Comparisons on Intrusion Detection and Prevention Systems in Distributed Databases

M.GUCLU, C.BAKIR, V.HAKKOYMAZ and B.DIRI

***Abstract*—With the use of distributed systems, different users can instantly access data from different locations and perform some operations on the data. However, the unauthorized access of multiple users to the system from different points at the same time can lead to dangerous results in terms of data security and confidentiality of the data. This study is based on intrusion detection and prevention systems built on distributed databases and classifies the methods used to analyze and evaluate successes comparatively. It is observed that the artificial immunity algorithm we have described in artificial intelligence techniques, which is one of the methods classified as three categories, gives more successful results compared to the other techniques mentioned in the data mining and statistical methods.**

Index Terms—**Artifical intelligence method, Data mining, Distributed database, Intrusion detection system, Intrusion prevention system, Statistical method.**

I. INTRODUCTION

DATA ACCESS and data communication have become very easy thanks to the developing technology. Today, as a result of increasing of the internet usage rate, expanding of the area used and increasing of the diversity of the work done, the issue of security has become a serious research topic. With the use of distributed systems, different users can gain instant access to data from different locations and perform a number of operations on the data. Especially in the area of security, a new threat is encountered every day, and accordingly, a rapid development in security measures is taking place. In this sense, many new applications have been developed for the purpose of ensuring the security of the computer systems, preventing unauthorized access, and developing mechanisms for authentication and access controls within the scope of the information security. The acceleration of development processes in internet and communication areas has also led to more systems that malicious attackers can harm, and accordingly, it also revealed the possibility that these attackers may have obtained much more information. For this reason, significant increases in the number of intrusions/attacks and in the use of intrusion detection and prevention methods have been observed.

Often the vast majority of intrusions are carried out by exploiting the vulnerabilities of the used systems. In order to prevent such intrusions, it is necessary to create a safe environment and ensure that intrusions are detected and prevented in a timely manner. No matter how secure a system is, what is important here is to detect attacks early and prevent possible infiltrations. In this context, there are various intrusion detection and prevention systems that have been developed for each possible type of attack from different sources and have been the subject of research and development studies from past to present. In this study, the point that the intrusion detection and prevention systems reached during their development process was mentioned, and by classifying the systems in question, within-class and between-classes performance evaluations were carried out. It should be also noted that in this study, the words “attack” and “intrusion” were used interchangeably.

The rest of the article is organized as follows: related studies in the 2nd section, data set introduction in the 3rd section, material and method in the 4th section, findings in the 5th section, evaluation in the 6th section, and the discussion and conclusion are given in the final section.

II. RELATED WORKS

Quickprop neural networks, a multi-agent statistical prediction system, have been developed to predict database attacks beforehand and to detect vulnerabilities before an attack occurs [1]. In this study, the Pearson correlation coefficient was used to calculate hidden layers and it was tried to identify users who did not have authority over a bank data. Abnormal and incorrect user behavior occurred in a short period was able to be detected.

MEHMET GUCLU, is with Department of Computer Engineering University of Yildiz Technical University, Istanbul, Turkey,(e-mail: mehmetguclu007@gmail.com).

https://orcid.org/0000-0001-2306-6008

CIGDEM BAKIR, is with Department of Computer Engineering University of Yildiz Technical University, Istanbul, Turkey, (e-mail: cigdem.bakir@igdir.edu.tr).

https://orcid.org/ 0000-0001-8482-2412

VELI HAKKOYMAZ, is with Department of Computer Engineering University of Yildiz Technical University, Istanbul, Turkey, (e-mail: vhkoymaz@yildiz.edu.tr).

https://orcid.org/ 0000-0002-3245-4440

BANU DIRI, is with Department of Computer Engineering University of Yildiz Technical University, Istanbul, Turkey, (e-mail: diri@yildiz.edu.tr).

https://orcid.org/ 0000-0002-4052-0049

Manuscript received August 12, 2019; accepted October 16, 2019.
DOI: 10.17694/bajece.605134

Another study carried out to detect incorrect user behaviors is the use of genetic algorithms [2]. Based on neural networks, genetic algorithm makes classification by the rules obtained by creating a variety of rules from network properties. The results of this study were presented as comparative to other studies. However, as in the previous one, this study also offered a solution for attacks that could occur in the short term.

The study of predicting and preventing intrusions by using the Hidden Markov Model is another study [3]. The Hidden Markov model is a classification algorithm performed to find hidden States based on given states. Distributed data communicate with each other on very large networks and therefore they are open to serious attacks. In this study, risk analysis was carried out using fuzzy technique and packet rate sent as dangerous was determined. In addition, attacks that would pose risks for distributed environments were attempted to be identified.

Deng and colleagues developed a Support Vector Machine (SVM)-based system for [4] wireless and ad hoc networks intrusion detection system. Security issues for wireless and ad hoc networks were attempted to be solved by the SVM method. In that study, hierarchical and complementary distributed systems were developed for two common Denial of Service (DoS) attacks and performance measurements were made. The effect of DoS attack on speed, distance changes between communication distances, and system performance of location information was observed.

For intrusion prediction and detection systems, Jemili et al. [5] introduced a new approach based on hybrid propagation that combines probabilities with the Bayesian network. The hybrid propagation approach is used to notice abnormal connections occurring both normally and in the network. The purpose of that study was to reveal possible attack plans, scenarios and the relationships between them. In addition, this study combined the host-based intrusion detection systems and the network-based intrusion detection systems to ensure data consistency.

Hu et al. [6] developed a method combining Particle Swarm Optimization (PSO) and SVM techniques for network attacks that often changed and were unnoticed due to the variances occurring in the network. They also attempted to ensure finding of the attacks for each node of the distributed database by using the Gaussian Mixture Model which was an Adaboost-based intrusion detection algorithm. However, this approach was not able to fully identify all types of attacks, particularly new types of attacks.

