

Construct an efficient distributed denial of service attack detection system based on data mining techniques

Dhurgham Kareem Gharkan^{1,2}, Amer A. Abdulrahman¹

¹Department of Computer, Informatics Institute for Post Graduate Studies, Iraqi Commission for Computers and Informatics, Baghdad, Iraq

²Department of Computer Systems, Technical Institute-Suwaira, Middle Technical University, Baghdad, Iraq

Article Info

Article history:

Received Jun 6, 2022

Revised Sep 24, 2022

Accepted Oct 11, 2022

Keywords:

CICDDoS2019 dataset

Data mining

DDoS attack

Distributed denial of service

Intrusion detection system

ABSTRACT

Distributed denial-of-service (DDoS) attack is bluster to network security that purpose at exhausted the networks with malicious traffic. Although several techniques have been designed for DDoS attack detection, intrusion detection system (IDS) It has a great role in protecting the network system and has the ability to collect and analyze data from various network sources to discover any unauthorized access. The goal of IDS is to detect malicious traffic and defend the system against any fraudulent activity or illegal traffic. Therefore, IDS monitors outgoing and incoming network traffic. This paper contains a based intrusion detection system for DDoS attack, and has the ability to detect the attack intelligently, dynamically and periodically by evaluating the set of attackers of the current node with its neighbors. We use dataset named CICDDoS2019 that contains on binary classes benign and DDoS. Performance has evaluated by applying data mining algorithms as well as applying the best features to discover potential attack classes.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Dhurgham Kareem Ghurkan

Department of Computer, Informatics Institute for Post Graduate Studies

Iraqi Commission for Computers and Informatics

Baghdad, Iraq

Email: ms202020630@iips.icci.edu.iq

1. INTRODUCTION

Denial-of-service (DoS) it is one of the types of malicious electronic attacks that make a network device like server, security resources, web element like CCTV, or dedicated secure network like web pages or any website temporarily or indefinitely unavailable to its legitimate users therefore, it prevents any connection, whether in the internet or the host [1]. Also, a sophisticated type of attack appeared to us, distributed denial-of-service (DDoS). It is considered one of the most dangerous electronic attacks. It has the ability to follow intrusive behavior on the network and poses a serious threat to network infrastructure and various network services. Moreover, it is very difficult to detect and track the real attackers. A DDoS attack targets the network by exhausting its resources and thus leads to denial of service and blocking of legitimate users. And it is increasing rapidly over the past few years, the duration of the attack has become shorter with the large growth of data and its widespread spread [2].

We'll go through some of the most recent and widely utilized strategies for detecting DDoS attacks in software defined network (SDN) systems. The support vector machine (SVM) detection algorithm was utilized by Ye *et al.* [3] to identify DDoS attacks in the SDN network. For the learning phase, the authors used six package features that can be obtained from the SDN controller. The dataset samples were obtained by modelling the SDN network with 5 virtual hosts using amount of dividends and flooding controller. Even during modeling stage, 3 independent DDoS scenarios are generated, including user datagram protocol (UDP),

transmission control protocol (TCP), synchronize (SYN), and internet control message protocol (ICMP) flood traffic. Oo *et al.* [4], Rahman *et al.* [5] used four distinct machine learning approaches to detect DDoS attacks in the context of SDN. To build the training and testing dataset, the researchers simulated both normal and malicious traffic. The hping3 software is used to produce two DDoS examples (TCP and ICMP floods). The findings of the trial revealed that the J48 is more accurate than the other approaches tested. To detect flow-table overflow attacks inside the SDN data plane, applied three distinct machine learning algorithms: SVM, Naive Bayes (NB), and neural network (NN). To create training data, the authors used the open flow protocol to extract tuple features from open flow switches. The Scapy utility is used to generate three types of flood traffic: TCP, UDP, and ICMP. Five characteristics are used in machine learning approaches, and the findings demonstrate that the SVM has a lower accuracy rate than the other two classifiers. Two detection algorithms for DDoS attacks on SDN networks were introduced [6]. The signature-based snort detection method was employed to collect network traffic in the first step. SVM and deep neural network (DNN) algorithms are used for attack classification in the final step. Trained the two detection modules using the KDDCUP'99 dataset, which had a total of 41 characteristics. The results of the experiment showed that the DNN outperforms the SVM, with accuracy rates of 92:30 and 74:30 percent, correspondingly.

