© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Fuzzy Generalized Hebbian Algorithm for Large-Scale Intrusion Detection System

Ahmed Hussein Ali[1*], Mohammad Aljanabi[2], Munef Abdullah Ahmed[3]

[1,2]AL Salam University College, Computer Science Dep. Baghdad, IRAQ

[1,2]Computer Science Dep, Education College, Al-Iraqias University, Baghdad, IRAQ

[3]Northern Technical University-Al Hawija Technical Institute-IRAQ

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2020.12.01.008
Received 29 October 2019; Accepted 22 January 2019; Available online 31 January 2020

Abstract: The huge number of irrelevant and redundant data used in building intrusion detection systems (IDS) is one of the common issues in network intrusion detection systems. This paper proposed the use of Fuzzy Generalized Hebbian Algorithm as a novel data reduction method to overcome this problem of data redundancy in IDS. Two methods for dimensionality reduction (GHA and Fuzzy GHA) were used and compared in this study. This allowed retaining the most relevant traffic data information from the network. Furthermore, the K Nearest Neighbor algorithm was applied for the classification of the test connections into 2 categories (attack or normal). The investigations were carried out on the KDDCUP '99 dataset and the results showed the Fuzzy GHA method to perform better than GHA in the detection of both U2R and DoS attacks.

Keywords: Generalized Hebbian Algorithm, Dimension reduction, Fuzzy GHA, Network security, Intrusion detection system

1. INTRODUCTION

Several computer techniques and mechanisms exist in recent times for improving the security and robustness of computer networks. The intrusion detection system (IDS) is one of such techniques which can be deployed for the detection of network anomalies [1, 2]. The IDS can detect threats to networks' security policy violation. The IDS is generally classified into anomaly-based (AB) and misuse-based (MB) categories [3, 4]. For the MB approach, network abnormalities are detected by comparing the observed network pattern to already established attacks patterns [5, 6]. Hence, there is a need to have an established database of attack signatures [7, 8]. The MB approach ensures a good detection of well-known network attacks [4, 9]. One major problem of the MB approach is that new or unfamiliar attacks may not be detected. For the AB approach, it attempts to determine "normal" network behaviors by generating an alarm when the difference between the pattern of an observed network behavior and that of a normal behavior has exceeded a pre-defined limit.

The purpose of this work is to improve the efficiency of the AB IDS through a reduction of the highly dimensional network traffic data using some techniques before deploying any AB algorithm. A common method of addressing the high dimensionality problem is the identification of the most relevant features which are related to all the connection records without affecting the classification quality. The Generalized Hebbian Algorithm (GHA) has proven to be a common approach with proven efficiency in many application as it allows the definition of the "eigenvectors" of the covariance matrix of the connection records distribution [3][10]. The variation between all the connection records can be calculated using these eigenvectors as features. The

*Corresponding author: msc.ahmed.h.ali@gmail.com definition of each connection is based on the eigenvectors that correspond to the highest eigenvalues; this depicts the most variance within a given set of connection records.

The GHA, just like any other statistical (multivariate) tool, is unfortunately sensitive to missing data, outliers, and the poor linear relationship between variables. Consequently, the GHA is influenced by data transformations.

The challenges of the GHA can be addressed by using one of the most exciting methods called Fuzzy Generalized Hebbian Algorithm (FGHA) whose main aim is input data fuzzification to minimize the influence of outliers. It uses Fuzzy C-Means algorithm to achieve this aim before reformulating GHA into FGHA. A novel feature extraction method from large-scale data, with the aim of finding a small set of features that can represent the most data variance, is proposed in this study. The rest of this paper is organized in the following manner: Section 2 is a review of related works while section 3 presents an overview of IDS. Section 4 presents a description of GHA and FGHA while section 5 presents the proposed approach in this study. Section 6 presents the experimental methods and result discussion while section 7 presents the conclusion of the study.

