

A. NSL_KDD dataset characteristic

The NSL_KDD is a benchmark dataset which has improved the KDD cup'99 by removing redundant and duplicate records. The dataset consists of 127953,22544 train and test records, respectively. Each record includes 41 features and the classes of records labeled as normal or anomaly [14]. The details of attributes are explained in Table 2.

B. Feature selection method

Research on the data used for training and testing the detection model has become a prime concern since better data quality can improve the offline intrusion detection. Feature selection methods are utilized to remove the unnecessary data with low information loss [5]. The process of these algorithms can be conducted with combination of a search technique for proposing new feature subsets, along with an evaluation measure which computes the merit of different feature subset.

TABLE 2. DETAILS OF FEATURES

Index	attribute	Description	Type
1	Duration	Length of connection	Continuous
2	protocol type	TCP, UDP, …	Discrete
3	Service	Destination service (ftp,telnet,…)	Discrete
4	Flag	Status of connection	Discrete
5	Source byte	Number of bytes from source to destination	Continuous
6	Destination byte	Number of bytes from destination to host	Continuous
7	Land	Is the source and destination address being same	Discrete
8	Wrong fragments	Number of wrong fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	Number of hot indicators	Continuous
11	Failed logins	Number of unsuccessful logins	Continuous
12	Logged in	Is logged in successfully?	Discrete
13	Number of compromised	Number of compromised conditions	Continuous
14	Root shell	Is a cmd with root account is running root	Continuous
15	Su attempt	Attempting to logged in with user credential	Continuous
16	Number of root	Number of root accesses	Continuous
17	Number of file creation	Number of file creation operations	Continuous
18	Number of shells	Number of shell prompts	Continuous
19	Number of access file	Number of operations on access control files	Continuous
20	Number of outbound cmd	Number of outbound commands in FTP sessions	Continuous
21	Is host login	Is the login on the host login list?	Discrete
22	Is guest login	Is the guest logged into the system?	Discrete
23	Count	Number of connections to the same host	Continuous
24	Server count	Number of connections to the same service as the current connection in the past two seconds	Continuous
25	Serror rate	Percentage of connection with SYN error	Continuous
26	Srv_serror rate	Percentage of connection with SYN error	Continuous
27	Rerror rate	Percentage of connection with REJ error	Continuous
28	Srv_error rate	Percentage of connection with REJ error	Continuous
29	Same_srv rate	Percentage of connection to the same service	Continuous
30	Dif_srv rate	Percentage of connection to the different service	Continuous
31	Srv_diff_host rate	Percentage of connection to the different host	Continuous
32	Dst_host count	Number of connections to the same destination	Continuous
33	Dst_host_srv count	Number of connections to the same destination with same services	Continuous
34	Dst_host_same_srv rate	Percentage of connection to the same destination with same services	Continuous
35	Dst_host_diff_srv rate	Percentage of connection to the different destination with same services	Continuous
36	Dst_host_same_src_port rate	Percentage of connection with the same source port	Continuous
37	Dst_host_srv_diff_host rate	Percentage of same service coming from different host	Continuous
38	Dst_host_serror rate	Percentage of connection to a host with S0 error	Continuous
39	Dst_host_srv_serror rate	Percentage of connection to a host and specific service with S0 error	Continuous
40	Dst_host_rerror rate	Percentage of connections to a host with an RST errors	Continuous
41	Dst_host_srv_rerror rate	Percentage of connections to a host and specific service with an RST errors	Continuous

According to the features value, the data was normalized before feature selection procedures.

$$X\ (i) = \frac{x_{(i)} - x_{\min}}{x_{\max -} x_{\min}} \quad (1)$$

Where, $X_{(i)}$ is the normalized data point, $x_{(i)}$ is each data point, $x_{\min}$ and $x_{\max}$ are the maximum and minimum value of each feature.

In this research, Wrapper method is used the naïve-Bayesian algorithm to evaluate feature subsets and Genetic algorithm (GA) is used as search method for finding the best subset. The training data is divided into 10 folds and then GA is invoked to search among the subsets. After that, Naïve-Bayesian is utilized to computes the merit of each subset and selects the best ones. The naïve-Bayesian is a probabilistic classifier based on the Bayes' Theorem [35, 36] with assumption of independence among the features. As denoted in Eq. 2, the posterior probability of class is calculated from $P(c)$, $P(x)$ and $P(x\ |c)$.

$$P(c\,|\,x) = \frac{P(c)P(x\,|\,c)}{P(x)} \quad (2)$$

Where, $P(c), P(x)$ and $P(x\ |c)$ are prior probability of class, prior probability of predictor and likelihood which is the probability of predictor given class, respectively. At the end of the naïve- Bayesian procedure, decision rules are used to select more probable class for each attributes[36].