Mohammed *et al.* [7] suggested a new architecture for DDoS attack detection on SDN. The NSL-KDD dataset was utilized to train the NB classifier, which included 25 characteristics. Combine three different selection algorithms (Genetic, Ranker, and Greedy) to choose the dataset's combined features. Precision, recall, and F1 score had average values of 0:81, 0:77, and 0:77, correspondingly. Most detection methods in the literature that simulated the SDN network to produce the DDoS attacks dataset only consider a small number of malicious activities, and only for IP or TCP protocols, ignoring any application layer DDoS attacks. The great resemblance of attacks and benign actions is one of the issues in detecting application layer DDoS attacks. As a result, there are few criteria available to define such attacks, and many detection algorithms are unable to detect them [8]. In furthermore, tools such as Scapy and Hping3 are used to produce the simulated traffic. As a result, the generated dataset is small and does not contain all of the traffic required to produce reliable results. Furthermore, existing strategies for training anomaly detection systems utilizing public datasets have a number of flaws. Most datasets, for example, seem to be out of date and do not include new types of attack flow. Furthermore, they offer only a few forms of attacks to meet all current internet trends. The evaluation of detection algorithms and approaches systems is greatly influenced by a large and valid dataset. To test our proposed model, we used the most recent publicly available dataset, CICDDoS2019 [9], which contains a wide range of DDoS attacks and fills in the gaps in existing datasets.

The purpose of our work is first to describe a new IDS dataset CICIDS2019, which contains 1048560 samples. Secondly, we analyzed the dataset to determine the best amount of feature set for attack detection and also implemented them. Some common data mining algorithms for evaluating this data set.

In this section we describe a dataset containing 88 features generated by the University of New Brunswick. The dataset is available to all researchers on the Canadian Institute of Cyber Security Intrusion Detection System (CICIDS 2019) website [10], [11]. Figure 1 shows the general structure of the DDOS attack collected in the current dataset.

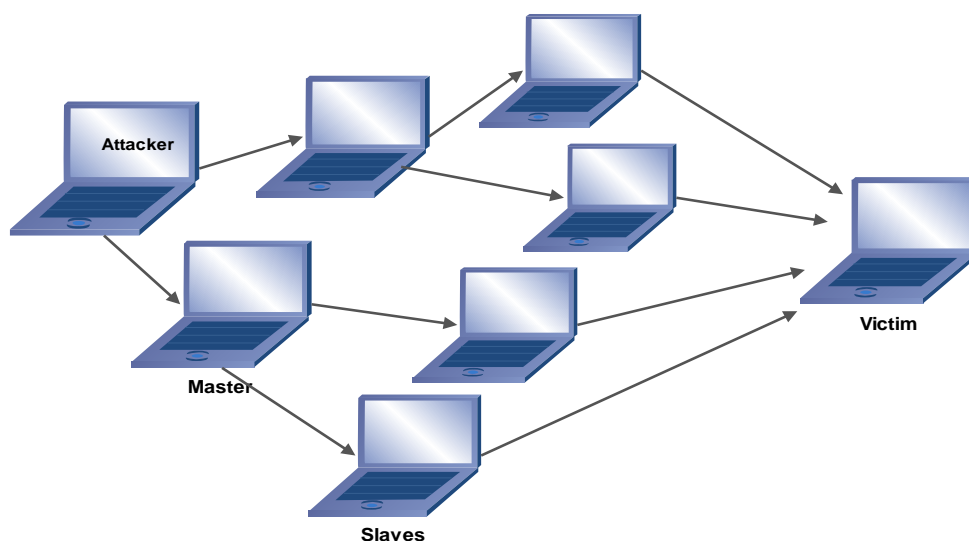


Figure 1. Architecture of DDoS attack

The data was collected for various types of attacks UDP, UDPLag, LDAP, MSSQL, and Portmap. If the request is made by the legitimate user, it will be flagged as "benign" otherwise it will be flagged as a custom attack name. This data set was generated for the purpose of analysis and organization with a continuous daily work, CIC records basic data including event logs and network traffic from each network server. The real dataset contained more than 88 features, but CIC worked on their own dimensionality reduction. They used CICFlowmeter-V3 and produced the most important 88 features for the analysis, and supplied csv files. They have shared PCAP files if someone wanted to extract the feature through the master files [12]. We used this binary data set to convert all types of attack to a unified name "DDOS" and the legitimate use is "benign" with the cancellation of a number of network-related features and in the event that they remain affecting the accuracy of the work and therefore the final number is 81 features in the data set that we relied on in building this model.