2. LITERATURE SURVEY

Several methods of handling noisy, large scale, and highly dimensional data have been proposed. Stanislaw et al [4] suggested a PCA neural network-based noise reduction method for noise and corrupt data elimination. The authors enumerated the main contribution of this work to include the development of a novel PCA-neural network-based coding/decoding method for noise removal, its suitability for image filtering without the need for more statistical knowledge of distorting random noise, as well as the implementation of the method in random noise reduction from 1D signals or 2D images. Furthermore, Uday et al [5] proposed an SQL implementation of a stored procedure for multisystem information transformation into Multiset Decision Tables. They proved the data reduction effectiveness of the MDT through performance evaluation on a large IDS data set. Abhishek et al presented a novel hyperspectral data dimension reduction method based on PCA. Zhang et al [6] developed a synthetic data dimension reduction approach by deploying analogy reasoning to define the similarity distance algorithm between two vectors. The benefit of the proposed reduction method was analyzed in a 3-D space. Finally, the distances between the sample were deduced from the sample sets. Xu et al [7] relied on the Bayesian fusion approach to propose a new IDS model which is comprised of several parts, including the feature level fusion, pixel level fusion, decision-making level fusion, etc. Bahrololum et al [8] applied Decision Trees (DT), Flexible Neural Tree (FNT), and Particle Swarm Optimization (PSO) to the KDD99 dataset for feature reduction purpose. Based on the comparison of the performance of the 3 methods on DARPA KDD99 dataset, DT achieved a better detection rate, cost per example, and false positive compared to FNT and PSO. Hence, DT performed better in almost all the attacks. Xiang et al [9] suggested a novel linear correlation feature reduction framework. This framework is beneficial in cases of marginally unrelated features which are jointly related to the response. A new approach was introduced for the removal of redundant features and it proved effective in reducing false selection rate during feature selection. Therese et al [10] evaluated the use of different feature selection (FS) methods on several datasets that are freely available. The evaluation focused on feature reduction and selection algorithms. Three FS algorithms which consist of a test method and an attribute evaluator were used. Deepa et al [11] comparatively studied 3 FS methods which are PCA, Folded PCA, and segmented PCA. The 3 techniques were deployed on hyperspectral images and observed for parameters. Nitika et al [12] proposed a novel PCA and feature ranking-based dimensionality reduction method. The dimensionality reduction performance of the proposed approach was evaluated on Breast Cancer dataset and the outcome showed the proposed method to effectively reduce the dimensionality of the chosen database without compromising the computational cost and classification accuracy. Min and Chan [13] proposed an integrated approach which is a combination of MI-based unsupervised feature transformation (UFT). The method employed PCA for the reduction of the dimensionality of the hybrid data. The UFT can also deploy Shannon’s Entropy and MI to provide proper numerical substitutions that can maintain the information harbored in the original symbolic features. Wang and De-Sheng [14] developed a PSO-based IDS model. The outcome of the experiments showed the model that the improved quantum particle swarm algorithm can improve particle convergence, achieve the minimal reduction, and reduce the chances of particles being trapped in the local minima. Rosanna et al [15] used the metric between distributions, the l2 Wasserstein distance to proposed a PCA-based method for distributional-valued data. Yining and Joe [16] proposed a dynamic inner PCA framework for the modeling of dynamic data through the maximization of the covariance between a component and the prediction based on its previous values. In this method, a dynamic latent variable model is first extracted to portray the most auto-covarying dynamics in a given dataset. The self-predictable variation of the data is contained in the components of the captured dynamic while after the extraction of the dynamic components, the residuals are least predictable, essentially uncorrelated, and can be handled using static PCA. Yao and Zhang [17] developed a novel feature reduction-based fuzzy logic in which each training data assigns a confidence weight to the training set as fuzzy points in the sample space. This paper applied an analysis method which is based on the PC analysis of the objective weight. The method is used to evaluate the artist's creation value. Although numerous data reduction frameworks exist, most of the existing frameworks are sensitive to missing data, outliers, and the poor linear correlation between poorly distributed variables despite being used for process monitoring.

3. INTRUSION DETECTION SYSTEM

In a computer network [18][11], intrusion detection is an important method of detecting different forms of network attack. It is a process of monitoring the pattern of action within a network. Intrusion detection techniques ensure the security of a network through monitoring, detecting, and responding to attacks. The major concern of an IDS is to identify both internal and external network threats [19][12]. In a common way, an IDS can be said to consist of hardware components which require compatible software to be run [20][13]. The IDS work as a network security guard. In the field of intrusion detection, there are 2 assumptions, i) computer systems monitor both user and program activities; ii) the behavior of intrusion and traditional activities can completely differ.