Abraham et al. [7] aimed the use of genetic algorithms in the intrusion detection systems. In the study, an automatic intrusion detection program was developed by using the given features. The output program was small and simple. Whereas all the features of many machine learning methods are used for intrusion detection, few features of genetic programming was used in the proposed study. A study developing an intrusion detection system for wired networks by using genetic programming in the future was proposed.

In their study, using the artificial neural network, Sağıroğlu et al. [8] aimed to determine what attack method the packets flowing on a network applied. In order to find "Neptune" and "the ping of death" from these attacks, an artificial neural network model with Multi-Layer Sensor was used. By taking the DARPA Data Sets as an example, they trained for the networks. As a result of that study, detection of DoS attacks that could come from the internet has been successfully achieved.

There are many methods developed and used to date for intrusion detection systems, such as data mining methods, artificial neural networks, and statistical methods.

III. DATASET DEFINITION

One of the most widely used datasets in applications related to intrusion detection systems is the "KDD Cup'99" (Knowledge Discovery and Data Mining Cup'99) dataset. On this dataset, the applicability of multilayer artificial neural networks has been tested and performance analyses have been conducted with parallel programming [9]. The KDD-Cup 1999 data is a data set containing different types of attacks, such as DoS, R2L, U2R, and Probe. This data set contains a total of 972780 samples. In DoS attack-type, the attacker keeps the server engaged by constantly sending fake requests to the server. In this case, the server becomes unable to respond to the requests of the users requesting by formal means. In the R2L (Remote to Local Attack) attack type, the attacker, who does not have access permission, sends packets over the network. Thus, he/she uses the necessary data without permission by providing access to the system. In U2R (User to Root Attack) attack type, the attacker seizes the user's password and accesses the system as if he/she is an administrator. In Probing Attacks, on the other hand, the attacker disrupts the security controls in the system and receives the data he wants over the network.

The KDD-Cup 1999 dataset is a dataset derived from the DARPA BSM (Defense Advanced Research Projects Agency Basic Security Module) data. The DARPA BSM data was generated as a result of simulation of a network traffic belonging to the American Air Force and has been widely used in Intrusion Detection. It includes a total of 38 different types of attacks, including 7 weeks of training and 2 weeks of test data. The contents of the data set consist of tcpdump files taken from night and day network listeners of the attacked machines, log records taken from the attacked machines, and files taken from the security module [10].

IV. MATERIAL AND METHODS

The methods used in the Intrusion Detection and prevention systems in distributed databases are gathered into three main groups as shown in Table 1: Data mining methods, Statistical methods, and Artificial Intelligence methods. The techniques in each method are included in the following subsections. In addition, the performance values of the techniques were evaluated as within-group and between-groups.

TABLE I
INTRUSION DETECTION AND PREVENTION SYSTEMS

	Methods Used
Data Mining	K-Mean Clustering k-Nearest Neighbor (k-NN) Decision Tree Support Vector Machine (SVM) Association Rules
Statistical Methods	Learning Vector Quantization (LVQ) Hidden Markov Models (HMM) Naive Bayes Fuzzy Logic
Artificial Intelligence Methods	Genetic Algorithm Artificial Neural Networks Artificial Immune Techniques

A. *Data Mining Methods*

Under this group, K-Mean Clustering, k-Nearest Neighbor (k-NN), Decision Tree, Support Vector Machine (SVM), Rules of Association are widely used.

1) *K-Mean Clustering*

The K-mean clustering technique is one of the non-educational learning methods that group objects according to their similarities. K-mean clustering, which is a division-based method, calculates N objects based on k distance of cluster center and incorporates the object into the cluster where it is near the center of the cluster. The cluster center is initially determined by averaging one or more random instances, and is recalculated in each iteration. Each time, similarities of all data is found according to the new cluster centers. In this way, similar objects are taken into the same cluster as other objects are taken into different clusters. These steps iteratively repeat. The steps end when the clustering error rate (objective function) is minimal. In attack detection and prevention, k-Mean divides N attack types into k amount of clusters. While the within-cluster similarities of the clusters obtained as a result of the division process are maximum, their between-cluster similarities should be minimum. The success and performance of this method vary according to the number of k clusters randomly selected at the beginning, cluster centers, and similarity criteria used. For this reason, it has not produced very successful results in finding the false alarm rate [11]. The false alarm rate is the proportion between classifying a normal data (un-attacked) falsely as attacked data and classifying an attacked data as normal data.

The K-mean method has been used to reduce computational complexity and increase the classification success in the intrusion detection and prevention systems [12]. In the studies conducted on the KDD-Cup 1999 data, the K-Mean Clustering method achieved approximately 96% success [11]. On the other hand, using the KDD-Cup 1999 data, Nadiammai et al. [13] achieved 92.05% success with the K-Mean Clustering method.

2) *k-Nearest Neighbor*

The k-Nearest Neighbor (kNN) method is the oldest and simplest educational classification method. It calculates the distance between the given input vectors and selects the class of k nearest neighbors. Different classes may exist for different k values. For this reason, the k parameter is very important for classification time and classification accuracy [10]. In this method, the distances of the data, whose class is intended to be found, to all known data are calculated. Euclidean distance is generally used as the distance criterion. Randomly, k neighbor number is determined. The class in which the data is included is determined by looking at the nearest k neighbors. Unlike the K-Mean Clustering method, the classes of the data set used are known in this method, and the newly arrived data is found by looking at these classes. In intrusion detection, the kNN algorithm has been used for the classification of data samples belonging to normal and aggressive species [10]. The distance of the data, whose class to be found, to all data is calculated. By looking at the mean of k data, the attack class of the data at hand is determined. The number of neighbors (k) and whether the number of classes of the data is balanced affect the success and performance of this method.