C. Proposed Anomaly-based IDS model

In this subsection, the proposed meta-heuristic algorithms are explained in details and then the implemented model for anomaly-based IDS is described.

1) Meta-heuristic algorithms

Meta-heuristic algorithms are applied to solve complicated optimization problems. The common factor of these algorithms is combining rules and randomness to imitate natural phenomena. They are based on certain physical and biological standards as inspire by natural phenomena. They are divided into two types, population and single-solution based algorithms. population based is considered more suitable because of their search capability and scape the local minimum. Different population-based algorithms are used in several domains. In this research, some of these algorithms like ICA, ACO, GA and PSO are used to detect the anomalies in networks.

ICA starts with an initial random country which includes two group: Imperialists and colonies. Over time, powerful imperialist take possession of their colonies and weak ones are eliminated from imperial population and become a colony. The power of each imperialist is obtained by their cost function. Two main phases of this algorithm are Assimilation and Revolution. In assimilation, colonies move towards imperialist states in different directions and the revolution changes the countries state randomly to improve the global search of algorithm [37]. During assimilation and revolution, a colony might receive to a better condition and might be replaced with imperialist if the cost function of imperialist is worse than the colony. Each imperialist try to expand its territory and strikes more colonies. Imperialist competition is executed as follows.

1. *Select some random points (countries) and initialized the empires.* 2. *Move colonies toward to the relevant imperialist (Assimilation).* 3. *Select new colonies using uniform random function and replace them with new generated countries (Revolution).* 4. *Swap the roles of the colonies and imperialist, if the colony is better than related imperialist.* 5. *Calculate the cost function of each empire. 6. Collapse the weakest colony from weakest empire and assign it to a powerful empire.* 7. *If an empire loses all colonies, then the empire will collapse. 8. If the stop condition is satisfied, stop, if not go to step 2.*

Ant colony optimization (ACO) is another population-based meta-heuristic algorithm which is inspired from ant's behavior when they are seeking the shortest path between their colony and source of food. Essentially, ants are blind, deaf and dumb. They communicate with each other using pheromone. Chemical pheromone that deposits on the path is consider as a weight in graph and the best way marked by strong pheromone concentrations. The optimization problem can be solved with ACO by transforming the problem of finding the best path on the weighted graph. When an ant finds a food source, it evaluates quantity and quality of the food and carries some of it back to the nest. These behavior patterns can provide some models for solving complex combinatorial optimization problems[38]. More details of the ACO algorithm are demonstrated below (Fig. 4).

ACO uses pheromone update and evaporation to find the best path between nest and food source. At the beginning, a constant amount of pheromone is assigned to all path (arc of graph), afterwards, the probability of choosing j is calculated as below.

$$p_{ij}^{k} = \frac{\tau_{ij}^{\alpha}}{\sum_{l \in N_k} \tau_{il}^{\alpha}} \text{ If } j \in N_i^{\,k} \text{ else } 0 \quad (3)$$

Where, $N_i^{\,k}$ is the neighborhood of ant k when in node i and τ_{ij} is the pheromone of node i to j.

The pheromone update and evaporation are computed by Eq. 4 and Eq. 5, respectively.

$$\tau_{ij} \longleftarrow\!\!\!\!\!\!- \tau_{ij} + \Delta\tau^{k} \quad (4)$$

$$\tau_{ij} \longleftarrow\!\!\!\!\!\!- (1-p)\tau_{ij} \quad , \forall (i\,,j) \in A \quad (5)$$

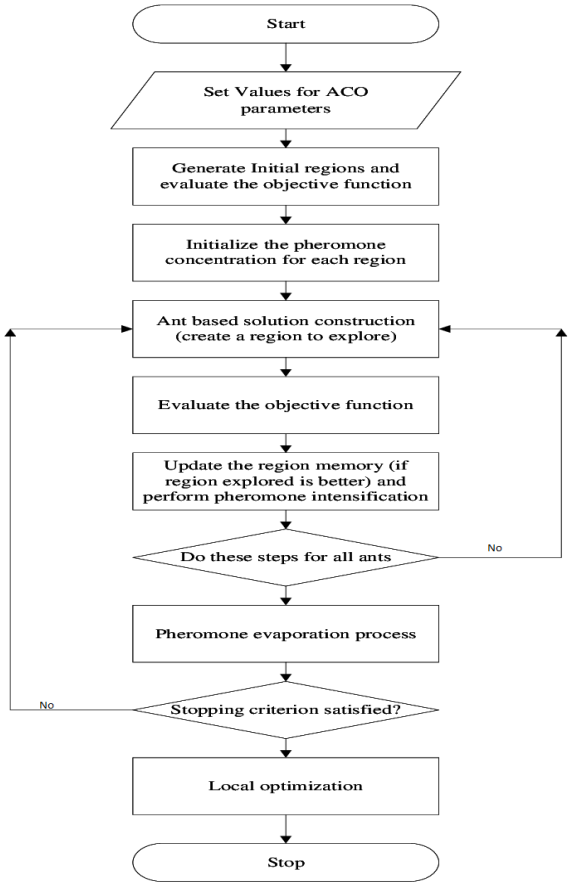

Fig. 4. ACO flowchart [38]