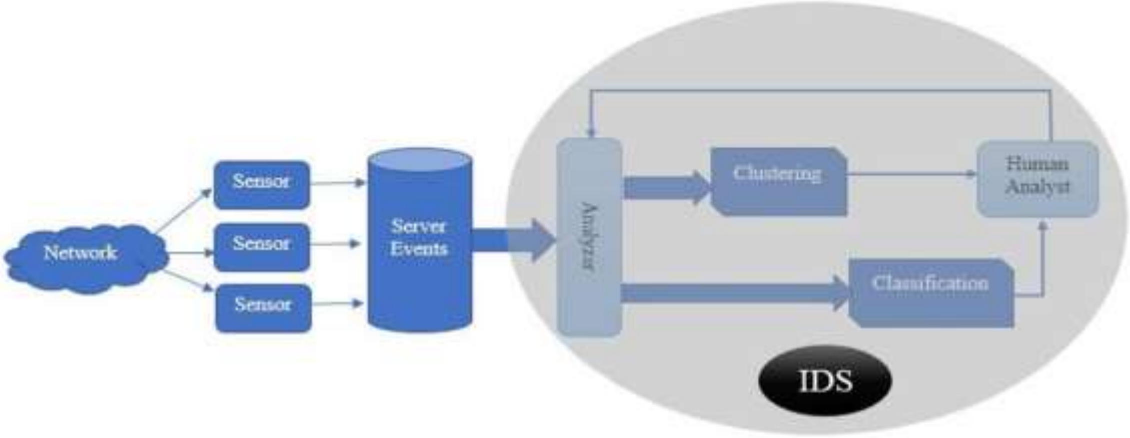

Fig. 1 - The framework of an intrusion detection system

A properly configured IDS should have the following features:

- Must be timeliness: A good IDS must detect an abnormal activity within a given time.
- Must have a high detection probability: Most of the abnormal network activities must be detected.
- Must have a low rate of false alarm.
- Must be specific: It must provide a detailed information of any detected attack in order to get a better response.
- Must be scalable to both large and small networks.
- Must have a low pre-information of an attack: The pre-information and strategies of an attack must not be known.

Figure 1 depicts the framework of an IDS.

4. GHA AND FUZZY GHA

The theoretical concepts GHA and FGHA are presented in this section.

4.1 Generalized Hebbian Algorithm (GHA)

The GHA is a linear feedforward NN framework for unsupervised learning[14, 15]. It is primarily applied in PCA. From the computational perspective, the GHA[16, 17] is beneficial because it can solve eigenvalue problems using iterative methods which requires no direct covariance matrix computation. This is more significant when there are many attributes in a given set[18, 19]. Oja and Karhunen[16] demonstrated an incremental solution to finding the first eigenvector from data arriving in the form of serial data items presented as vectors[7, 20]. Later, Sanger[21] generalized this to the use of GHA to find the first N eigenvectors. The algorithm converges on the exact eigen data decomposition with a probability of 1. The simple Hebbian learning rule is the basic concept of these algorithms:

$$U_n(t+1) = U_n(t) + \lambda * (U_n^T * A_j) * A_j$$

where U_n = n^{th} column of U i.e., the n'th eigenvector (refer to Eq. 2), λ = learning rate, A_j = j'th column of the training matrix A, t = timestamp.

To extend to multiple eigenvectors, the only modification required is that each U_n has to shadow any lowerranked $U_m(m > n)$ through the removal of its projection from the input Aj. This will guarantee an orthogonal and ordered ranking of the resulting eigenvectors. Consider a dataset M with connection vectors v1, v2, v3….,$v_{M;}$ let each of the connection vector be represented by N features. The PCs are calculated following these steps:

Step 1: Determine the average μ of the dataset.

$$\mu = \frac{1}{M}\sum_{i=1}^{M} v_i \qquad (1)$$

Step 2: Deviation from the average is defined as:

$$\theta_i = v_i - \mu \qquad (2)$$

Step 3: The sample covariance matrix of the dataset is defined as:

$$C_{n\times n} = \frac{1}{M}\sum_{i=1}^{M} \theta_i \theta_i^T = \frac{1}{M}AA^T \qquad (3)$$

Where $A = [\theta_1, \theta_2, \theta_3, \ldots, \theta_n]$.

Step 4: Let Uk represent the k^{th} eigenvector of C, λ_k represent the associated eigenvalue; also let $Un\times d$ = U_1U_2
…U_d represent the matrix of these eigenvectors. Then,

$$CU_k = \lambda_k U_k \qquad (4)$$

Step 5: The eigenvalues are ordered in a descending order before selecting the eigenvectors or PCi with the largest eigenvalues. The number of selected PCs is dependent on the inertia ratio expressed as:

$$\tau = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \qquad (5)$$

This ratio is an expression of the level of information withheld from the rough input data by the associated eigenvalues.

Step 6: Let t represent a new sample column vector; the projection of t onto the new subspace traversed by these PCi is based on the rule:

$$y_i = U_i^T t \qquad (6)$$

4.2 Fuzzy GHA

The major concept of this method is the fuzzification of the input data in order to obtain a fuzzy membership for each data and reformulate GHA into FGHA. Assume M to be a set of connection with vectors v1, v2, v3,….,$v_{M,}$ and each vector is represented by N features.