DARPA attempted to detect samples belonging to the attacker class by looking at the frequency of system calls on the BSM data and provided a low false positive rate. The false positive rate is the rate of finding an attack-free data as an attack-type, that is, it is the rate of false classification of a data.

kNN classifier has been used to reduce the false alarm rate that will identify false and non-normal data in intrusion detection and prevention [14]. In this study, 5 close neighborhoods of false alarm models of normal and aggressive data were taken and a high degree of success was achieved. In the studies conducted on the KDD-Cup 1999 data, kNN (the nearest neighbor method) achieved approximately 97% success [15]. In addition, in the study conducted by Aburomman et al. [16] on KDD-Cup 1999 data, 96% success was also achieved. On the other hand, Chen and his colleagues [17] achieved 91.96% success in their study carried out on the KDD-Cup 1999 data.

3) *Decision Tree*

The decision tree technique is one of the first classification algorithms used in data mining. Classification results are obtained more easily and more quickly. In this technique, columns in each data set show features, while values for those features are defined in each row. In addition, classes are defined for each record by depending on the values of the features. The first approach in the decision tree is to select features. Then the classes are divided according to this selected feature and this process is iteratively repeated. Each node shows a feature and these nodes have child nodes [18]. In short, the decision tree technique is performed in two stages. In the first stage, a tree consisting of root and child nodes is created. In the second stage, various classification rules are issued according to the structure of this tree. These classification rules show the

nodes remaining between the root node of the tree and its leaves. Figure 1 shows how to create a sample decision tree.

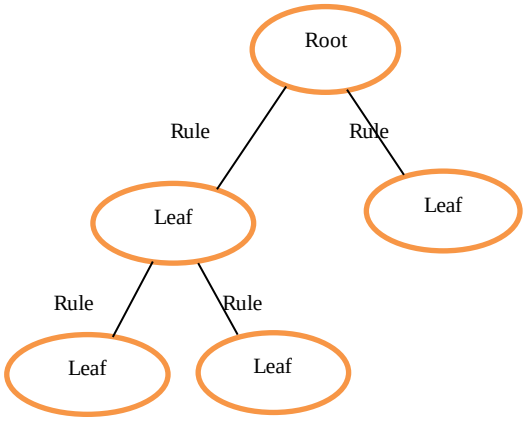

Fig.1. The creation of the decision tree

In network-based intrusion detection and prevention systems, each node displays the types of attacks or normal events in the user or data. The decision tree solves the classification problem by modeling the data set. It predicts future types of attacks based on the model created in false intrusion detection. Provides high performance in real-time intrusion detection. It uses a variety of rule-based models. In the detection of new types of attacks, it gives quite successful results. In the studies conducted on the KDD-Cup 1999 data, the decision tree method achieved approximately 96% success [18]. Rachburee et al. [19] carried out an intrusion detection system on the KDD-Cup 1999 data. In that study, feature selection was made in this data set by the Chi-Square method, and attack types were classified by Decision Trees and Artificial Neural Networks.

4) Support Vector Machines

In analyzing data, the support vector machines (SVM) is a technique based on the supervised learning used in classification. Although SVM was initially used to solve the two-class data problems, it was later expanded and used in the multi-class data problems. In a two-class data problem, a model is created to separate the two classes from each other. This model is obtained by creating a function. Which class the new incoming data belongs to is determined according to this function. The goal in the SVM technique is to find the function that will obtain the most appropriate hyperplane that will separate the two classes from each other. In addition, the support vectors belonging to the two classes must be as maximum as possible to separate the classes from each other.

SVM is a classification algorithm that determines the class of each training vector in high-dimensional space. SVM determines the output of the system as well as classes and hyperplane that will determine the support vectors of the data. At the time of training, it determines support vectors by linear, polynomial, or sesamoid functions. The fact that SVM separates the classes varies depending on various parameters [20].

SVM was used to classify types of attacks on DARPA BSM data in attack detection and Prevention [21]. SVM has two important roles in attack detection. The first and most important is to train the system by providing real-time performance for attack detection, and to calculate the success of the system. The second role is to overcome the scaling problems that may occur in the system. Furthermore, SVM provides very successful results in high-dimensional space and complex classification problems [21]. In the studies conducted on the KDD-Cup 1999 data, the SVM method achieved approximately 99% success [15]. Chen et al. [17] achieved 92.46% success in their study conducted on the KDD-Cup 1999 data. In addition, 93.9% success was also achieved in the study carried out on the KDD-Cup 1999 data by Aburomman et al. [16].

5) Association Rules

The association rules algorithm is a data mining technique that examines association behaviors between data by looking at the historical data. Assuming that T is transaction, let us express all transactions in the database with { T_1, T_2,…,T_n } and transaction objects with { i_1, i_2,…,i_m }. The rule between the data is X →Y (c,s). Here, **s** refers to the support between the data and **c** refers to the confidence interval. Support indicates the frequency of using an object for all objects. Confidence indicates the frequency of using of an object with the other object. **s** specifies the percentage of X and Y transactions performed together; **c** specifies the ratio [22]. Association rules are most commonly used in product sales. For example, more sales can be made to the customer by finding the products sold together in the markets. The first step in this algorithm is to find frequently used data objects for each transaction and set a threshold value for the **s** confidence interval. The second step is to create appropriate rules for the data set. The main problem in the apriori algorithm is to construct a large number of rules. However, rules created on these data objects can be restricted by selecting frequently used data objects. The algorithm steps are as follows:

- Minimum support and confidence interval are determined
- Support value of each object is found.
- The smallest support value is compared to the support value of each object, and those smaller than the smallest value are discarded from the set of objects.
- Binary association rules are created and the same operations are repeated.