Genetic algorithm (GA) is an optimization algorithm that reflected the natural selection procedures. Three main phases in GA algorithm are considered as: selection, crossover and mutation. In the first stage, the parents (two individuals) are selected among best genes. Roulette wheel is a selection method which chooses the best parents based on their fitness function. In crossover phase, two individual Genes (parents) are chosen and the value of the gens are replaced with each other. There are three types of the crossover: single point crossover, two point cross over and monotonous crossover[39]. In this paper a random number is generated between (1, 2, 3) and one of the three types of crossovers is selected in each iteration to change the gene values. In mutation, some of the bits are flipped randomly. Mutation occurs to maintain diversity within the population and prevent premature convergence. The main procedure of proposed GA algorithm is as follow.

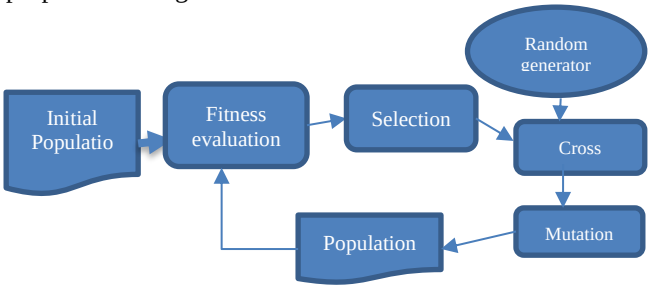

Fig. 5. Proposed GA procedure

Particle swarm optimization (PSO) is a stochastic population-based algorithm that is inspired from the nature social behavior and dynamic movements with communications of insects, birds and fish. In comparison to GA, PSO can be implemented easily with combination of self-experience and social experience. PSO is initialized with random population and then moving around the search space to find the optimal solution [40]. The particles are coordinated with two "best" parameters: P-best and G-best. The best experience of each particle is stored in the P-best and the best experience of population is saved in G-best. In each iteration, the position of particles is updated as follows.

$$x_{(t+1)} = x_{(t)} + v_{(t+1)} \quad (6)$$

$$v_{(t+1)} = wv_{(t)} + c_1 r_1 (Pbest - x_t) + c_2 r_2 (Gbest - x_t) \quad (7)$$

Where v is the velocity of the particle, $x_{(t)}$ shows the position of the particle in time t , the parameters w ,c_1,c_2 are user supplied coefficients, r_1 and r_2 are random variable that regenerate for each velocity.

IV. PROPOSED NON-LINEAR REGRESSION MODEL

In order to detect the intrusions in network, 4-step procedures are implemented. In the first step, the benchmark dataset (NSL-KDD) which consists of network packets is selected. The features of the dataset are optimized using naïve-bayesian feature selection method in second step and then, the selected features are considered as an input for the proposed model. The proposed model is based on combination of sigmoid function and non-linear regression. At the end, the best parameters of model are obtained using meta-heuristic algorithms and the precision of intrusion detection model is computed using the train and test process of meta-heuristic algorithms.

Sigmoid function is an applicable function which is extremely used in machine learning, natural language processing and optimization problems. The output of this function is bounded between (0,1) and it is defined for all real values. In this paper, the composition of sigmoid function and non-linear regression is modeled as follows.

$$f(x) = \frac{1}{1+e^{-(\sum_{i=1}^{N} \alpha_i x_i + \sum_{j=1}^{N} \sum_{k=j+1}^{N} \alpha_{jk} x_j x_k + \beta)}} \quad (8)$$

Where, N shows the number of the selected features, α denotes the population of meta-heuristic algorithms which should be optimized, β is bias number and x shows the input vector of features. As mentioned before, the output of sigmoid function is bounded between (0, 1), therefore, we used following hypothesis to train and test the algorithms.

$$class\_label = \begin{cases} 1 & f(x) \geq 0.5 \\ 0 & f(x) < 0.5 \end{cases} \quad (9)$$

Finally, mean square error (MSE) is calculated for each record of dataset.

$$MSE = \frac{\sum_{z=1}^{N} (class\_label(z) - f(s))^2}{N} \quad (10)$$

Where, $f(s)$ shows the desire output and N represent the number of rows in the dataset.

V. RESULTS

After required data preprocessing and feature selection, our model trained with the best features subset and then the test data was entered to the model to evaluate the model performance. The best feature subset is obtained using the naïve-bayesian method and the subset is classified using the decision tree to prove the subset efficiency. As shown in Fig. 6, the 16 most important features are chosen among 41 features. The selected features are shown by red points and the merit [10] of each feature is denoted in Y-axis. The reduction of features and selection of the most important ones, can lead to the accuracy increase and the process time decrease.