1) The first process will be the fuzzification of the dataset by applying FCM algorithm on the input dataset to obtain centroids (V) and membership matrix R. 2) Next is fuzzy covariance matrix calculation:

$$C_{fpca} = \frac{1}{M} \sum_{i=1}^{M} V_i V_i^T \qquad (7)$$

3) Determine the number of eigenvectors, then, calculate U and λ using Eq. 4. 4) Project the data to U using Eq. 6.

5. PROPOSED APPROACH OF OUR SYSTEM

This work aims to develop an efficient IDS with improved network performances. The proposed approach is composed of the following steps:

Step 1: Dataset

KDDcup99 dataset (used in this study) a commonly used dataset in ID studies as it gives a predictive model the opportunity to differentiate normal network activities from attacks. Its training dataset comprised of approximately 5,000,000 connection records; 10% of this training dataset consists of 494,021 connection records, including 97,278 normal connection records (19.69%). There are 41 attributes in each connection record and each of these attributes presents different features of the associated connection. The connection value is designated as either one type of attack or as normal. Each type of attack belongs to one attack category such as denial-of-service (DOS), R2L, U2R, or probing.

There are 311,029 connections in the KDDCUP99 test dataset; it also includes some specific forms of attack that does not exist in the training dataset. It contains 24 types of training attacks with additional 14 types in the test data. Only 10% of this dataset was used in this work.

Step 2: Data preprocessing

Data preprocessing step mainly aims at achieving a standard attribute format prior to the application of any form of dimensionality reduction. As such, the discrete attributes values of the dataset were first converted into continuous values following the method earlier used by [26]. Assume that a discrete attribute i has m possible values; we correspond m coordinates for each discrete attribute and associated one coordinate for each possible attribute value. The corresponding coordinate to the attribute value then has a value of 1 while the rest of the coordinates has a value of 0. For instance, consider a type of attribute with any of the following discrete attributes: tcp, udp or icmp. Based on the idea earlier described, this attribute will have 3 coordinates. Consequently, assume a connection record with a tcp (resp. udp or icmp) protocol, it will have the following coordinates: (1,0,0) (resp. (0,1,0) or (0,0,1)). Each connection in the datasets will after the conversion have 128 coordinates (comprised of 3 different values for protocol_type, 11 for flag attribute, 70 for service, and 0 or 1 for the remaining 6 attributes) instead of only 41 attributes.

Step 3: Dimensionality features reduction

GHA and FGHA were used in this stage to minimize the high data dimensionality (both for testing and training datasets) while maintaining the maximum differences in the original dataset.

Step 4: Classification

The KNN classifier was deployed at this stage to classify in order to check their status as either attack or normal.

6. EXPERIMENTS AND RESULTS

All the experiments and the results achieved are presented in this section. A total of 1900 normal connection, 900 Probing, 900 DOS, 52 U2R, and 900 R2L randomly selected from the utilized 10% of the KDDCUPP99 training dataset. For the testing data, 900 normal connections, 900 Probing, 900 DOS, 900 R2L, and 52 U2R were selected in a random manner from the utilized dataset. The successfully classified intrusions were denoted as true positives (TP), correctly predicted normal connections were denoted as true negatives (TN), wrongly classified normal connections were denoted as false positives (FP), and wrongly classified intrusions were denoted as false negative (FN). The performance of the developed model was evaluated using four performance measures which are the detection rate (DR, also known as recall), false positive rate (FPR), F-measure, and precision. The average of these performance measures was calculated using 10-fold cross-validation to results validity.

$$DR = \frac{TP}{TP + FN} \times 100 \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \times 100 \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{10}$$

$$F - measure = \frac{2\times TP}{2\times TP+FP+FN} \times 100 \tag{11}$$

A well-built IDS must possess a high detection rate, low FPR, high precision, and high F-measure. Initially, two experiments were conducted in this study to determine the optimum parameters for achieving a maximum DR, the precision value, and F-measure value. During the first experiment, the number of PCs was fixed at 2 while the number of nearest neighbors was widely diversified. As earlier stated in Figure 1, k = 2 nearest neighbors produced the best results with the maximum DR, F-measure, Precision, and the least FPR. This experiment mainly aims to find the best number of PCs that will guarantee an enhanced DR. In the second experiment, the number of nearest neighbors was fixed at 2 while the number of generalized PCs was varied to establish the best number of k neighbors to achieve the best DR. As depicted in Figure 2, both the first and second PCs gave the best results. Based on the 2 experiences earlier stated, the number of nearest neighbors' and PCs were fixed at their optimal values for the computation and establishment of the DR for each type of attack.

