IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 12, 2020, accepted October 20, 2020, date of publication November 2, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035250

Epigenetic Algorithm-Based Detection Technique for Network Attacks

MEHDI EZZARII[1], HAMID EL GHAZI[1], HASSAN EL GHAZI[1], AND FAISSAL EL BOUANANI[2], (Senior Member, IEEE)
[1]National Institute of Posts and Telecommunications, Rabat 10100, Morocco
[2]ENSIAS College of Engineering, Mohammed V University, Rabat 713, Morocco

Corresponding author: Faissal El Bouanani (f.elbouanani@um5s.net.ma)

ABSTRACT Nowadays, the cybersecurity issue involves new strategies to protect against advanced threats and unknown attacks. Intrusion detection system (IDS) is considered a robust system dealing with attacks detection, particularly unknown attacks and anomalies. Several IDS-based algorithms have been recently inspected in the literature, among them the well-known strengthen algorithms, i.e. Genetic algorithm (GA). Moreover, Epigenetic-based algorithm (EGA) is known as an improved version of GA ensuring high performance with reduced computational complexity. Its main goal is to converge within a short time towards an optimal solution by acting on genetic operators, namely mutation and crossover. In this article, we propose a new classifier based on EGA for IDS. Especially, based on a database of network traffics, EGA is applied to classify attacks. The results, performed through EGA simulation, show that the performance of the proposed technique outperforms the ones of GA classifier by obtaining a high detection rate up to 98% and a faster processing time than that of GA and other algorithms that we have compared in this article.

INDEX TERMS Epigenetic algorithm, genetic algorithm, intrusion detection system, network, security.

I. INTRODUCTION

Intrusion Detection Systems (IDS) is one of the main techniques used to ensure security in a network or computing environment. It is defined as either software or hardware systems that monitor and analyze events occurring in a computer system or network so as to detect malicious activities or intrusions [1]. IDS are systems that have proven their efficiency in serious security constraints [2]. Owing to this fact, they attract the researchers' attention by proposing new approaches aiming to improve the system security robustness against new potential unknown attacks. Actually, three types of methods used in IDS can be distinguished, namely (i) signature-based method [3], (ii) anomaly-based method [4], and (iii) hybrid signature/anomaly-based method to get a complementary intrusion detection. While signature-based IDS technique matches the presented attack's signature with a database of known attacks [5], the anomaly-based IDS can effectively identify unknown attacks whose signatures do not exist in database, by learning about certain normal behaviors in the network. To this end, it raises alerts or block traffics once an abnormal behavior in the network is detected [5]. Although Anomaly-based IDS can detect efficiently unknown attacks, there still a big challenge to ensure high accurate performance by maximizing the detection rate and minimizing the false positive one [6]. The GA algorithm is categorized as an anomaly-based IDS and is one of the best-known heuristic methods and evolutionary algorithms dealing with the solution of an optimization and classification problem [7]. Firstly, GA was combined with a multi-agent system so as to improve the IDS detection efficiency [8]. In [9], GA was used to find the optimal parameters of fuzzy functions multi-agent. Later, GA was applied to classify and generate the best rules for intrusion detection purposes [10]–[12]. Nevertheless, the use of GA made the training procedure costly as it requires more data and time. Although its efficiency, finding a fitness function still the major concern of GA in IDS [13]. To remedy the limitations of GAs, Epigenetic Algorithm (EGA) is presented as a concurrent algorithm allowing to reach the optimal solution within a respected time. For that, EGA have attracted researchers' attention and shown their effectiveness in solving some problems such as GSM mobile planning frequency [14] and Inverse Kinematics problem [15]. Essentially, it relies on the control of the randomness of gene activities. This can be ensured by the use of additional factors aiming to enhance both mutation

The associate editor coordinating the review of this manuscript and approving it for publication was Yanjiao Chen.

and crossover operations in classical GA, and consequently contribute to its convergence acceleration [16]. Inheritance, crossover and mutation operations exist in EGA structure, and they are called as epigenetic inheritance, epicrossover, and epimutation [17]. This new operations acting on individuals' genes including inactive ones, called epigenetic factors (EF), to build the optimal next population. Such a list can be built relying on one of the feature selection methods such as correlation features selection (CFS) [18], [19], InfoGain [20], or gain ratio (GR) [21].

A. RELATED WORKS

In the context of IDS, table 1 outlines the most known approaches utilized in the literature, dealing with the considered problem! Essentially, among these approaches, we site statistical [22], [23], datamining-based methods [24], machine learning (ML) techniques [25], [26], artificial neural network (ANN) [27], fuzzy logic (FL) [28], support vector machine (SVM) [29], [30], and genetic algorithm (GA). We can then classify this number of detection approaches according to adopted database, performance result, and approach's limitations. Specifically in [23], the presented model was based on a statistical approach by collecting communication activities, and then conducting a joint analysis to detect the host malicious behavior. However, this model give a low accurate. The Bayesian-based [31] model provide a moderate accuracy because the focus is on classifying the classes for the instances, not the exact probabilities [32]. While ML algorithm has been used in anomaly-based malware detection techniques [33], the ANN-based algorithm has been presented as a detection technique of both known and unknown distributed denial of service (DDoS) attacks [27]. As heuristic and evolutionary learning, ANN-based algorithm requires more data and time that why it made the training procedure costly [34]. Furthermore, FL and SVM-based methods strengthening the intrusion detection in the network were investigated in [28] and [30], respectively. FL provide high accuracy and high false alarms [32]. In the new data set that contains new attacks (like ADFA-WD and Bot-IoT) SVM provide a good accuracy but it was not as good as for every machine learning technique used [35]. In [36], the wavelet neural network (WNN) model applied to the IDS gave results with a moderate accuracy and high computational complexity because it is necessary to reduce the size of the wavelet decomposed data. As for convolutional neural network (CNN) algorithm [37], this model gives a moderate accuracy and a high cost of computational in front of a complex architecture and a diversity of the data in real time [38].

B. CONTRIBUTION

Motivating by the above, we propose a new scheme based on EGA to detect intrusions and network attacks in IDS. Explicitly, the presented algorithm improves the detection and false negative rates of IDS to achieve high accurate performance while reducing the running time. Specifically, the main contributions of this article can be summarized as follows:

- Based on a training dataset, we propose an epigenetic algorithm to detect and classify attacks,
- We optimize the proposed algorithm's parameters so that to enhance the proposed classifier's reliability,
- We provide useful insights into the performance of the EGA-based classifier for IDS.

The remainder of this article is structured as follows. Section II presents the proposed EGA-based algorithm for IDS and discussed the effects of key parameters of the system's security. Simulation results and insightful discussions gained into the IDS performance are summarized in Section III. Lastly, section IV reports closing remarks that outlines the current contribution.

TABLE 1. Classification of related works for various intrusion detection approaches.