According to Fig. 6, the 16 most important features with their merit are selected. In order to prove the efficiency of the feature subset, the DT classifier is utilized and the outputs are demonstrated in Table 3. The accuracy of the feature selection method is evaluated based on the precision and recall. The $F-measure$ (F1 score) is defined as the weighted harmonic mean of the precision and recall (where β=1). In some applications, recall is more important than precision[41]. In order to weigh recall higher than precision the value 2 is considered for β. The F-value of a classifier is desired to be as high as possible. Therefore, the high value (close to 1) of F-value shows the performance of classifier.

TABLE 3. FEATURE SUBSET EVALUATING USING DT

	Normal	**Abnormal**	Weighted AVG.
TP rate	0.993	0.994	0.993
FP rate	0.006	0.007	0.007
Precision	0.995	0.992	0.993
Recall	0.993	0.994	0.993
F1	**0.994**	**0.993**	**0.993**
F_β	0.993	0.993	0.993
ROC Area	0.998	0.998	0.998

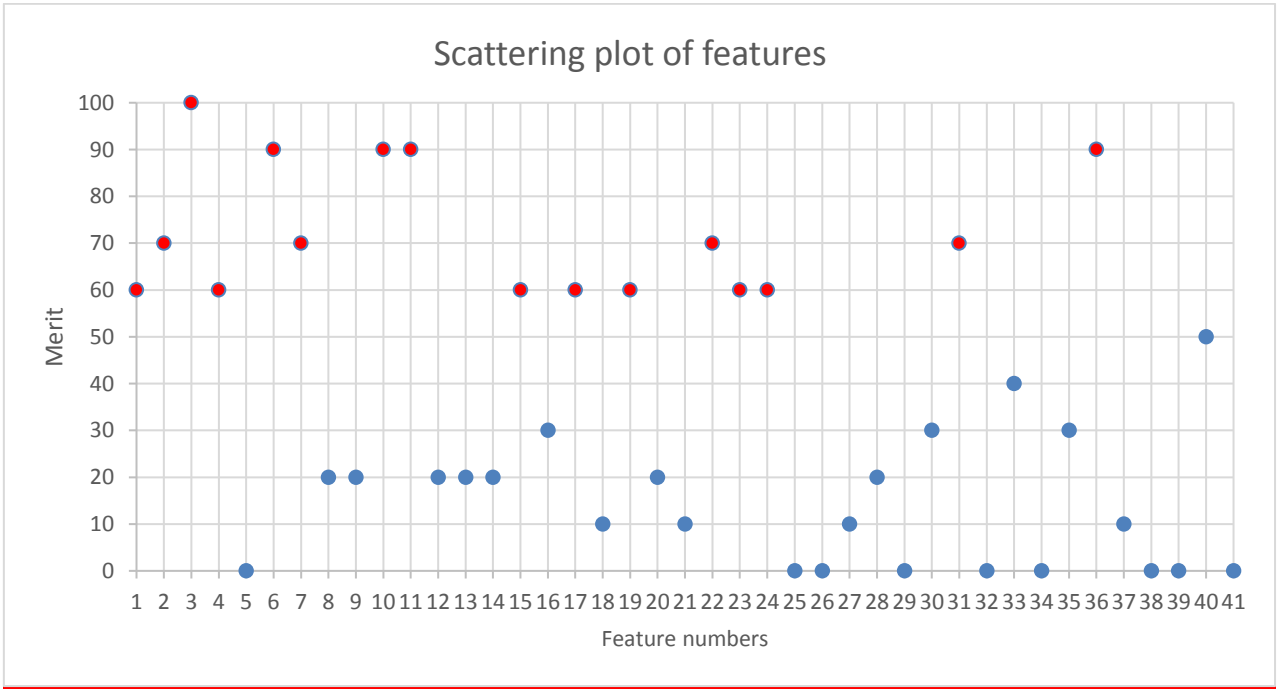

Fig. 6. Scattering of features

$$\Pr ecision = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = 2 \times \frac{\Pr ecision.\text{Re}call}{\Pr ecision + \text{Re}call} \quad (13)$$

$$F_\beta = (1 + B^2)\frac{Precision.Recall}{\left(\beta^2.Precision\right) + Recall} \quad (13\text{-}1)$$

Where, TP shows the number of classes that are correctly labeled, FP denoted the number of normal classes that are labeled as anomaly and FN is the number of anomaly classes that are incorrectly labeled as normal. As shown in Table 3, the selected features deliver a significant improvement in classification performance. The unlabeled data is classified by 99.3% accuracy. The selected features and their merit are shown in Table 4.