A new approach was carried out on the KDD-99 data set using association rules for intrusion detection and prevention [22]. With the set of fuzzy relational rule, a new classification approach was created that determines different classes. They also aimed to perform an effective algorithm that could make measurements on new data objects.

In the study conducted by Tajbakhsh et al. [22], 91% success was achieved. In the studies carried out on the KDD-Cup 1999 data, the association rule method achieved approximately 96% success [7].

Security analyses were conducted on event records by establishing relational rules between read and write transactions [23]. In this study, the performance of the system was measured by trying to determine reliable read and write transactions on a very large database.

B. *Statistical Methods*

Under this group, Learning Vector Quantization (LVQ), Hidden Markov Models (HMM), Naive Bayes, Fuzzy Logic techniques are used widely.

1) *Learning Vector Quantization*

Learning Vector Quantization (LVQ) is a supervised classification method. This technique is accomplished according to the Kohonen learning rule. Unlike other classification algorithms, the LVQ technique finds that the n-dimensional input vector can be represented by vectors of which process element. The finding of this vector varies depending on the learning rate and the maximum number of training. The process element closest to the input vector is rewarded, and the class of the input vector is approximated to that process element. Thus, this processing element is rewarded. If not, it is punished by being removed. This way feature vectors are updated.

In LVQ intrusion detection and Prevention Systems, a layer is used to classify the attack types given as input to the system. This layer is independent of the input vectors and collects the input vectors, which are similar to each other, in the same class [24]. LVQ networks have two layers, the first and the next layers. The first layer trains the attack types given as input vectors for classification. The second layer is converted to a target layer specified by the user. The classes that are trained show the subclasses of the system and the classes that must occur at the target.

LVQ has been successfully used in many applications from telecommunications to robotics [25]. Also, this classification algorithm is intuitive and has several forms such as LVQ2, LVQ3. The cost function is determined by calculating the distance between classes [25]. The learning rate and number of iterations determine performance and it is selected iteratively.

LVQ was used to classify attack types [24]. They used two layers to classify 5 types of attacks: Normal, DoS, U2R, R2L, and probe. These layers determine the subclass and the master class. The LVQ method has achieved about 81% success in the studies conducted on the KDD-Cup 1999 data [26]. Degang et al [27] achieved 76.3% success by classifying attack types through using LVQ after normalizing data on the same data set and making feature selection.

2) *Hidden Markov Model*

In the Hidden Markov Model (HMM), it is attempted to predict future states that may occur when present states are given as input to the system. HMM is a stochastic process since it produces a different output each time it is run. Also, in Markov models, the system can move from its own state to another state, depending on the probability distribution, or remain in the same state. The probabilities that occur in the state are called transition probabilities. In HMM, unlike the normal Markov model, states are not seen by the observer. However, transitions that depend on the state can be seen.

Intrusion detection systems are defense mechanisms that detect bad packets in network traffic in the distributed systems. The intrusion prevention systems are divided into two, network-based and host-based intrusion prevention systems. The main purpose of the intrusion prevention systems is to observe suspicious flow and suspicious packets in normal network traffic. Also, by rearranging the path of suspicious network traffic, it ensures the preventing of attacks such as DoS attacks. It also observes all network performance and high packet processing rates and attempts to reduce false positive rates at DoS stage [3].

HMM is widely used in many areas such as bioinformatics, handwriting recognition, image processing, and audio processing. HMM was used for attack prediction prevention in distributed database and risk assessment was done [3]. There are two stochastic processes in HMM. One is the state of the system (x_t; t=1,2,…) and the other is the observable processes (x_t; t=1,2,…). **T** refers to observations that are consecutive among intrusion detection representatives.

HMM units are as follows [3]:

- S= { s_1, s_2,…,s_N } identifies possible states in the system. There are 4 states in this study: Normal (N), Intrusion Attempt (IA), suspicious activities occurring in the network-Intrusion in Progress (IP), and occurring one or more attacks in the system-Successful Attack (SA).
- The observations that occur in the system are expressed by V= {v_1, v_2,...,v_M}. There are three observations in this study: observation of no suspicious activity (N), observation of a suspicious activity in the network (P), and observation of the successful realization of a suspicious activity (SA).
- The first distributed vector is defined in the system as π= { π_i } ve π_i= P (x_i=i). This system is assumed to be N-state.
- The transition probability matrix is P= { p_{ij} } ve p_{ij}=P (x_t=j I x_{t-1}=i). This shows the interaction between the system and the users entering the system without permission.
- The observable probability matrix defines security or quality for representatives of each intrusion detection system.

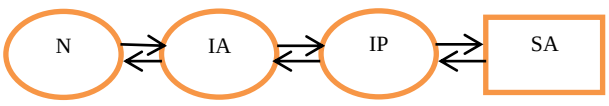

Fig.2. Security states used in HMM

The security modeling of the Hidden Markov model in networks is as in Figure 2 [3]. States indicate security states in circles. If there is a damage to the system, it is shown with SA. Observations are independent of current states. From each iteration result, the calculated probability distributions are updated [28].

3) Naive Bayes Classifiers

The Naive Bayes technique is a technique based on Bayesian probability. The Naive Bayes classifier handles events independent of each other. The most important reason for this is that the probability of a feature is not affected by the probability of the other features. It is widely used in classifying text documents and in classifying whether incoming mails to an e-mail are spam or not. It calculates the probabilities of all data to determine which class the data belongs to. The probability of one data depends on the probability of the occurrence of the realizing state of the other data and the probability of all data being. Whichever class has a greater probability value, the data is included in that class. The results of the Naive Bayes classifier technique are generally correct. However, the fact that the data is noisy can produce incorrect results due to some reasons, such as variance. By reducing input features, Naive Bayes allows important features to be found. Thus, effective and active intrusion detection system can be created [31].