Detection approach	*Examples*	*Used dataset*	*Research findings and limitations*	*Authors and year*
Statistical-based	Statistics	Customizable dataset UDP traffic.	Low accurate. Not effective for detecting advanced and complicated attacks	Fragkiadakis, 2012; Hamed, 2017.
	Bayesian-based	KDD_NSL, DARPA, KDD Cup99	Moderate accuracy	Stavroulakis, 2010; Khraisat, 2019.
Data Mining-based or Rule-based [39]	k-means and k-nearest neighbour	KDD_NSL, DARPA, KDD Cup99	Moderate accuracy. Not easily to create and update [35]	Azad, 2013; Liao, 2013.
ML and Heuristic-based	FL	KDD_NSL, DARPA 98	High accuracy. High false alarms	Botha, 2003; Khraisat, 2019.
	GA	KDD_NSL, DARPA, KDD Cup99	Moderate accuracy	Davahli, 2020.
	SVM	KDD_NSL, DARPA 98, ADFA-WD [34], Bot-IoT [40]	Usually, good performance for a binary class problem	Liao, 2013.
	ANN	KDD_NSL, DARPA 98	Moderate accuracy. Self-learning with fault tolerant [31]	Saied, 2017; Stavroulakis, 2010.
	WNN	KDD99	Moderate accuracy. High computational complexity.	Hamid, 2018.
	CNN	KDD_NSL, KDD99	Moderate accuracy and a high cost of computational.	Vinayakumar, 2017.

TABLE 2. Performance metric of IDS.

Metric	Value
D_r	$\frac{N_P^{(T)}}{N_P^{(T)}+N_N^{(F)}}$
F_p	$\frac{N_P^{(F)}}{N_P^{(F)}+N_N^{(T)}}$
P	$\frac{N_P^{(T)}}{N_P^{(T)}+N_P^{(F)}}$
F_N	$\frac{N_N^{(F)}}{N_N^{(N)}+N_P^{(T)}}$
S	$\frac{N_P^{(T)}}{N_P^{(T)}+N_N^{(F)}}$
S_p	$\frac{N_N^{(T)}}{N_N^{(T)}+N_P^{(F)}}$
A	$\frac{N_P^{(T)}+N_N^{(T)}}{N_P^{(T)}+N_P^{(F)}+N_N^{(T)}+N_N^{(F)}}$

II. EPIGENETIC-BASED ALGORITHM FOR IDS

A. KEY METRICS FOR IDS

The effectiveness of an IDS is related to several key rates parameters, namely Detection Rate (D_r), False Positive (F_p), Precision (P), False Negative Rate (F_N), Accuracy (A), Sensitivity (S), and Specificity (S_p) [41].

The following variables are defined as follows:

- $N_P^{(T)}$: number of intrusions successfully detected,
- $N_P^{(F)}$: number of normal traffic wrongly detected an intrusion,
- $N_N^{(T)}$: number of normal traffics successfully labeled as non-intrusive,
- $N_N^{(F)}$: number of intrusions labeled as normal traffic.

The aforementioned key metrics of IDS are summarized in Table 2 [42], [43]. It is worth mentioning that these metrics will be able to assess the performance of the proposed EGA-based algorithm for IDS. Also, they are mandatory to decide on the appropriate parameters related to the IDS process.

B. PROPOSED EPIGENETIC ALGORITHM FOR IDS

By applying EGA, the best population of rules is given as an input to IDS. Indeed, each rule is is a part of a decision maker allowing to verify whether the traffic is either normal or attack. Noteworthy that the no-active genes, which do not participate in attack detection, are gathered in a particular list, called the Epigenetic factor list (EFL).

Let's introduce the following notations:

- $T_t^{(i)} = \left\{g_{i,j}^{(t)}\right\}_{1\leq j\leq p}$, $T_r^{(i)} = \left\{g_{i,j}^{(r)}\right\}_{1\leq j\leq p}$, and $T_s^{(i)} = \left\{g_{i,j}^{(s)}\right\}_{1\leq j\leq p}$ be the ith test, training, and solution message represented by a set of p genes, e.g. Networks, Endpoint host, industrial platform, respectively.
- $E_t = \left\{T_t^{(i)}\right\}_{1\leq i\leq M_t}$, $E_r = \left\{T_r^{(i)}\right\}_{1\leq i\leq M_r}$, and $E_s = \left\{T_s^{(i)}\right\}_{1\leq i\leq M_s}$ denote the set of test, training, and messages, respectively, where M_t, M_r, and M_s account for the cardinal of these sets.
- Each $T_i^{(\alpha)}$ with $\alpha \in \{t, r, s\}$ is a rule and composed by two parts, (i) $\mathcal{C}\left(T_i^{(\alpha)}\right) = \left\{g_{i,j}^{(\alpha)}\right\}_{1\leq j\leq p-1}$ referring to the traffic content or the condition to check, and (ii) the decision $\mathcal{A}\left(T_i^{(\alpha)}\right) = g_{i,p-1}^{(\alpha)}$ on the traffic's state (i.e. attack, normal). For the ease of exposition, the following encoding for the traffic's state is employed

$$\mathcal{A}\left(T_\alpha^{(i)}\right) = \begin{cases} normal, & 1 \\ attack, & 0 \end{cases} \tag{1}$$

- Each gene $g_{i,j}^{(\alpha)}$ can be either active or inactive. To formalize such state, the following function is introduced.

$$\mathcal{S}\left(g_{i,j}^{(\alpha)}\right) = \begin{cases} active, & 1 \\ inactive, & 0. \end{cases} \tag{2}$$

  Basically, all messages, irrespective of their categories, are assumed to have the same states at the same indices. Owing to this fact, the above function can be redefined, for the sake of simplicity, by only the position index in the message, i.e. $\mathcal{S}(j)$.
- For a set of mutation and crossover probabilities p_c and p_m, and the list $\mathcal{L}$, $E_s^{(GA)}(p_m, p_c)$ and $E_s^{(EGA)}(\mathcal{L})$ refer to the set of GA and EGA solutions, respectively, and are expressed as:

$$E_s^{(GA)}(p_m, p_c) = \left\{T_{s(GA)}^{(i)}\right\}_{1\leq i\leq M_s} \tag{3}$$

$$E_s^{(EGA)}(\mathcal{L}) = \left\{T_{s(EGA)}^{(i)}\right\}_{1\leq i\leq M_s} \tag{4}$$

Fig. 1 depicts the flowchart of our proposed algorithm. The first step consists of preparing the set E_s of the solutions, provided by EGA-based algorithm on the training set E_r. Then, for each element $T_t^{(i)}$ of E_t, the closest solution $T_s^{(i,*)}$ maximizing the following objective function is selected

$$T_s^{(i,*)} = \underset{1\leq k\leq M_s}{\arg\max} f\left(T_t^{(i)}, T_s^{(k)}\right), \tag{5}$$

with

$$f\left(T_t^{(i)}, T_s^{(k)}\right) = \sum_{j=1}^{p-1} h\left(g_{i,j}^{(t)}, g_{k,j}^{(s)}\right), \tag{6}$$

and

$$h\left(g_{i,j}^{(t)}, g_{k,j}^{(s)}\right) = \begin{cases} 1, & g_{i,j}^{(t)} = g_{k,j}^{(s)} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Specifically, such a solution is kept iff $\left(T_t^{(i)}, T_s^{(i,*)}\right) \geq f_{min}$ with f_{min} is a fixed threshold below $p-1$.