TABLE 4. LIST OF THE SELECTED FEATURES

Feature	Merit
service	100
dst_bytes	90
hot	90
num_failed_logins	90
dst_host_same_src_port_rate	90
protocol_type	70
land	70
is_guest_login	70
srv_diff_host_rate	70
duration	60
flag	60
su_attempted	60
num_file_creations	60
num_access_files	60
count	60
srv_count	60

Meta-heuristic algorithms which are used in this paper are trained with the selected features. Sigmoid function (Eq. 8) is considered as a cost function of the proposed method. Training results of the ACO, ICA, PSO and GA are demonstrated in Fig. 7-10. The descending pattern of all algorithms shows value of the mean square error (MSE) in each iteration. In order to reduce the MSE, the met-heuristic algorithms are conducted to find the optimal coefficients in Eq. 8. At the start of the first iteration, a random population which concluded various coefficients are used to predict the packet classes (-1 or 1). In the next iterations, the best coefficients are utilized to enhance the power of the algorithm's prediction. This procedure is evaluated by MSE. The X-axis of figures shows the number of iterations that is needed for algorithms to receive in a stable condition. According to the figures, the algorithm with the lower MSE must have a better performance. As shown in Y-axis of figures, the stable point of algorithms are as follows: ACO: 0.058594, GA: 0.061882, PSO: 0.084, ICA: 0.055941. The best number of iteration are obtained using trial and error technique.

After training procedure, the class label of test data is predicted and compared with the desire outputs. The performance of model is evaluated using accuracy of test data (Eq. 14). All of the experimental results shown are the average of 5 runs.

$$Accuracy\ (train\ /\ test) = \frac{\sum_{i=1}^{N} N_{(predict_i = desire_i)}}{N_T} \times 100 \quad (14)$$

Where, $N_{(predict_i = desire_i)}$ shows the total number of predicted classes which are classified truly and N_T is the total number of instances. The train and test accuracy of algorithms are shown in Table 5.

TABLE 5. ACCURACY OF DIFFERENT ALGORITHMS

	Train Acc.	Test Acc.
GA	92.8118	92
ACO	94.9706	94.25
ICA	**96.4059**	**96.0875**
PSO	91.6	91.175

In this paper, Table 5 confirms that ICA has more efficiency in comparison to other algorithms for the detection of intrusions. The robustness of the proposed model is proved by low FP and FN rates that obtained 0.05,0.02, respectively. Additionally, the fast convergence of the ICA proves the efficiency of the proposed model in finding the optimal solution.

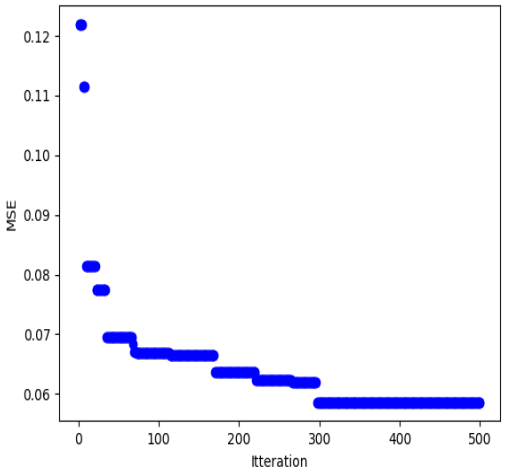

Fig. 7. ACO training chart

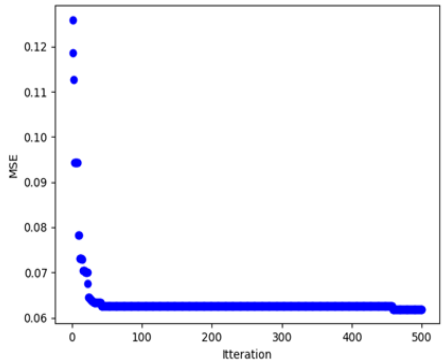

Fig. 8. GA training chart

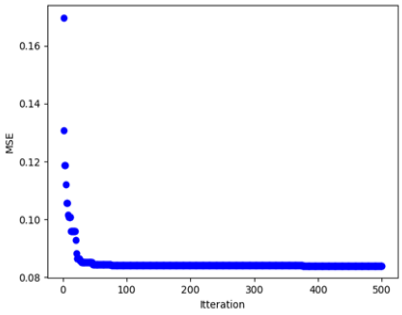

Fig. 9. PSO training chart

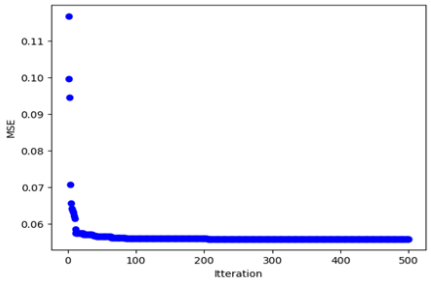

Figure 10. ICA training chart

In order to verify the performance of the proposed model, some related researches in anomaly-based domain are investigated in Table 6. The accuracy rate of the models shows that the proposed model in this paper is much better than the suggested models in [42] and [43] . According to table 6, our proposed model increases the accuracy about 8% and 13% compared to the SEIDS model and hybridized GA and Tabu search model, respectively. certainly, the difference between the accuracy in the proposed method and mentioned researches is originated from the following reasons.