With the Naive Bayes classification method, it was attempted to detect attacks occurring on the network [32]. They achieved very successful results in the study that they conducted in order to detect new types of attacks occurring in the network on the KDD'cup'99 data set. Deshmukh et al. [29] achieved 88.02% success in their study carried out on the KDD-Cup 1999 data [29]. The Naive Bayes method have achieved about 85% success in the studies performed on the KDD-Cup 1999 data [33].

4) Fuzzy Logic

Fuzzy Logic techniques have been used in the field of computer security since the 90s. There is no concept of certainty in this technique. A data is included in more than one class according to the degree of membership. Degree of membership is between 0 and 1. For example, the fact that air or water is warm, cold or mild varies according to people. For this reason, certain value ranges cover these three classes. The steps of Fuzzy Logic technique are given below:

- All data is defined.
- Membership functions are determined.
- The rules to be applied are defined.
- The rules are evaluated and combined.
- The data is classified according to the membership values shown.

In mixed systems, it continuously analyzes incoming data to ensure computer security. Additionally, this technique attempts to detect user signatures or attacks with classical pattern recognition. It detects events that are faulty or abnormal. It consists of two stages. The first stage is rule generation, while the second stage is detection and prevention of attacks. Because Fuzzy Logic addresses low, medium, and high attack types in a way to cover each other, it overcomes many sharp boundary problems and reduces errors that occur as false positives. It ensures system optimization in real-time systems [34].

Using the Fuzzy Logic technique, Tian et al. [35] divided the large data sets into sub-datasets and performed a performance analysis of the data set by following the TCP data for different data sets. In the studies conducted on the KDD-Cup 1999 data, the Fuzzy Logic method has achieved approximately 94% success [30]. Nadiammai et al. [13], on the other hand, achieved 81.54% success using the K-average clustering method on the KDD-Cup 1999 data.

C. Artificial Intelligence Methods

Under this group, Genetic Algorithm, Artificial Neural Networks, and Artificial Immune Techniques are widely used methods.

1) Genetic Algorithm

Genetic algorithms are methods based on biological processes and used in solving search and optimization and also used to be able to make modeling.

Genetic algorithms are used to solve problems that are difficult or impossible to solve by conventional methods. Since genetic algorithms are a random search method, they work over multiple sets of solutions instead of searching a single solution for the optimal solution of the problem. Therefore, results in problem solving are not always best. The reason why genetic algorithm is preferred is that genetic algorithm does not need any information about the nature of the problem. The basic steps of the genetic algorithm are as follows [36]:

- A random population is created.
- By applying genetic processes (selection, crossover), new individuals are created from this population. Selection and crossover processes create new individuals by exchanging genes from individuals in the population.
- Among these individuals, the most appropriate individuals who can solve the problem are selected.
- The population length is the same for all iterations, and for future iteration the highest probability of the function is chosen. Iteration stops when it comes to the optimal threshold.

In distributed systems, the network structure is denoted by G=(N,E) weighted graph. E indicates communication links between nodes, while N indicates nodes. In multiple network structures, the variable T represents cost, while the network structure is denoted by T= (N_T,E_T). P_T (s,u) denotes the path from s source to u destination node. The cost of T is calculated as in Equation 1 [36].

$$C(T_e) = \sum e \varepsilon E_T \, C(e), e \varepsilon E_T \qquad (1)$$

On the other hand, the minimum bandwidth from s source node to u target node is calculated as in Equation 2.

$$B_T = \min(B(e), e\mathcal{E}E_T) \quad (2)$$

In studies conducted on the KDD-Cup 1999 data, the genetic algorithm has achieved approximately 85% success [37].

2) Artificial Neural Networks

Artificial Neural Networks (ANN) are information systems that model the human brain and classify data through learning. It was developed based on the working principle of the human brain. In other words, ANN is an information processing structure developed with a logic similar to biological neural networks and linked to each other by weights.

An ANN consists of input, output, and hidden layers. While the data is imported to the artificial neural network by the input layer, it was exported by the output layer. The layers between the input and the output layer constitutes the hidden layers.

Neurons in feedforward artificial neural networks are connected only forward. The neuron refers to all data connected together. Each layer of the neuron network contains the connection of the next layer, and these connections are not backward. That is, there is a hierarchical structure between neurons, and neurons in one layer only transmit data to the next layer. The forward transition consists of activation flow reaching the output layer and input samples. Activation functions such as Sigmoid and Gaussian function can be used. In the back-transition phase, the actual output on the network is compared to the target output and the error in the output units is calculated [38]. For intrusion detection and prevention systems, this structure is calculated as in Equations 3, 4, and 5 [39].

$$x(t) = f\left(W^A x_c(t) + W^B u\,(t-1)\right) \quad (3)$$

$$x_c(t) = x(t-1) \quad (4)$$

$$y(t) = g\left(W^C X(t)\right) \quad (5)$$

where x(t) is the hidden layer output, y(t) is the output of the output layer, u (t-1) is the input of the network, WA is the weight of the connection between units and the hidden layer, WB is the weight of the connection between the input and output layer, WC is the weight of the connection between the hidden and output layer, f (.) and g (.) are the activation code between the hidden layer and the output layer [39].

The backpropagation network shows how the neuron is trained. ANN is a type of the supervised learning. When the supervised method is used, the network is provided with both sample inputs and expected outputs. The expected outputs for the given networks are compared to the actual outputs. If the expected outputs are used, the error is calculated and the weights of the various layers are adjusted backward from the output layer to the input layer. The network updates its coefficients to obtain the expected output. In ANN, the error in the output layer is calculated as a result of each iteration and this error is transmitted from the output layer to the input layer toward all neurons and the weights are rearranged according to the margin of error. This margin of error is distributed to the neurons preceding the neuron itself in proportion to their weight.