Mainly, the algorithm presented in Fig.1 aims to calculate the variables $N_P^{(T)}, N_P^{(F)}, N_N^{(T)}$, and $N_N^{(F)}$ Typically, such a process is structured as follows:

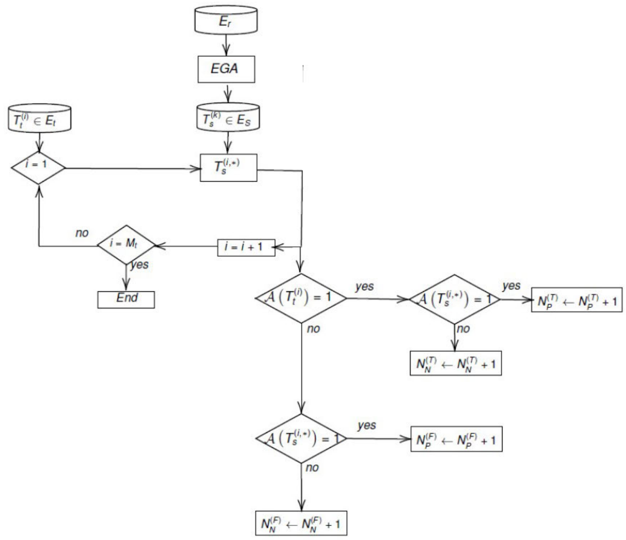


FIGURE 1. Flowchart of evaluation of proposed algorithm.

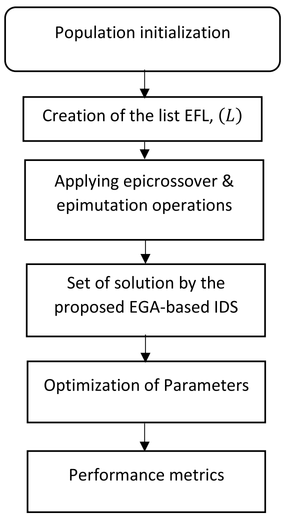


FIGURE 2. Flowchart of the overall proposed approach process.

- Based on the training set E_r, the set of solutions E_s is obtained with the help of EGA algorithm,
- Next, for each element $T_t^{(i)}$ from E_t, we calculate the closest solution $T_s^{(i,*)}$, as detailed in equations (3)-(5), by looking for the message having the maximum number of features obtained by matching between the $T_t^{(i)}$ and a k element $T_s^{(k)}$ of E_s above f_{min},
- The metrics' variables are then calculated by comparing the last genes of $T_t^{(i)}$ and $T_s^{(i,*)}$; i.e., attack or normal,
- Such steps are repeated M_s times (i.e., by going through all the messages of E_s).

C. EGA ALGORITHM FOR IDS

The proposed approach is presented in flowchart Fig. 2. While Algorithm 1 presents the EGA algorithm, a part of the IDS-EGA flow-chart presented in Fig. 2. In the sequel, the detailed steps of such algorithm are provided.

Algorithm 1: Epigenetic Algorithm (EGA) for IDS

Input: E_r
Output: E_s
// Last generation provided at the N_gth step
parameter: $N_e, N_g, M_s, M_r, p, \mathcal{L}, N_{EF}, w_1, w_2, \phi_{\min}$
$P^{(0)} \leftarrow$ `InitialPopulation` (E_r, M_s);
// $P^{(0)} = \{I_i\}_{i \leq M_s}$
$k \leftarrow 1$; // Index of the next generation to build
;
while $k \leq N_g$ **do**
  for $m \leftarrow 1$ **to** M_s **do**
    $\phi(I_i) \leftarrow$ `Fitness` $(I_i, E_r, M_r, w_1, w_2)$;
  end for
  $P_s^{(k-1)} \leftarrow$ `DecreasingSort` $(\boldsymbol{\phi}, P^{(k-1)})$;
  // $\boldsymbol{\phi} = \{\phi(I_i)\}_{i \leq M_s}$, $P_s^{(k-1)} = \left\{I_i^{(s)}\right\}_{i \leq M_s}$
  // copy the N_e best ones to the next generation
  for $m \leftarrow 1$ **to** N_e **do**
    $P_s^{(k)}(m) \leftarrow I_m^{(s)}$;
  end for
  // complete the remaining $M_s - N_e$ individuals of the next generation
  $m \leftarrow N_e + 1$;
  while $m \leq M_s$ **do**
    // Select 2 parents based on *SUS* method with fitness $\leq \phi_{\min}$
    $\{I_\alpha, I_\beta\} \leftarrow$ `SUS` $\left(P_s^{(k-1)}, \phi_{\min}\right)$;
    // Apply Epimutation
    $I_\alpha \leftarrow$ `EpiMut` $(I_\alpha, \mathcal{L}, N_{EF})$;
    $I_\beta \leftarrow$ `EpiMut` $(I_\beta, \mathcal{L}, N_{EF})$;
    // Apply Epicrossover and create 2 new individuals
    $\{C_1, C_2\} \leftarrow$ `EpiCross` $(I_\alpha, I_\beta, \mathcal{L}, N_{EF})$;
    $P_s^{(k)}(m) \leftarrow C_1$;
    $P_s^{(k)}(m+1) \leftarrow C_2$;
    $m \leftarrow m + 2$;
  end while
  $k \leftarrow k + 1$;
end while
$E_s \leftarrow P_s^{(N_g)}$;

1) INITIAL POPULATION

The initial population is a key factor contributing to the convergence [44]. Here, the initial population $P^{(0)}$ contains M_s individuals $I_i = \left\{g_{i,j}\right\}_{i \leq j \leq p}$, imported from E_r, such that they are evenly distributed between normal and attack messages

$$\Pr(\mathcal{A}(I_i) = b) = \frac{1}{2}; \quad b = 0, 1$$

while these attack messages are uniformly distributed over four categories: (i) deny of service (DoS) attacks, (ii) user to root (U2R), (iii) remote to local (R2L), and (iv) Probe, namely occurs with probability $\frac{1}{8}$. Each individual contains p alphanumeric information (i.e. genes) on the traffic. The above individuals are imported from E_r. The last population is retrieved by iterating a set of steps N_g times where N_g refers to the maximum number of generations.

TABLE 3. Encoding data.