1. Dataset is a most important part of machine learning process and without data, all research and automation will go vain. All datasets have irrelevant information that must eliminate from dataset. In our proposed method the NSL-KDD is used for train and test that is more efficient than the KDD-Cup dataset which is used in [40]. Authors in [39] are using whole of NSL-KDD for training that is a weakness of this research.

2. Incorrect features removing or participating can affect the classification accuracy. In [40], ten features are selected as the most important features that show lower detection accuracy. According to the result, some important features are removed from the features subset.

3. Nonlinear regression model that is used in the proposed model shows acceptable performance in discovering the complex relation between features in the subset. Therefore, the detection/prediction accuracy is highly related to the modeling. According to [20] the hybrid meta-heuristic algorithms reduced the number of features and time-consumption efficiently but lack of robust modeling and using common classifiers leads to low detection accuracy compare to our model.

TABLE 6. COMPARISON ANALYSIS

Model	Feature selection method	IDS technique
Hybrid method[42]	_	SEIDS (88.45%)
Hybridized of GA and Tabu search [43]	GA, Tabu search and KNN	Tabu search and GA (83.56%) GA only (77.17%)
Hybrid metaheuristic[20]	Hybrid meta-heuristic algorithms	J.48 as the best classifier (92.80%)
Dynamic differential annealed optimizer (DDOA) [44]	DDOA	DDOA by cost function (94.7%)
Proposed model	Naïve-Bayesian (99.3%)	**ICA (96.0875%)** ACO (94.25%) GA (92%) PSO (91.175%)

CONCLUSION

In this paper, the dual problem of accuracy and efficiency have been considered to suggest efficient and robust anomaly-based IDS. At first, the most important features are selected from a benchmark dataset (NSL_KDD). The dataset is comprised of 41 features which 16 main features are chosen using Naïve-Bayesian for training the meta-heuristic algorithms. The selected feature subset is evaluated with DT and the obtained accuracy proves the efficiency of subset. The detection methods are based on the combination of sigmoid function and non-linear regression method. The parameters of proposed model are obtained using meta-heuristic algorithms. As shown in the results, The ICA algorithm confirms better accuracy in comparison with other meta-heuristic algorithms compared in this paper. It should be said that the ICA is converged to the optimal value faster than other algorithms. This model can detect the intrusions in just 500 iterations, then the proposed model can be very time efficient. In addition, after one training, the proposed method can detect the new intrusions with stable accuracy and this can prove the method robustness.

REFERENCES

[1] W.-C. Lin, S.-W. Ke, and C.-F. J. K.-b. s. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," vol. 78, pp. 13-21, 2015.

[2] N. Hubballi and V. J. C. C. Suryanarayanan, "False alarm minimization techniques in signature-based intrusion detection systems: A survey," vol. 49, pp. 1-17, 2014.

[3] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *computers & security*, vol. 28, no. 1-2, pp. 18-28, 2009.

[4] R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas, and Y.-L. J. I. S. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," vol. 378, pp. 484-497, 2017.

[5] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. J. J. o. i. s. Molla, "Feature selection for intrusion detection using random forest," vol. 7, no. 03, p. 129, 2016.

[6] M. Xiao, H. J. I. Nagamochi, and Computation, "Exact algorithms for maximum independent set," vol. 255, pp. 126-146, 2017.

[7] J. Wang, M. Yin, and J. J. T. C. S. Wu, "Two approximate algorithms for model counting," vol. 657, pp. 28-37, 2017.

[8] E. K. Burke, M. Hyde, G. Kendall, G. Ochoa, E. Özcan, and J. R. Woodward, "A classification of hyper-heuristic approaches," in *Handbook of metaheuristics*: Springer, 2010, pp. 449-468.

[9] G. L. Pappa *et al.*, "Contrasting meta-learning and hyper-heuristic research: the role of evolutionary algorithms," vol. 15, no. 1, pp. 3-35, 2014.

[10] M. Babagoli, M. P. Aghababa, and V. J. S. C. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," pp. 1-13, 2018.

[11] X.-S. Yang, *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[12] J. M. Framinan, J. N. Gupta, and R. J. J. o. t. O. R. S. Leisten, "A review and classification of heuristics for permutation flow-shop scheduling with makespan objective," vol. 55, no. 12, pp. 1243-1255, 2004.