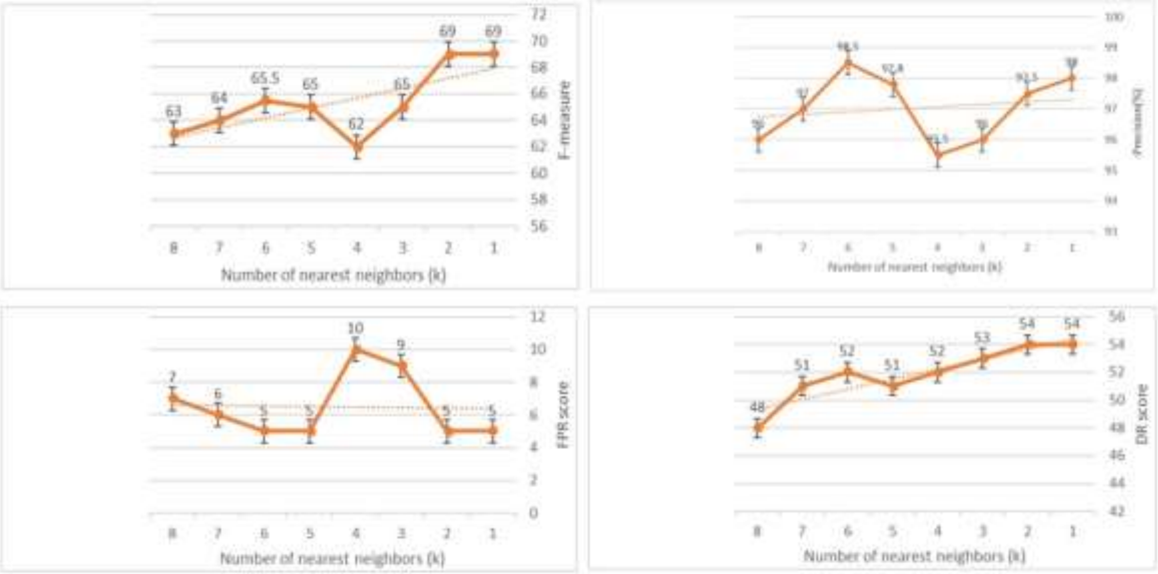

Fig. 2 - F-measure, Precision, FPR, DR, vs. k

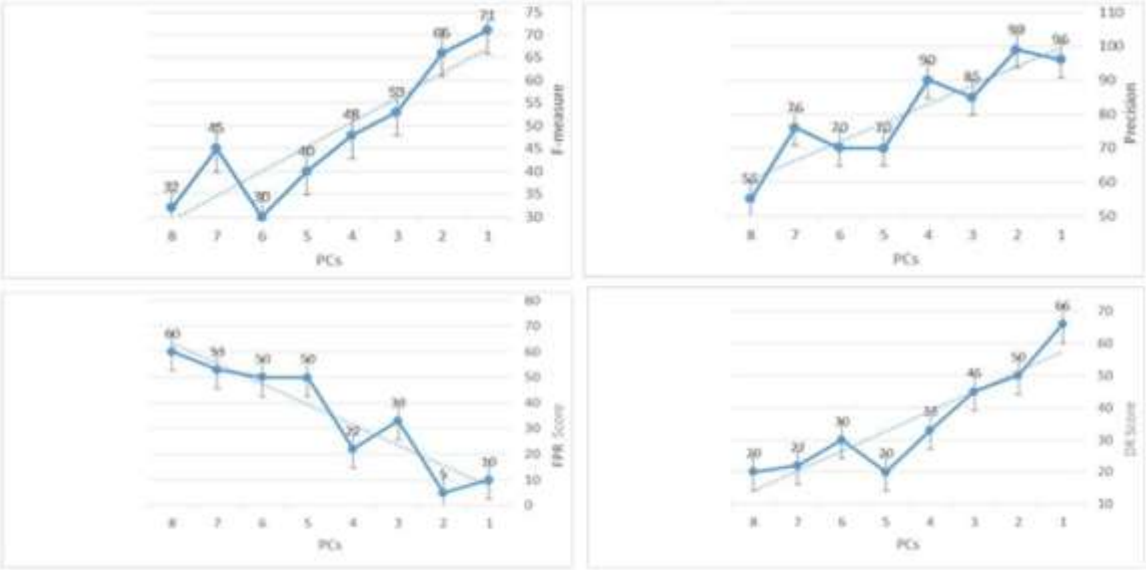

Fig. 3 - F-measure, Precision, FPR, DR, vs. number of PCs

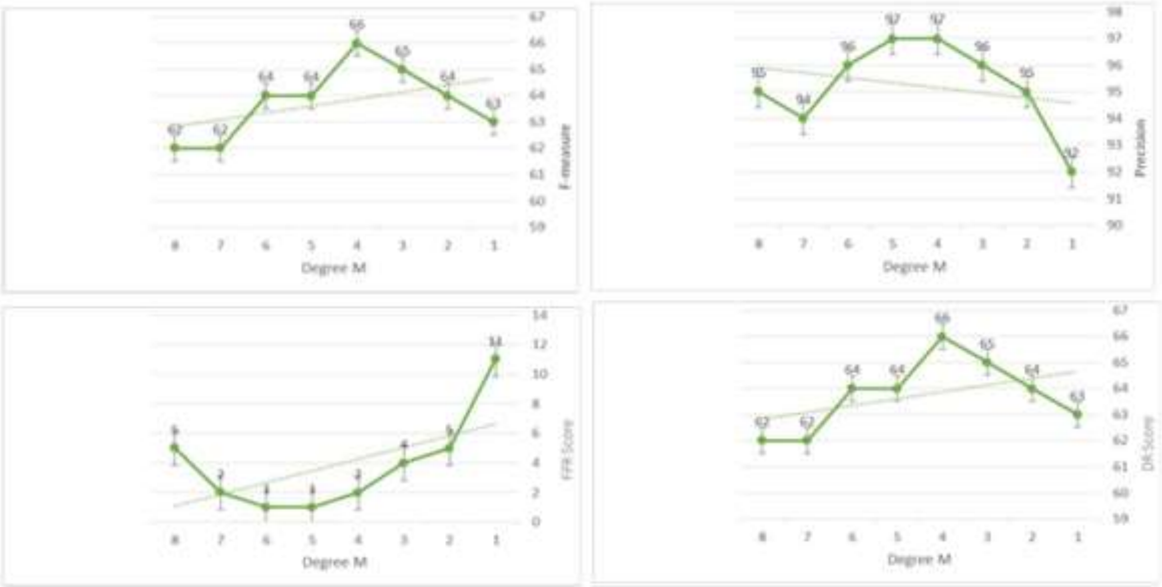

Fig. 4 - F-measure, Precision, FPR, DR, vs. Degree M

The third experiment involved the evaluation of the efficiency of FGHA in intrusion detection. As such, the number of generalized PCs and k was fixed at 2 to determine the degree of membership M that can give the optimal results. As clearly shown in Figure 3, M = 9 produced the optimal results.

The subsequent experiments we focus on comparing the two methods (GHA and FGHA). Figures 4 and 5 showed that FGHA performed better than GHA at the first and second PCs in attack detection. However, GHA achieved fewer FPR rate compared to FGHA.