Feature	*Coded value*
Normal	0
Attack	1
Protocol-type: TCP	2
Protocol-type: UDP	3
Protocol-type: ICMP	4
Flag: OTH	5
Flag: REJ	6
Flag: RSTO	7
Flag: RSTOS0	8
Flag: RDTR	9
Flag: S0	10
Flag: S1	11
Flag: S2	12
Flag: S3	13
Flag: SF	14
Flag: SH	15
Other services	116 to 81

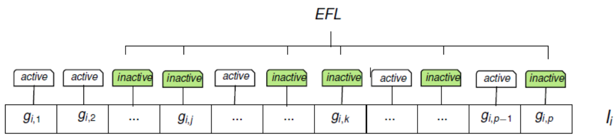


FIGURE 3. Epigenetic Factors List (EFL).

2) DATA PREPARATION: ENCODING PROCESS

The sets E_r and E_t are imported from the KDD_NSL [45] that is an example of a dataset used by researchers to compare different detection intrusion methods. Each message T from this dataset contains 41 features (duration, protocol type, service, dynamic indicator, etc.). This data contains numeric and text values. To then implement the proposed algorithm, these values have to be encoded into numeric features. Table 3 shows the transformation used for each nominal features of KDD_NSL.

3) EPIGENETIC FACTORS

The EF list (EFL) contains the indices of inactive genes among $\{1 \ldots p\}$ as shown in Fig. 3. Its establishment is mandatory before orchestrating crossover and mutation operations. Such list can be built relying on CFS, InfoGain, or GR methods. The best method among these is selected be applying a test algorithm, e.g., NaiveBayes and J48, as detailed in results section. Mathematically speaking, EFL can be defined as follows

$$\mathcal{L} = \{j \in \{1 \ldots p\}, \mathcal{S}(j) = 0\}. \tag{8}$$

4) FITNESS FUNCTION AND SELECTION PHASE

The first step in generating the next population is to sort the individual of the current population in decreasing order of their fitnesses. To this end, such function shall be chosen carefully to efficiently classify the individuals. In our work, the following fitness function is considered [46].

$$\phi(I_i) = N^{(i)} \left(\frac{w_1}{M_r} + \frac{w_2}{N_{\mathcal{C}}^{(i)}} \right), \tag{9}$$

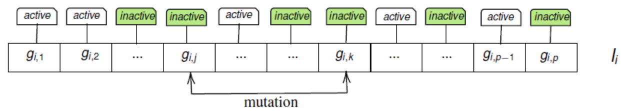


FIGURE 4. Epimutation operation.

with $N_{\mathcal{C}}^{(i)}$ is the number of messages in E_r having the same condition $\mathcal{C}(I_i)$, i.e.,

$$N_{\mathcal{C}}^{(i)} = \left| j \in \{1 \ldots M_r\}, \quad \mathcal{C}(I_i) = \mathcal{C}\left(T_j^{(r)}\right) \right|, \tag{10}$$

$N^{(i)}$ is the number of messages in E_r having the same condition and action as those of I_i, namely

$$N^{(i)} = \left| j \in \{1 \ldots M_r\}, \mathcal{C}(I_i) = \mathcal{C}\left(T_j^{(r)}\right) \& \mathcal{A}(I_i) = \mathcal{A}\left(T_j^{(r)}\right) \right|, \tag{11}$$

while the two weights w_1 and w_2 are two positive constant coefficients. Obviously, $N^{(i)} \leq N_{\mathcal{C}}^{(i)} \leq M_r$. That is, $\phi(I_i) \leq w_1 + w_2$ and assumed less than 1. For a normalization purpose, $w_1 + w_2$ is set unity. Moreover, to keep a balance between the two terms in 9, w_1 is chosen above 0.5 as $M_r > N_{\mathcal{C}}^{(i)}$. As a result, $w_2 < 0.5$.

Next, at each step $k \leq N_g$, the N_e best top individuals (i.e., elites) are reproduced into the next generation, i.e. $P^{(k)}$. To this end, the remaining $M_r - N_e$ individuals in $P^{(k)}$ are completed by selecting repeatedly and randomly, two best elements I_α and I_β with significant higher fitness, from which two childrens will be generated by performing both epimutation and epicrossover operations. Specifically, such fitness must be above a certain threshold $\phi_{\min}$ to ensure the minimum requirements, namely

$$\phi(I_\Theta) \geq \phi_{\min}, \quad \Theta = \alpha, \beta. \tag{12}$$

Promisingly, Stochastic Universal Sampling (SUS) selection method is considered here for performance enhancement purposes [25].

5) EPIMUTATION AND EPICROSSOVER OPERATORS

In EGA, both epimutation and epicrossover operators are applied exclusively to the non-active genes as presented in Fig. 4 and 5. If N_{EF} denotes the cardinal of EFL then the epimutation operation, depicted in Fig. 4, swaps the values of two selected genes $g_{i,j}$ and $g_{i,k}$ of the same individual I_i with the probabilities $\frac{1}{N_{EF}}$ and $\frac{1}{N_{EF}-1}$, respectively with j and k are two indices in EFL, chosen randomly and uniformly from $\mathcal{L}$ and $\mathcal{L} \backslash \{j\}$, respectively, i.e.,

$$z \leftarrow g_{i,j}; \quad g_{i,j} \leftarrow g_{i,k}; \; g_{i,k} \leftarrow z,$$

In the same manner, the epicrossover interchange the values of two selected genes between the two parents I_α and I_β, as shown in Fig. 5 as

$$\begin{cases} z_1 \leftarrow g_{\alpha,i}; \, g_{\beta,l} \leftarrow g_{\alpha,j}; \, g_{\alpha,i} \leftarrow z_1 \\ z_2 \leftarrow g_{\alpha,j}; \, g_{\beta,k} \leftarrow g_{\alpha,j}; \, g_{\alpha,j} \leftarrow z_2 \end{cases}$$

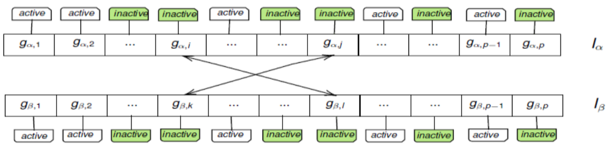


FIGURE 5. Epicrossover operation.

Again the positions of two non-active genes i, j, k, and l are selected with the same above probabilities. Thereby, performing the two aforementioned operations, two individuals will be created in the next population.

6) THE EGA AGORITHM

After having explained the main phases of the EGA algorithm in the previous sections, we summarize Algorithm 1 in the following steps:

1) Initialize step to 1,
2) Construction of the initial population $P^{(0)}$,
3) Calculation of the fitness of each individual among M_s ones,
4) Copy the N_e best individual to the next generation after classifying individuals' fitness in a decreasing manner.
5) For the rest of the $M_s - N_e$ individuals, the following operations are performed,
   - Selection of two parents based on the SUS method,
   - Applying the epimutation,
   - Applying the epicrossover,
   - Adding two children to the next generation,
6) If step $< N_g$ then go to 3.