ANN is a method through which security vulnerabilities can be solved in distributed systems. In distributed computations, ANN consists of nodes, connections between nodes, and processing units. The connection between the two units consists of one unit's weights that affect the other unit. These units move from the input nodes to the output nodes by gathering and passing through the threshold value. ANN is implemented in different network securities, medicine, marketing, banking and finance, telecommunication, operations management and other industries [38].

Tong and colleagues used the ANN model in the intrusion detection system. With Elman neural networks, they have attempted to detect both faulty and anomaly attacks. These neural networks have the content of the nodes; the content of each node receives its input from the single hidden layer, and the output for each node is connected to the hidden layer [39]. Mahit et al. [40] achieved about 94% success in their study conducted with ANN on the KDD-Cup 1999 data. In addition, 98.5% success was achieved in the study carried out on the KDD-Cup 1999 data by Aburomman et al. [16].

3) Artificial immune system

We can identify the immune system as a protective mechanism that brings the body in defense by protecting it against diseases. The artificial immune system (AIM), which has emerged by inspiring from the biological definition of the immune system and its working logic, has been observed to be frequently involved in research studies, especially in recent years as an artificial intelligence-based method. One of these areas of research is also the use of AIM in intrusion detection. It can be said that the first source of inspiration for this system, which we call artificial immunity, is computer viruses.

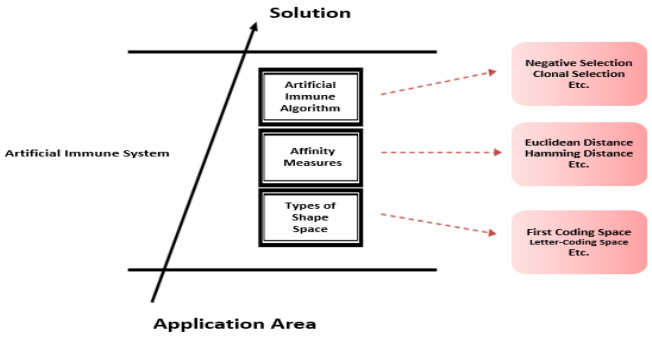

Fig.3. The Layered Structure of AIM [41]

The artificial immune system (AIM) has a layered structure and it is shown in Figure 3 [41]. The similarity between antigen and antibodies used as the basic part of the system is ensured by affinity measurement, such as recognition and representation in the shape space. Euclidean, Hamming, or Manhattan distances are used for the affinity measurement. By mutating the individuals with a low-affinity measurement more, the best antibody that will recognize an antigen is found, and the antibody is replicated by mutation at certain rates (clonal selection algorithm). The reproduced cells are added to the antibody cells. The antigen is presented to the antibody cells. Antibodies that best identify the antigen and are above a certain threshold value are taken and raced among them (similarity rates are calculated with Affinity measurement). As a result of the race, a certain number of individuals with low affinity are taken and added to the memory cell. In this way, the memory cell that will represent the best antibody cells against an antigen is produced. This process is also applied separately for other antigens to be presented to the system [42]. The artificial immune method achieved approximately 99% success in the study of Shem et al. [42] conducted on the KDD-Cup 1999 data. Chen et al. [17] also achieved 92.41% success in the study they carried out on the KDD-Cup 1999 data.

V. EXPERIMENTAL STUDY

In this study, distributed database intrusion detection and prevention methods were compared according to some metrics such as accuracy, speed, performance, and size of the data set used. Accordingly, the advantages and disadvantages of the used methods are shown in Table 2.

TABLE II
ADVANTAGES AND DISADVANTAGES OF THE METHODS USED IN DISTRIBUTED SYSTEMS

Data Mining Methods	**Advantage:** it is very useful for small data and its success is high. It is safe, and its results are close to precise.
	Disadvantage: Its application to the system brings with it a number of computational and time complexities.
Statistical Methods	**Advantage:** It generally does not need prior information to observe Normal movements. It is therefore active in finding new types of attacks.
	Disadvantage: Determining the parameters and metrics required to calculate the success of the system are quite difficult and vary for each set of data.
Artificial Intelligence Methods	**Advantage:** It is flexible and can be applied to almost all data sets.
	Disadvantage: It consumes a high amount of resources for the intrusion detection and prevention systems.

Artificial intelligence techniques and data mining methods provide more successful results than statistical methods. In addition, artificial intelligence methods are more easily implemented for big data. By combining these techniques, it may be possible to achieve safer and faster data transmission.

VI. RESULTS AND EVALUATION

With the development of technology, data access and data communication has become quite easy. Thanks to the distributed systems, quick and easy access to data from anywhere can be ensured. However, the fact that more than one user wants to access the system from different locations at the same time presents a number of problems such as data security, data confidentiality, service continuity, authorization, and secure storage of the data. Problems of the users' access to data, listening to the network, blocking the service, insecure networks, obtaining and storing of the important information by unauthorized people constitute a few of these problems. In order to overcome these problems, various systems developed based on the intrusion detection and prevention techniques applied in the field of data mining, statistical and artificial intelligence are used.

The performance results of the techniques, which are used for intrusion detection and prevention systems in distributed databases, on the KDD-Cup-1999 data are given in Table 3 by compiling from various studies. In the studies conducted, it was observed that the use of noise removal, feature extraction, and selection methods increased classification success, and with these methods, the data at hand is made more useful. Therefore, the results of the studies vary depending on the techniques of normalization, noise removal, feature extraction, and feature selection.