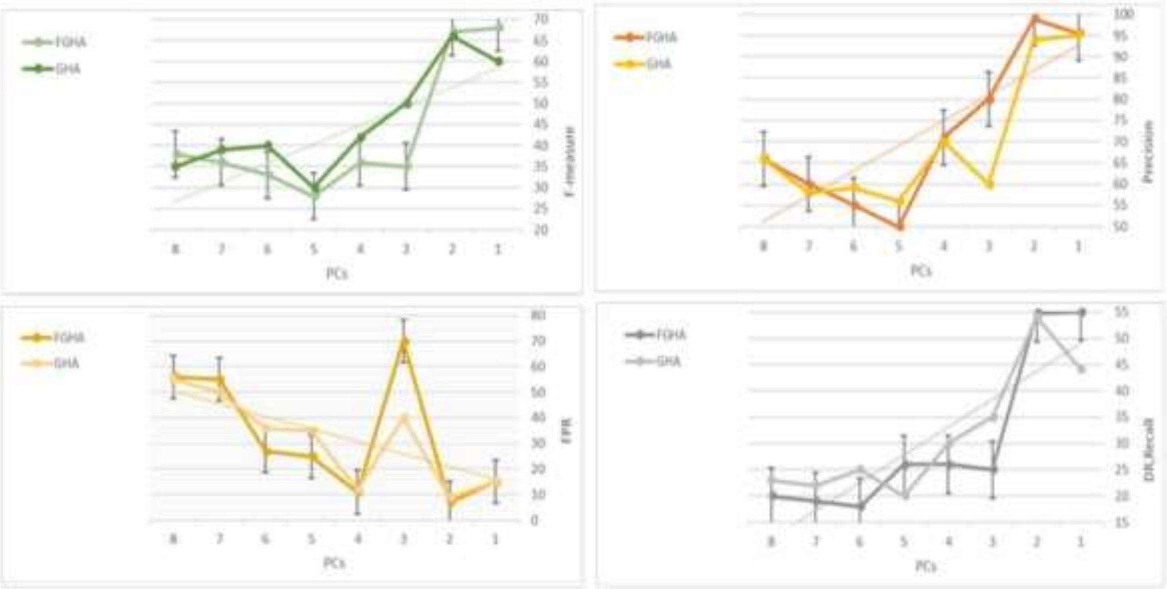

Fig. 5 - F-measure, Precision, FPR, DR, vs. PCs of GHA and FGHA

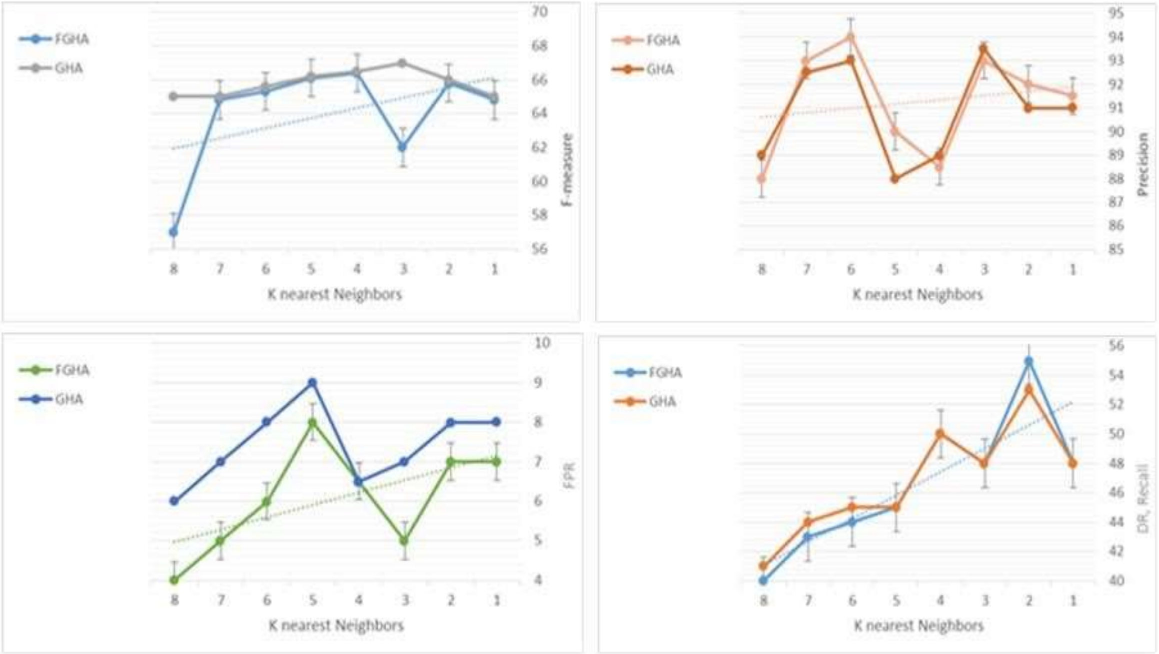

Fig. 6 - F-measure, Precision, FPR, DR, vs. K of GHA and FGHA

A comparison of the DRs of each type of attacks done for both GHA and FGHA to get more realistic results (Table 1). The global DRs of FGHA for U2R and DOS are better compared to those of GHA for the same attack types.

Table 1 - Attacks detection rate of GHA and FGHA

	Preprocessing algorithm	Probing	R2L	U2R	DOS
Detection Rate	GHA	90.363	5.2	7.221	71.133
	FGHA	96.038	5.016	15.239	74.653

Finally, it was determined that FGHA performed better on KDDCUP99 dataset compared to GHA when the number of PCs is varied from 1 to 8. However, FGHA achieved a lower false alarm rate compared to GHA despite its better DR

7. CONCLUSION

The approach presented in this paper mainly aims at reducing the high volume of input data features associated with ID connections records while retaining the important information. This aim was achieved using both GHA and the suggested FGHA. To enhance the DR and reduce false alarm rate, there is a need to build a strong IDS. From the evaluations performed in this work, FGHA performed better in detecting U2R and DoS attacks. The future studies will focus on the hybridization of FGHA with other machine learning methods, as well as enhancing the performance of IDS on the latest datasets using some evolutionary optimization methods.

ACKNOWLEDGMENTS

The authors are grateful to Al-Iraqia University and Al-Salam University College for their moral support. Special thanks to the anonymous reviewers for their valuable suggestions and constructive comments.

REFERENCES

[1] M. A. Ahmed, R. A. Hasan, A. H. Ali, and M. A. Mohammed, "The classification of the modern arabic poetry using machine learning," *Telkomnika,* vol. 17, no. 5, 2019.