III. RESULTS AND DISCUSSIONS

In this section, we investigate the performance of the proposed IDS EGA-based algorithm for various parameters to achieve the best classifier. First, we start deleting any redundancy reported in the dataset, and then we build both training and test sets by following this process: import E_t and E_r from KDD-NSL[1] choosing to import "KDDTrain ±20Percent" for E_r representing 20% of the global training file subset. The full NSL_KDD test set that includes all attack-type will be the E_t. Initially, the EFL has been set based on the comparison of D_r provided by numerous selection methods evaluated on J48 [45] and Naive Bayes (NB) [47] decision tree methods with the help of Weka software. To this end, CFS selection method, applied on NB, is being the optimum method, as outlined in Table 4, providing a better DR and precision, allowing to select active genes participating mostly in the intrusion's detection [48].

The parameters values for the simulation throughout the paper are summarized in Table 5. To find the optimum parameters' values for EGA, one can start by finding those of GA algorithm. The effect of different EGA parameters on the overall performance alongside a comparison between EGA and GA results are provided and discussed.

[1]http://www.unb.ca/cic/datasets/nsl.html

TABLE 4. D_r and P for CFS, InfoGain and RG using $J48$ and NB.

	J48		NB	
Metric	D_r	P	D_r	P
CFS	0.965	0.96	0.986	0.998
InfoGain	0.907	0.913	0.959	0.96
Gain Ratio	0.925	0.93	0.96	0.975

TABLE 5. Parameter settings for the proposed EGA-based algorithm.

Parameter	*Value/interval*
M_s	$[10, 1000]$
N_g	$[100, 1000]$
p	41
$\phi_{\min}$	$]0, 1[$
N_e	$]10, 1000[$
$\mathcal{L}$ by CFS	$\{3, 4, 5, 6, 12, 26, 29, 30, 37, 38\}$
N_{EF}	5
(w_1, w_2)	$(0.8, 0.2)$

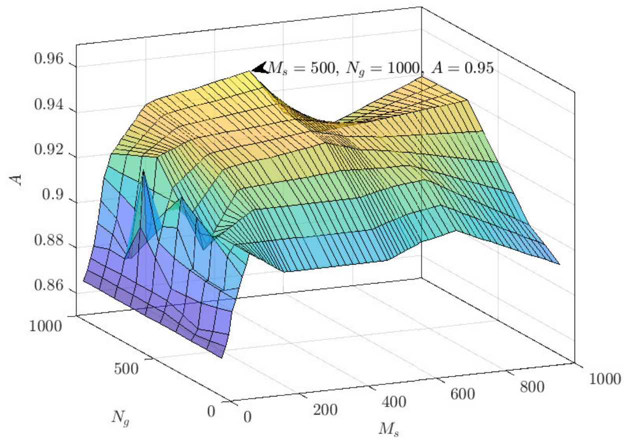


FIGURE 6. GA's accuracy vs N_g and M_s.

A. OPTIMUM PARAMETERS FOR GA

1) EFFECT OF M_s AND N_g ON GA

Fig. 6 shows that the accuracy of GA versus the population size M_s and the number of generation N_g. Obviously, such a metric is enhanced with the increase of both parameters. Particularly, it can be ascertained that its highest value is reached for $N_g = 1000$ and $M_s = 500$, representing the optimal values of these two parameters. Interestingly, in this latter interval, a slight steady of the accuracy metric which intervals of both N_g and M_s.

2) EFFECT OF MUTATION AND CROSSOVER PROBABILITIES IN GA

The impact of the probabilities of mutation p_m and that of crossover p_c on the GA's accuracy is presented in Fig. 7. The optimum values such two parameters are those maximizing the accuracy. That is, $p_m = 0.024$ and $p_c = 0.5$ allowing to reach the maximum value of A, i.e., $A_{\max} = 0.95$.

B. OPTIMUM PARAMETERS FOR EGA

1) EFFECT OF M_s AND N_g ON EGA

Fig. 8 shows that the accuracy of EGA versus the population size M_s and the number of generation N_g. It is clearly seen

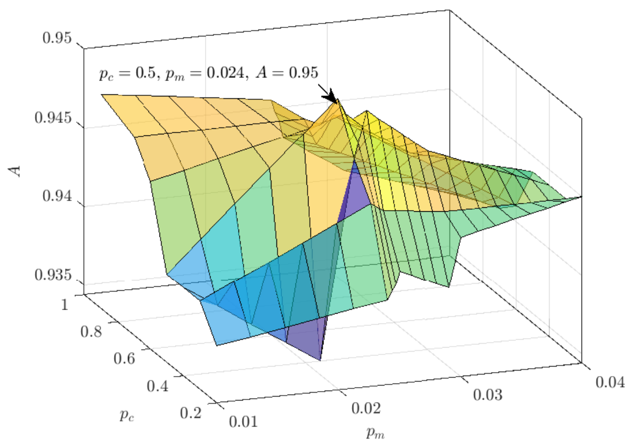


FIGURE 7. Mutation and Crossover probabilities for GA.

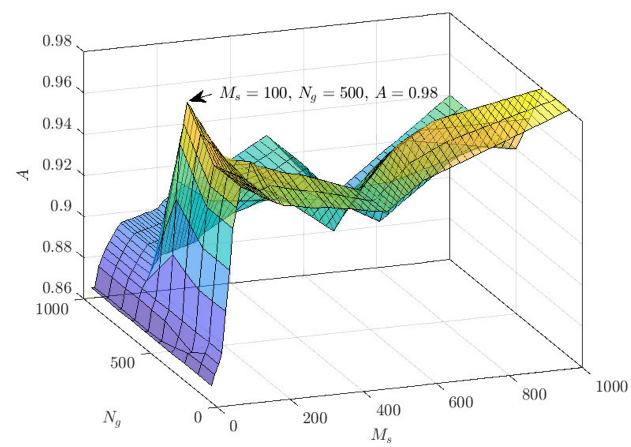


FIGURE 8. Optimum solution target for EGA.

that the values $N_g = 500$ and $M_s = 100$ represent the optimal values ensuring the maximum value of the system's accuracy.

2) IMPACT OF EFL

The EGA performance using two different EFLs, namely $\mathcal{L}_1 = \mathcal{L}$ outlined in Table 5, and its complementary $\mathcal{L}_2 = \{1 \ldots p\}\backslash\mathcal{L}$ is summarized in Table 6. Of note, $\mathcal{L}_1$ is optimized with the help of CFS method. The choice of CFS as the best feature selection method is justified in Table 4 by calculating the D_r and P metrics through J48 and NB algorithms. The results obtained show that the metrics of CFS are the most interesting compared to those of InfoGain and Gain Ratio. This well-chosen selection method will help select the best candidates for active genes. Evidently, the use of EFL, specifically $\mathcal{L}_2$, is contributing to the enhancement of both accuracy and detection rate with 99% provided by our proposed EGA as shown at Table 6. The $\mathcal{L}_2$ list then relates to the list of inactive genes through which the mutation and crossover operations will be applied. In this way we increase the precision of the algorithm by increasing the number of precise rules of the E_s in addition to those obtained through the active genes.