[13] A. M. Shaheen, S. R. Spea, S. M. Farrag, and M. A. J. A. S. E. J. Abido, "A review of meta-heuristic algorithms for reactive power planning problem," vol. 9, no. 2, pp. 215-231, 2018.

[14] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, 2009, pp. 1-6: IEEE.

[15] A. Waskita, H. Suhartanto, P. Persadha, and L. T. Handoko, "A simple statistical analysis approach for intrusion detection system," in *Systems, Process & Control (ICSPC), 2013 IEEE Conference on*, 2013, pp. 193-197: IEEE.

[16] N. Moustafa and J. J. I. S. J. A. G. P. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," vol. 25, no. 1-3, pp. 18-31, 2016.

[17] J. Jabez and B. J. P. C. S. Muthukumar, "Intrusion Detection System (IDS): Anomaly detection using outlier detection approach," vol. 48, pp. 338-346, 2015.

[18] H. Sadreazami, A. Mohammadi, A. Asif, K. N. J. I. T. o. S. Plataniotis, and I. P. o. Networks, "Distributed-Graph-Based Statistical Approach for Intrusion Detection in Cyber-Physical Systems," vol. 4, no. 1, pp. 137-147, 2018.

[19] C. Manikopoulos and S. J. I. C. M. Papavassiliou, "Network intrusion and fault detection: a statistical anomaly approach," vol. 40, no. 10, pp. 76-82, 2002.

[20] O. Almomani, "A Hybrid Model Using Bio-Inspired Metaheuristic Algorithms for Network Intrusion Detection System," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 68, no. 1, pp. 409-429, 2021.

[21] S. Hosseini and B. M. H. Zade, "New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN," *Computer Networks*, vol. 173, p. 107168, 2020.

[22] C. I. Rene and J. J. I. Abdullah, "Malicious Code Intrusion Detection using Machine Learning And Indicators of Compromise," 2017.

[23] C. Kruegel and T. Toth, "Using decision trees to improve signature-based intrusion detection," in *International Workshop on Recent Advances in Intrusion Detection*, 2003, pp. 173-191: Springer.

[24] R. R. Patel and C. S. Thaker, "Zero-day attack signatures detection using honeypot," in *International Conference on Computer Communication and Networks (CSI-COMNET)*, 2011.

[25] S. O. Al-Mamory and H. Zhang, "A survey on IDS alerts processing techniques," in *Proceeding of the 6th WSEAS international conference on information security and privacy (ISP'07), Spain*, 2007, pp. 69-78.

[26] P. A. Porras, M. W. Fong, and A. Valdes, "A mission-impact-based approach to INFOSEC alarm correlation," in *International Workshop on Recent Advances in Intrusion Detection*, 2002, pp. 95-114: Springer.

[27] M. J. I. J. o. C. A. Gupta, "Hybrid Intrusion Detection System: Technology and Development," vol. 115, no. 9, 2015.

[28] I. Dutt, S. Borah, I. K. Maitra, K. Bhowmik, A. Maity, and S. Das, "Real-Time Hybrid Intrusion Detection System Using Machine Learning Techniques," in *Advances in Communication, Devices and Networking*: Springer, 2018, pp. 885-894.

[29] C. Amza, C. Leordeanu, and V. Cristea, "Hybrid network intrusion detection," in *Intelligent Computer Communication and Processing (ICCP), 2011 IEEE International Conference on*, 2011, pp. 503-510: IEEE.

[30] C. Estan and G. Magin, "Interactive Traffic Analysis and Visualization with Wisconsin Netpy," in *LISA*, 2005, vol. 5, pp. 17-17.

[31] D. Santoro, G. Escudero-Andreu, K. G. Kyriakopoulos, F. J. Aparicio-Navarro, D. J. Parish, and M. J. M. Vadursi, "A hybrid intrusion detection system for virtual jamming attacks on wireless networks," vol. 109, pp. 79-87, 2017.

[32] N. V. Abhishek, T. J. Lim, B. Sikdar, and A. Tandon, "An Intrusion Detection System for Detecting Compromised Gateways in Clustered IoT Networks," in *2018 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)*, 2018, pp. 1-6: IEEE.

[33] H. Bostani and M. J. C. C. Sheikhan, "Hybrid of anomaly-based and specification-based IDS for Internet of Things using unsupervised OPF based on MapReduce approach," vol. 98, pp. 52-71, 2017.

[34] S. Aljawarneh, M. Aldwairi, and M. B. J. J. o. C. S. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," vol. 25, pp. 152-160, 2018.

[35] G. J. N. I. D'Agostini, S. Methods in Physics Research Section A: Accelerators, Detectors, and A. Equipment, "A multidimensional unfolding method based on Bayes' theorem," vol. 362, no. 2-3, pp. 487-498, 1995.