TABLE III
SUCCESFULL OF THE METHODS USED IN DISTRIBUTED SYSTEMS

Methods	Techniques	Success Rate (%)
Data Mining Methods	Support Vector Machines	99.18
	k-Nearest Neighborhood	97.04
	Decision Tree	96.83
	Association Rules	96.9
	K-Mean Clustering	96.9
Statistical Methods	Fuzzy Logic	94.92
	Hidden Markov Models	93.4
	Naive Bayes	88.02
	Learning Vector Quantization	81
Artificial Intelligence Methods	Artificial Immune Techniques	99.74
	Artificial Neural Networks	94
	Genetic Algorithm	85.7

VII. DISCUSSION AND CONCLUSION

Today, computers are widely used in many parts of education, economy, military and business life. A lot of valuable information that is confidential, private and needs to be protected is shared among users over the network. The rapid development of the technology and the sharing of information leads to occurring of attacks against the network and computer security. Emerging of many security problems such as encryption, authorization, damaging the system by unauthorized persons, crashing the system also necessitates

taking the required security measures. In this case, the intrusion detection and prevention systems have emerged to detect users who are abusing the system and to prevent abnormal behaviors.

As a result, artificial intelligence techniques, particularly the artificial immune technique, have given highly successful results in the intrusion detection and prevention systems for distributed databases. The effective use of artificial intelligence techniques in the intrusion detection and prevention is of great importance for future studies. It should not be ruled out that more successful results can be achieved by using hybrid methods, including especially improved artificial intelligence techniques.

REFERENCES

[1] P. Ramasubramanian, A. Kannan, "Multi-Agent based Quickprop Neural Network Short-term Forecasting Framework for Database Intrusion Prediction System", CiteSeerX, 2014.

[2] P.Romasubramanian, A. Kannan, A. "A genetic-algorithm based neural network short-term farecasting framework for database intrusion prediction system", Soft Computing, Vol., 8, pp. 699-714, 2006.

[3] K.Haslum, A. Abraham, "Disp: A framework for distributed intrusion prediction and prevention using hidden markov models and online fuzzy risk assessment", 3rd International Symposium on Information Assurance and Security, pp.183-190, 2007.

[4] H.Deng, Q. Zeng, Q, "SVM-baseed detection system for wireless ad hoca networks", Vehicular Technology Conference, Vol.3, pp. 2147-2151, 2003.

[5] F. Jemili, M. Zaghdoud, "Hybrid Intrusion Detection and Prediction multiAgent System, HIDPAS", (IJCSIS) International Journal of Computer Science and Information Security, Vol.5, 1, pp. 62-71, 2009.

[6] W. Hu, G. Jun, "Online Adaboost-Based Parameterized Methods for Dynamic Distributed network Intrusion Detection", IEEE Transactions on CyberNetics, Vol.44, 3, pp. 66-82, 2014.

[7] A. Abraham, C. Grosan, C. Martiv-Vide, "Evolutionary design of intrusion detection programs", Int. Journal of Network Security, Vol. 4, pp. 328-339, 2007.

[8] Ş.Sağıroğlu, E.N. Yolaçan, U. Yavanoğlu, "Zeki Saldırı Tespit Sistemi Tasarımı ve Gerçekleştirilmesi", Ankara, 2011.

[9] M.Z.Yıldırım, A.Çavuşoğlu, B. Şen, İ. Budak, İ, "Yapay Sinir Ağları ile Ağ Üzerinde Saldırı Tespiti ve Paralel Optimizasyonu", XVI. Akademik Bilişim, Mersin, 2011.

[10] Y.Liao, V.R.Vemuri., "Use of K-Nearest Neighbor classifier for intrusion detection", Elsevier Computers&Security, Vol. 21, 5, pp.439-448, 2002.

[11] M.Jianliang, "The Application on Intrusion Detection based on K-Means Cluster Algorithm", International Forum on Information Technology and Applications, pp. 150-152, 2009.

[12] K.M. Faraoun, A.Boukelif, "Neural Networks learning improvement using the K-Means clustering algorithm to detect network intrusions", International Journal of Computer and Information Engineering, Vol. 1, 10, pp. 3138-3145, 2007.

[13] G.U.Nadiammai, M.Hemalathen, "An evaluation of clustering technique over intrusion detection system", ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 1054-1060, 2002.

[14] K. Law, F.Kwok, "IDS False Alarm Filtering using KNN Classifier, Springer Information Security Applications Lecture Notes in Computer Science", pp.114-121, 2004.

[15] A. Adetunmbi, "Network Intrusion Based on Rough set and k-Nearest Neighbour", International Journal of Computing ICT Research, Vol. 2, 1, 2008.

[16] A.Aburonman, M. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system", Elseiver Applied Soft Computing, Vol.38, pp. 360-372, 2016.

[17] M. Chen, P. Chang, J. Wu, "A population-based incremental learning approach with artificial immune system for network intrusion detection", Elseiver Engineering Applications of Artificial Intelligence, 51, pp. 171-181, 2016.

[18] A. Peddabachigari, A. Abraham, "Intrusion detection systems using decision trees and support vector machines", International Journal of Applied Science and Computations, pp.1-16, 2004.

[19] N. Rachburee, N. Punlumjeak, "Big Data Analytics: Feature Selection and Machine Learning for Intrusion Detection on Microsoft Azure Platform", Journal of Telecommunication Electronic and Computer Engineering, Vol. 9, 1-4, pp. 1-5, 2017.

[20] A. Sung, S. Mukkamala, "Identifying import features for Intrusion Detection using Support Vector Machines and Neural Networks", Proceedings of the 2003 Symposium Applications and the Internet (Saint'03), 2003.

[21] S. Mukkamala, G. Janoski, " Intrusion Detection using Neural Networks and Support Vector Machines", IJCNN'02 Proceedings of the 2002 International Joint Conference on, Vol. 2, pp. 1702-1707, 2002.