TABLE 6. Performance of EGA for different EFLs.

EFL	D_r	A
$\mathcal{L}_1$	86%	96%
$\mathcal{L}_2$	99%	99%

TABLE 7. Optimum values for the proposed EGA parameter.

Parameter	*Value*
M_s	100
N_g	500
$\phi_{\min}$	0.6
N_e	60
Feature Selection Method	CFS
$\mathcal{L}_2$	$\{1..41\}\backslash\{3, 4, 5, 6, 12, 26, 29, 30, 37, 38\}$
f_{min}	4

TABLE 8. Comparison of the performance of various algorithms-based detection technique.

Algorithms	D_r	A	F_p
EGA	98%	98%	17%
GA	86%	89%	8%
Naivebayes	87%	89%	9%
J48	90%	90%	2%
SVM	94%	96%	2%
WNN	93%	93.3%	0.1%
CNN	98%	93.1%	0.31%

3) COMPARISON GA VERSUS EGA

First, we performed the simulation for GA-based algorithm for IDS by looking for the best optimal value in terms of accuracy by varying jointly N_g and M_s. The values $A = 0.95$, $N_g = 1000$, and $M_s = 500$ show that the optimum GA attained for a high number of iterations (i.e, N_g), a high value of p_c, and a small p_m's value. In a second step, we optimized the performance of EGA by varying mutually N_g and M_s as shown in Fig. 8. One can ascertain that the maximum value of the accuracy (i.e. $A = 0.98$) is reached for $M_s = 100$ and $N_g = 500$. It is worthy to mention that the optimum value of N_g obtained for EGA-based algorithm is lesser than the one in its GA counterpart, while the accuracy is further enhanced, proving the usefulness of such proposed algorithm in terms of both performance and computational complexity.

In a similar manner, all the remaining EGA parameters are optimized. For the sake of simplicity, the corresponding figures are omitted, whereas the optimum values obtained by simulation are summarized in Table 7.

The receiver operating characteristic curve (ROC), presented in Fig. 9, measures the performance of the sensitivity S (also known as the true positive rate) versus the $\overline{S}_p$, given in Table 1, for both GA and EGA. Such a metric has been evaluated relied on the optimum values of both algorithms' parameters obtained in the previous phases. Obviously, the sensitivity computed based on EGA algorithm outperforms that evaluated relied on its GA counterpart over the entire range of $\overline{S}_p$.

4) COMPARISON BETWEEN EGA AND OTHER ALGORITHMS

Table 8 outlines various metrics, namely D_r, A and F_p for numerous algorithms-based detection techniques, i.e., EGA, GA, Naivebayes,J48, and SVM. One can ascertain that EGA demonstrates high values of both accuracy and D_r compared

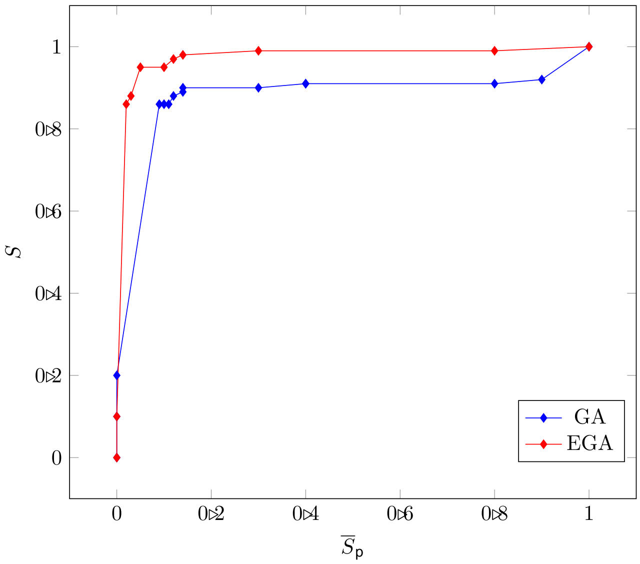


FIGURE 9. RoC curve comparison EGA and GA.

to the remaining algorithms. Thus, EGA is a good candidate for detection and prevention against unknown attacks. Furthermore, the false positive rate, i.e., F_p obtained using EGA, is slightly higher than the ones corresponding to other considered methods. Nevertheless, this undesirable value can be improved by further reducing the number of iterations. However, this can impact negatively on the previous two metrics. To this end, we thought that these values represent a good trade-off between detection performance and the false positive rate. Also in the field of detection of attacks in the IDS, the risk that is not desirable is an increase in the false negative $F_N R$ because if we consider normal traffic as an attack it is less critical than if we consider malicious traffic as normal. The $F_N R$ rate is calculated by equation in Table 2 and the value obtained for proposed EGA algorithm is 3%.

On the other hand, the computational complexity is a another element of paramount importance that highlights the reliability of any proposed algorithm. Towards this end, Fig.9 depicts the running time of EGA-based algorithm compared with three main concurrent ones, namely GA, Fuzzy [49], SVM [50], Wavelet Neural Network (WNN) [36] and Convolutional Neural Network (CNN) with one layer [37]. Mainly, again EGA outperforms its counterpart ones in terms of time complexity, which makes from it a promising candidate for the unknown attack detection.

To summarize this section and through the experimentation phase presented, we tried to explain the important phases to find the optimized parameters of GA and EGA in order to build a correct basis of comparison. These phases took into account the Effect of M_s and N_g on EGA and GA, the Effect of mutation and crossover probabilities in GA and the impact of the EFL of the EGA. Once these parameters are defined, the comparison started first with the EGA and the original GA algorithm. This comparison shows through the RoC curve and the two graphs of Fig. 8 and Fig. 9 that the EGA outperforms in terms of accuracy and detection rate, as well as the execu-

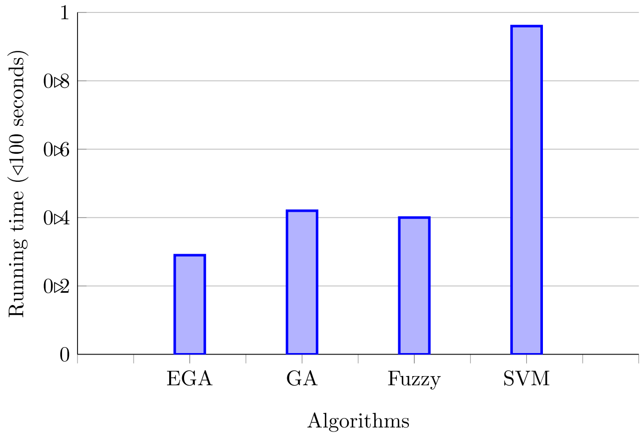


FIGURE 10. Processing time comparison.

tion time. We extended in a second step the comparison with several other algorithms shown in Table 8 and Fig. 10.