[36] J. Chen, H. Huang, S. Tian, and Y. J. E. S. w. A. Qu, "Feature selection for text classification with Naïve Bayes," vol. 36, no. 3, pp. 5432-5435, 2009.

[37] S. Shamshirband, A. Amini, N. B. Anuar, M. L. M. Kiah, Y. W. Teh, and S. J. M. Furnell, "D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks," vol. 55, pp. 212-226, 2014.

[38] W. Feng, Q. Zhang, G. Hu, and J. X. J. F. G. C. S. Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks," vol. 37, pp. 127-140, 2014.

[39] D. J. S. Whitley and computing, "A genetic algorithm tutorial," vol. 4, no. 2, pp. 65-85, 1994.

[40] P. N. Suganthan, "Particle swarm optimiser with neighbourhood operator," in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, 1999, vol. 3, pp. 1958-1962: IEEE.

[41] B. S. Bhati and C. Rai, "Ensemble Based Approach for Intrusion Detection Using Extra Tree Classifier," in *Intelligent Computing in Engineering*: Springer, 2020, pp. 213-220.

[42] Z.-H. Chen and C.-W. Tsai, "An Effective Metaheuristic Algorithm for Intrusion Detection System," in *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, 2018, pp. 154-159: IEEE.

[43] O. C. Abikoye, T. O. Aro, R. O. Obisesan, and A. N. Babatunde, "Hybridized Intrusion Detection System Using Genetic and Tabu Search Algorithm," 2017.

[44] A. J. Wilson and S. Giriprasad, "A Feature Selection Algorithm for Intrusion Detection System Based On New Meta-Heuristic Optimization," *Journal of Soft Computing and Engineering Applications*, vol. 1, no. 1, 2020.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۲، شماره ۱، بهار ۱۴۰۲، صفحه ۲۹-۴۰
دریافت: ۱۳۹۹/۰۳/۱۷، بازنگری: ۱۴۰۰/۰۴/۱۳، پذیرش: ۱۴۰۰/۰۶/۰۹

تکنیک تشخیص نفوذ اکتشافی مبتنی بر رگرسیون غیرخطی و تابع سیگموئید

شهریار محمدی*[۱]، مهدی باباگلی[۲]
۱*- دانشکده صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران
۲- دانشکده صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران

[1*]mohammadi@kntu.ac.ir , [2]Mehdi.babagoli@email.kntu.ac.ir

* نشانی نویسنده مسئول: شهریار محمدی، سید خندان، دانشگاه خواجه نصیرالدین طوسی، دانشکده صنایع/فناوری اطلاعات

چکیده- گسترش فن آوری های اینترنتی طی دهه های گذشته به وابستگی فعالیت های کاربران در فضای مجازی به خدمات ارائه شده توسط شبکه های رایانه ای منجر شده است. در این فضا سیستمی به نام سیستم تشخیص نفوذ(IDS) وجود دارد که ترافیک شبکه را برای تشخیص رفتارهای غیرطبیعی و همچنین فعالیتهای ناهنجاری کنترل می کند. استحکام و کارآیی زمان IDS به عنوان یک مسئله اساسی در شبکه ها در نظر گرفته می شود. در این مقاله ، مدل جدیدی مبتنی بر الگوریتم های فرا اکتشافی برای شناسایی بسته های غیر طبیعی به کار گرفته شده است. به منظور توسعه استراتژی با کارایی بالا از موارد ذیل استفاده شده است: یک مجموعه داده مرجع (NSL-KDD) ، روش انتخاب ویژگی با دقت بالا و چهار الگوریتم فرا اکتشافی. این مجموعه داده شامل ۱۵۰۴۹۰ بسته نرمال و غیر طبیعی است که از یک شبکه نظامی ضبط شده است و ۱۶ ویژگی مهم با استفاده از روش انتخاب ویژگی wrapper از این مجموعه داده استخراج می شوند. روش انتخاب ویژگی ذکر شده از روش Naïve-Bayesian برای ارزیابی زیر مجموعه های ویژگی استفاده می کند. پس از فرآیند انتخاب ویژگی ، از چهار الگوریتم فرا اکتشافی برای تشخیص ناهنجاری در اتصالات شبکه استفاده می شود. پارامترهای تابع هزینه (ترکیب رگرسیون غیر خطی و سیگموئید) با استفاده از الگوریتم های فرا اکتشافی بهینه می شوند. نتایج به دست آمده نشان می دهد که الگوریتم رقابت استعماری از لحاظ دقت نسبت به سایر الگوریتم های فرا اکتشافی بهتر است و همچنین همگرایی قابل قبولی جهت پیدا کردن جواب بهینه دارد.

واژه‌های کلیدی: امنیت شبکه ، سیستم تشخیص نفوذ ، فرا اکتشافی ، بیزین ، رگرسیون غیر خطی ، تابع سیگموئید