[22] A. Tajbakhsh, M. Rahmati, "Intrusion detection using fuzzy assocation rules", Elsevier Applied Soft Computing, Vol. 9, pp. 462-469, 2009.

[23] Y. Hu, B. Panda, " A data mining approach for Database Intrusion Detection", ACM Symposium on Applied Computing, pp. 711-716, 2004.

[24] R. Noum, Z. Al-Sultani, " Learning Vector Quantization (LVQ) and k-Nearest Neighbor for Intrusion Classification", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, 3, pp. 105-109, 2012.

[25] B. Hamman, D. Hoffman, "Learning vector Quantization for (dis-)-similarities", Elsevier Neurocomputing, Vol. 131, pp. 43-51, 2014.

[26] E. Soleiman, A. Fetarat, "Using Learning Vector Quantization (LVQ) in Intrsuion Detection Systems", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol. 1, 10, 2014.

[27] Y. Degang, C. Guo, C, "Learning Vector Quantization Neural Network Method for Network Intrusion Detection", Wuhan University Journal of Natural Sciences, Vol. 12, 1, pp. 147-150, 2007.

[28] L.R. Rabier, "A tutorial on Hidden Markov Models and Selected applications speech recognition", Ready in Speech Recognition, pp. 267-296, 1990.

[29] D. Deshmukh, T. Ghorpade, P. Padiya, " Improving Classification Using Preprocessing and Machine Learning Algorithms on NSL-KDD Dataset", 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 2015.

[30] R. Shanmugavadivu, N. Nagarajan, "Network Intrusion Detection System using Fuzzy Logic", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2, 1, pp. 101-111, 2014.

[31] D.S. Mukherjee, N. Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Elsevier Procedia Technology, Vol. 4, pp. 119-128, 2012.

[32] S. Sharma, "An Improved Network Intrusion Detection Technique based on k-means clustering via Naive Bayes Classification", IEEE-International Conference on Advances In Engineering, Science and Management (ICAESM-2012), pp. 417-422, 2012.

[33] M. Panda, M. Patra, "Network Intrusion Detection using Naive Bayes", IJCSNS International Journal of Computer Science and Network Security, Vol. 7, 12, pp. 258-263, 2007.

[34] A. El-Semany, "A Framework for Hybrid Fuzzy Logic Intrusion Detection Systems", IEEE International Conference on Fuzzy Systems, pp. 325-330, 2005.

[35] J.Tian, "Intrusion detection combining Multiple Decision Trees by Fuzzy Logic", Proceedings of the sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'05), 2005.

[36] S. Janakiraman, V. Vasudevan, "An Intelligent Distributed Intrusion Detection System using Genetic Algorithm", JCIT Journal of Convergence Information Technology, Vol. 4, 1, 2009.

[37] M. Hassan, "Network Intrusion Detection System Using Genetic Algorithm and Fuzzy Logic", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, 7, pp. 435-1445, 2013.

[38] W. Chen, S.H. Hsu, "Application of SVM and ANN for intrusion detection", Elsevier Computers&Operations Research, 32, 2617-2634, 2005.

[39] X. Tong, Z.Wang, "A research using hybrid RBF/Elman neural networks for intrusion detection system secure model", Elsevier Computer Physics Communications, Vol. 180, pp.1795-1801, 2009.

[40] D. Mahit, "Using Artifical Neural Network Classification and Invertion of Intrusion in Classification and Intrusion Detection System,

International Journal of Innovative in Computer and Communication Engineering, Vol. 3, 2, pp. 1102-1108, 2015.
[41] L.Castro, J. Timmis, "Artificial immune systems as a novel soft computing paradigm", Soft computing, Springer, Vol. 7, 8, pp. 526–544, 2003.
[42] J.Shen, J. Wang, "Network Intrusion Detection by Artificial Immune System", IEEE Power and Energy General Meeting, pp.1-8, 2011.
[43] C. Bakir, V. Hakkoymaz, " Veritabanı Güvenliğinde Saldırı Tahmini ve Tespiti için Kullanıcıların Sınıflandırılması", ISCTurkey2015 8.Uluslararası Bilgi Güvenliği ve Kriptoloji Konferansı (VIII. Int'l Conference on Information Security and Cryptology), pp. 28-33, 2015.

BIOGRAPHIES

MEHMET GUCLU received the B.S. degrees in computer engineering from the University of Yildiz Technical University, in 2009 and the M.S. degree in computer engineering from Yildiz Technical University, İstanbul, in 2013. He started his PhD in 2013 at Yildiz Technical University and still continues. His research interests include information security, data mining and machine learning.

CIGDEM BAKIR received the B.S. degrees in computer engineering from the University of Sakarya, in 2010 and the M.S. degree in computer engineering from Yildiz Technical University, İstanbul, in 2014. She started his PhD in 2014 at Yildiz Technical University and still continues. Since 2012, she was a Research Assistant with the Yildiz Technical University. She works a Research Assistant at Igdir University. Her research interests include recommendation systems, information security, machine learning and data mining.

VELI HAKKOYMAZ received B.S. degrees in computer engineering from Hacettepe University, in 1987 and M.S.degree in Computer Science from University of Pittsburgh (PA), In 1992, Ph.D.degree in CWRU (OH) in 1997. In 2011, he received the title of Associate Professor. His research interests include database management systems, computer architecture, operating systems and distributed systems.

BANU DIRI received the B.S. degrees in computer engineering from the University of Yildiz Technical University, in 1987 and the M.S. degree in computer engineering from Yildiz Technical University, in 1990, Ph.D.degree in Yildiz Technical University in 1999. She works as a professor at the Yildiz Technical University. Her research interests include data mining, natural language processing, machine learning and artificial intelligence.