IV. CONCLUSION AND FUTURE WORKS

In this article, an EGA-based detection technique for arbitrary attacks was presented and optimized. This EGA algorithm was applied to the IDS using the KDD-NSL dataset. To define the epigenetic factor list, we adopted CFS method which proved its effectiveness rather than its GR and InfoGain counterparts. Next, the crossover and mutation operators are limited to the genomes defined in such list, based on it, the remaining optimum EGA parameters are retrieved. The numerical results prove that the IDS accuracy and detection rate obtained based on the proposed algorithm outperforms its GA counterpart, even with smallest rate (i.e., 3%), allowing to strengthening the security when dealing with destructive attacks. Even if the gap is only 2% to 3% more, but in terms of security, this gap is very considerable to face the smallest flaw leading to destructive attacks. Moreover the computational complexity of the EGA-based IDS is lesser than the one of GA, even most other concurrent algorithms, rending from it a a suitable algorithm for various secure applications. In the future research directions, we will improve the rate of false positive by reducing the number of iterations significantly and ensuring the other performance metrics (D_r and A). We will also try to combine more methods to build a more precise EFL by selecting genes as the best candidates. We also plan to apply our approach to other datasets other than KDD_NSL.

REFERENCES

[1] Q. M. Alzubi, M. Anbar, Z. N. M. Alqattan, M. A. Al-Betar, and R. Abdullah, "Intrusion detection system based on a modified binary grey wolf optimisation," *Neural Comput. Appl.*, vol. 32, pp. 6125–6137, Feb. 2019.

[2] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Comput. Netw.*, vol. 76, pp. 146–164, Jan. 2015.

[3] S. H. Amer and J. Hamilton, "Intrusion detection systems (IDS) taxonomy-a short review," *Defense Cyber Secur.*, vol. 13, no. 2, pp. 23–30, 2010.

[4] B. A. Tama, M. Comuzzi, and K. Rhee, "TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," *IEEE Access*, vol. 7, pp. 94497–94507, 2019.

[5] M. Sazzadul Hoque, "An implementation of intrusion detection system using genetic algorithm," *Int. J. Netw. Secur. Appl.*, vol. 4, no. 2, pp. 109–120, Mar. 2012.

[6] A. Zarrabi and A. Zarrabi, "Internet intrusion detection system service in a cloud," *Int. J. Comput. Sci.*, vol. 9, no. 5, p. 2, Sep. 2012.

[7] A. Davahli, M. Shamsi, and G. Abaei, "Hybridizing genetic algorithm and grey wolf optimizer to advance an intelligent and lightweight intrusion detection system for IoT wireless networks," *J. Ambient Intell. Humanized Comput.*, Apr. 2020. [Online]. Available: https://rd.springer.com/article/10.1007/s12652-020-01919-x

[8] M. Crosbie and G. Spafford, "Applying genetic programming to intrusion detection," in *Proc. AAAI Fall Symp. Genetic Program.*, Cambridge, U.K., Feb. 1995, pp. 1–8.

[9] M. Hossain, S. M. Bridges, and R. B. Vaughn, "Adaptive intrusion detection with data mining," in *Proc. IEEE Int. Conf. Syst., Man Cybern. Conf. Theme - Syst. Secur. Assurance*, Washington, DC, USA, Oct. 2003, pp. 3097–3103.

[10] W. Lu and I. Traore, "Detecting new forms of network intrusion using genetic programming," in *Proc. Congr. Evol. Comput. (CEC)*, 2003, pp. 475–494.

[11] Y. Gong, Y. Fang, L. Liu, and J. Li, "Multi-agent intrusion detection system using feature selection approach," in *Proc. 10th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Kitakyushu, Japan, Aug. 2014, pp. 528–531.

[12] A. Goyal and C. Kumar, "GA-NIDS: A genetic algorithm based network intrusion detection system," Elect. Eng. Comput. Sci., North West Univ., Potchefstroom, South Africa, Tech. Rep., 2008.

[13] P. G. Majeed and S. Kumar, "Genetic algorithms in intrusion detection systems: A survey," *Int. J. Innov. Appl. Stud.*, vol. 5, no. 5, pp. 233–240, Mar. 2014.

[14] S. Birogul, "EpiGenetic algorithm for optimization: Application to mobile network frequency planning," *Arabian J. Sci. Eng.*, vol. 41, no. 3, pp. 883–896, Oct. 2015.

[15] B. Kenwright, "Epigenetics & genetic algorithms for inverse kinematics," *Experim. Algorithms*, vol. 9, no. 4, p. 39, Aug. 2014.

[16] P. Cubas, C. Vincent, and E. Coen, "An epigenetic mutation responsible for natural variation in floral symmetry," *Nature*, vol. 401, p. 157–161, Sep. 1999.

[17] B. Kirkpatrick, "Computer algorithm uses epigenetics to identify aging genes," Epigenetic Biotechnol. Company EpiGentek, New York, NY, USA, Tech. Rep., May 2014.

[18] F. Gottwalt, E. Chang, and T. Dillon, "CorrCorr: A feature selection method for multivariate correlation network anomaly detection techniques," *Comput. Secur.*, vol. 83, pp. 234–245, Jun. 2019.

[19] A. Mark, "Correlation-based feature selection for machine learning," M.S. thesis, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.

[20] B. Azhagusundari and A. S. Thanamani, "Feature selection based on information gain," *Int. J. Innov. Technol. Exploring Eng.*, vol. 2, no. 2, pp. 18–21, 2013.

[21] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manage.*, vol. 2, no. 2, pp. 271–277. Dec., 2010.

[22] A. G. Fragkiadakis, E. Z. Tragos, T. Tryfonas, and I. G. Askoxylakis, "Design and performance evaluation of a lightweight wireless early warning intrusion detection prototype," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 1, pp. 1–18, Dec. 2012.

[23] T. Hamed, R. Dara, and S. C. Kremer, "Intrusion detection in contemporary environments," in *Computer and Information Security Handbook*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 109–130.

[24] C. Azad and V. K. Jha, "Data mining in intrusion detection: A comparative study of methods, types and data sets," *Int. J. Inf. Technol. Comput. Sci.*, vol. 5, no. 8, pp. 75–90, Jul. 2013.

[25] M. Cordy, S. Muller, M. Papadakis, and Y. Le Traon, "Search-based test and improvement of machine-learning-based anomaly detection systems," in *Proc. 28th ACM SIGSOFT Int. Symp. Softw. Test. Anal. (ISSTA)*, New York, NY, USA, Jul. 2019, pp. 158–168.

[26] K. Huang, Q. Zhang, C. Zhou, N. Xiong, and Y. Qin, "An efficient intrusion detection approach for visual sensor networks based on traffic pattern learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 10, pp. 2704–2713, Oct. 2017.

[27] A. Saied, R. E. Overill, and T. Radzik, "Detection of known and unknown DDoS attacks using artificial neural networks," *Neurocomputing*, vol. 172, p. 385–393, Jan. 2016.