[2] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, 2018, pp. 140-145: IEEE.

[3] M. A. Mohammed *et al.*, "A Focal load balancer based algorithm for task assignment in cloud environment," in *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2018, pp. 1-4: IEEE.

[4] Z. H. Salih, G. T. Hasan, M. A. Mohammed, M. A. S. Klib, A. H. Ali, and R. A. Ibrahim, "Study the Effect of Integrating the Solar Energy Source on Stability of Electrical Distribution System," in *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, 2019, pp. 443-447: IEEE.

[5] R. A. Hasan, I. Alhayali, A. Royida, N. D. Zaki, and A. H. Ali, "An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark," *Telkomnika*, vol. 17, no. 6, 2019.

[6] A. H. Ali, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering,* vol. 11, no. 6, pp. 138-150, 2019.

[7] S. A.-b. Salman, A.-H. A. Salih, A. H. Ali, M. K. Khaleel, and M. A. Mohammed, "A New Model for Iris Classification Based on Naïve Bayes Grid Parameters Optimization."

[8] Z. F. Hussain *et al.*, "A new model for iris data set classification based on linear support vector machine parameter's optimization," *International Journal of Electrical & Computer Engineering (2088-8708),* vol. 10, 2020.

[9] A. H. Ali and M. Z. Abdullah, "A novel approach for big data classification based on hybrid parallel dimensionality reduction using spark cluster," *Computer Science*, vol. 20, no. 4, 2019.

[10] S. A. Abed, H. K. Sulaiman, and Z. A. H. Hassan, "Reliability Allocation and Optimization for (ROSS) of a Spacecraft by using Genetic Algorithm," in *Journal of Physics: Conference Series*, 2019, vol. 1294, no. 3, p. 032034: IOP Publishing.

[11] A.-H. A. Salih, A. H. Ali, and N. Y. Hashim, "Jaya: An Evolutionary Optimization Technique for Obtaining the Optimal Dthr Value of Evolving Clustering Method (ECM)."

[12] M. A. Mohammed and N. ŢĂPUŞ, "A novel approach of reducing energy consumption by utilizing enthalpy in mobile cloud computing," *Studies in Informatics and Control,* vol. 26, no. 4, pp. 425434, 2017.

[13] R. A. Hasan, M. A. Mohammed, Z. H. Salih, M. A. B. Ameedeen, N. Ţăpuş, and M. N. Mohammed, "HSO: A Hybrid Swarm Optimization Algorithm for Reducing Energy Consumption in the Cloudlets," *Telkomnika,* vol. 16, no. 5, pp. 2144-2154, 2018.

[14] S.-J. Lin, Y.-T. Hung, and W.-J. Hwang, "Efficient hardware architecture based on generalized Hebbian algorithm for texture classification," *Neurocomputing,* vol. 74, no. 17, pp. 3248-3256, 2011.

[15] R. A. Hasan and M. N. Mohammed, "A krill herd behaviour inspired load balancing of tasks in cloud computing," *Studies in Informatics and Control,* vol. 26, no. 4, pp. 413-424, 2017.

[16] K. Samiee, A. Iosifidis, and M. Gabbouj, "On the comparison of random and Hebbian weights for the training of single-hidden layer feedforward neural networks," *Expert Systems with Applications,* vol. 83, pp. 177-186, 2017.

[17] R. A. Hasan, M. A. Mohammed, N. Ţăpuş, and O. A. Hammood, "A comprehensive study: Ant Colony Optimization (ACO) for facility layout problem," in *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 2017, pp. 1-8: IEEE.

[18] M. Z. A. Ahmed Hussein Ali, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering*, 2018.

[19] S. A. Dheyab, M. N. Abdullah, and B. F. Abed, "A novel approach for big data processing using message passing interface based on memory mapping," *Journal of Big Data,* vol. 6, no. 1, pp. 1-17, 2019.

[20] M. A. Mohammed and R. A. Hasan, "Particle swarm optimization for facility layout problems FLP—A comprehensive study," in *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP),* 2017, pp. 93-99: IEEE.

[21] M. Rizk and E. Koosha, "A Comparison of principal component analysis and generalized hebbian algorithm for image compression and face recognition," in *Computer Engineering and Systems, The 2006 International Conference on*, 2006, pp. 214-219: IEEE.

[22] A. H. Ali, "An Efficient Model for Data Classification Based on SVM Grid Parameter Optimization and PSO Feature Weight Selection," International Journal of Integrated Engineering, vol. 11, no. 8, pp. 135-142, 2019.