[28] M. Botha and R. von Solms, "Utilising fuzzy logic and trend analysis for effective intrusion detection," *Comput. Secur.*, vol. 22, no. 5, p. 423–434, Jul. 2003.

[29] B. Hu, C. Zhou, Y. C. Tian, Y. Qin, and X. Junping, "A collaborative intrusion detection approach using blockchain for multimicrogrid systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 8, pp. 1720–1730, Aug. 2019.

[30] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," *Knowl.-Based Syst.*, vol. 136, pp. 130–139, Nov. 2017.

[31] P. Stavroulakis and M. Stamp, *Handbook of Information and Communication Security*. New York, NY, USA: Springer-Verlag, 2010.

[32] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, Dec. 2019.

[33] N. Peiravian and X. Zhu, "Machine learning for Android malware detection using permission and API calls," in *Proc. IEEE 25th Int. Conf. Tools with Artif. Intell.*, Herndon, VA, USA, Nov. 2013, pp. 300–305.

[34] G. Creech and J. J. Hu, "A semantic approach to host-based intrusion detection systems using contiguousand discontiguous system call patterns," *IEEE Trans. Comput.*, vol. 63, no. 4, pp. 807–819, Apr. 2014.

[35] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, p. 16–24, Jan. 2013.

[36] Y. Hamid, F. A. Shah, and M. Sugumaran, "Wavelet neural network model for network intrusion detection system," *Int. J. Inf. Technol.*, vol. 11, no. 2, pp. 251–263, 2018.

[37] J. Kim, H. Kim, M. Shim, and E. Choi, "CNN-based network intrusion detection against denial-of-service attacks," *Electronics*, vol. 9, p. 916, Jun. 2020.

[38] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, 2017, pp. 1222–1228.

[39] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly detection in wireless sensor networks: A survey," *J. Netw. Comput. Appl.*, vol. 34, pp. 1302–1325, Jul. 2011.

[40] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Gener. Comput. Syst.*, vol. 100, pp. 779–796, Nov. 2019.

[41] N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, and H. F. M. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," in *Proc. Australas. Comput. Sci. Week Multiconf.*, Jan. 2018, pp. 1–6.

[42] M. Ezzarii, H. El Ghazi, H. Elghazi, and T. Sadiki, "Performance analysis of a two stage security approach in cloud computing," in *Proc. Int. Conf. Cloud Technol. Appl. (CloudTech)*, Marrakesh, Morocco, Jun. 2015, pp. 1–7.

[43] S. E. Benaicha, L. Saoudi, S. E. B. Guermeche, and O. Lounis, "Intrusion detection system using genetic algorithm," in *Proc. Sci. Inf. Conf.*, London, U.K., Aug. 2014, pp. 564–568.

[44] J. S. Vithalpura and H. M. Diwanji, "Analysis of fitness function in designing genetic algorithm based intrusion detection system," *Int. J. Sci. Res. Develop.*, vol. 3, no. 1, pp. 86–92, 2015.

[45] U. Bashir and M. Chachoo, "Performance evaluation of J48 and Bayes algorithms for intrusion detection system," *Int. J. Netw. Secur. Appl.*, vol. 9, no. 4, pp. 1–11, Jul. 2017.

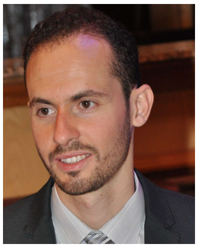
[46] J. McHugh, "Intrusion and intrusion detection," *Int. J. Inf. Secur.*, vol. 1, pp. 14–35, Jan. 2014.

[47] S. Mukherjee and N. Sharma, "Intrusion detection using naive Bayes classifier with feature reduction," *Procedia Technol.*, vol. 4, pp. 119–128, Jan. 2012.

[48] T. Tollefsbol, *Handbook of Epigenetics: The New Molecular and Medical Genetics*. Amsterdam, The Netherlands: Elsevier, 2017. [Online]. Available: https://www.elsevier.com/books/handbook-of-epigenetics/tollefsbol/978-0-12-805388-1?aaref=https%3A%2F%2Fwww.google.co.ma%2F

[49] Y. Danane and T. Parvat, "Intrusion detection system using fuzzy genetic algorithm," in *Proc. Int. Conf. Pervas. Comput. (ICPC)*, 2015, pp. 1–5.

[50] I. Ahmad, M. Hussain, A. Alghamdi, and A. Alelaiwi, "Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components," *Neural Comput. Appl.*, vol. 24, p. 1671–1682, Jun. 2013.



MEHDI EZZARII received the Engineering degree in the specialty of security and network systems from the National Institute of Posts and Telecommunications (INPT), Morocco, in 2009, where he is currently pursuing the Ph.D. degree with the Communications Systems Department. He has worked in a multinational company integrating IT security solutions. His research interests include cybersecurity, networking, cloud, and cognitive radio networks.

HAMID EL GHAZI received the Ph.D. degree in computer science from Paris1 University. He has worked as a Senior Consultant for international companies such as ALSTOM and THALES. He is currently an Assistant Professor with INPT. His research interests include information systems security, big data, and SOA. He is a PC Member of the ICEIT and ICTMOD conferences.

HASSAN EL GHAZI received the M.S. degree in wireless communications and the Ph.D. degree in electrical engineering from the University of Polytechnique Hauts-de-France, France, in 2004 and 2008, respectively. He is currently an Associate Professor with the Communications Systems Department, INPT, Morocco. He has advised many Ph.D. and graduate students at both INPT and M5U. So far, his research contributions have culminated in more than 50 articles in a wide variety of international journals and conferences. His main research interests include cyber-physical security, smart grid systems, and cognitive radio networks. He has served as a Reviewer for IEEE Access, Elsevier, and Springer. He has served as the General Chair for IWTSC 2018 conference and the Conference Chair for NISS 2019 conference.

FAISSAL EL BOUANANI (Senior Member, IEEE) was born in Nador, Morocco, in 1974. He received the M.S. and Ph.D. degrees in network and communication engineering from Mohammed V University (M5U), Rabat, Morocco, in 2004 and 2009, respectively. He has served as a Faculty Member with the University of Moulay Ismail, Meknes, from 1997 to 2009, before joining the National High School of IT/ENSIAS College of Engineering, M5U, in 2009, where he is currently an Associate Professor. He has advised many Ph.D. and master's students at both M5U and Moulay Ismail University. So far, his research efforts have culminated in more than 75 articles in a wide variety of international journals and conferences. His current research interests include channel coding and performance analysis of wireless communication systems. His Ph.D. thesis was awarded the best one by M5U, in 2010. He has served as the TPC Chair for the ICSDE conferences, the General Co-Chair for the ACOSIS 2016 and CommNet 2018 conferences, and the General Chair for the 2019 and 2020 CommNet conferences. He is also an Associate Editor of IEEE Access and *Frontiers in Communications and Networks* journals.