2. METHOD

2.1. Preprocessing dataset

Pre-processing of data is a very important part that is used in a phase to deal with actual data. We use it in the event that the data is in an unclear and incomprehensible form, and usually heterogeneous data (existence of errors, outlier values) so this data is incomplete, before implementing data mining algorithms it is necessary to apply preprocessing methods improve the accuracy and quality of data as well as the efficiency of data mining techniques [13]. The operations conducted by the pre-processing is considered an important and essential matter in analyzing and surveying data traffic because of the different traffic patterns in terms of dimensions and coordination. Operation used for preprocessing, as data reduction, data cleaning, data discretization, data integration and data transformation (normalization). Most of the operations used in data normalization such as z-score, decimal scaling normalization and min-max [14]. Attribute data contains values of different numerical. If the model is trained directly on the original data set, a classification error occurs. Then the model consumes a lot of time during training, so we normalize the data set where the upper bound value is one and the lower value is zero. The following equation shows the normalization method used.

$$w = \left(\frac{N_i - \min(N)}{\max(N) - \min(N)} \right) \quad (1)$$

Note that, N_i it is the data element, $\min(N)$ is the minimum for whole data values, and $\max(N)$ is the maximum of whole data values, W is a new value. CICIDS 2019 dataset contains a portion of missing values, so an error occurs in the normalization process. The missing value was processed before the normalization process was performed.

2.2. Feature selection

Feature selection is an important measure for determining the minimum number of appropriate and necessary features for the application of basic classification techniques [15], [16]. Conceptually, defining a subset of features is a search for all possible subsets of features. Multiple types of different search techniques can be used, but it is necessary that the search technique is computationally inexpensive, and it must find feature sets that are close to optimal or optimal. Often it is not possible to achieve both requirements, and therefore trade-offs are necessary [17]. There are a lot of ways to define supervised features and they can be categorized into filter, wrapper and embedded models. Table 1 shows the top 11 important features from the 81 features of the CICIDS2019 data set that we extracted for our assessment based on extra trees classifier importance.

Table 1. Feature selection

No	Feature name	Weight
1	Inbound	0.815
2	Source IP	0.667
3	URG flag Count	0.434
4	Destination IP	0.413
5	CWE Flag count	0.413
6	Fwd PSH Flags	0.413
7	RST Flag count	0.393
8	Bwd packet length mean	0.393
9	Bwd packet length min	0.364
10	Avg Bwd Segment Size	0.336
11	Flow ID	0.316

2.3. Feature weighting

Feature weighting is an important alternative to retaining or removing the feature. It gives a higher weight to the more important feature, while giving less weight to the less important features. Features with large weights play an important role in building the model and results in higher accuracy. Often these weights are determined based on the domain's knowledge about the relative importance of features. Alternatively, it may happen to be selected automatically [9]. Figure 2 illustrates the working mechanism of the method used to select features.



Figure 2. Flowchart of a feature subset selection [16]

2.4. Data mining algorithms

Data mining provides a lot of algorithms that take inputs and convert data into actionable knowledge. "A predictive model is utilized for tasks that include, as the name implies, the prediction of one value using other values in the dataset [18]. The learning algorithm attempts to discover and model the relationship between the objective feature and the other features. The processing of training predictive model is known as supervised learning or classification". There are a lot of supervised learning algorithms like "Decision trees, Naive Bayes, neural networks, support vector machine, and random forests."

In this work, the results were derived by constructing four models. Naive Bayes, random forests, decision tree and logistic regression, they were compared and the best model was chosen. Random forest (RF) algorithm, they are used in many cases and have a powerful approach to data analysis, predictive modeling and data exploration.

This technique was suggested by Breman in 2001. Results (RF) can be obtained from the results of individual decision trees that are constructed from a set of independently learned decision trees. An ensemble classifier (building a set of classifiers by learning algorithms and then classifying new data by a (weighted) rating system for predictions. RF is groups of trees, this technique consists of many decision trees and the output is done through individual trees [19]. It has many features like:

- Provides great and effective services for missing data and ways to deal with it.
- In some decision trees, problem of over processing occurs, so this technique is best solution for it [20].

Naïve Bayes (NB): this algorithm is the core of its currency in Bayes theory and is used when the input dimensions are high. A Bayesian classifier has the ability to compute the output from the input. And also add new data at runtime and earn on best probability classifier. A Naive Bayes classifier states that the presence (or absence) of a feature assigned to a class is not correlated with the presence (or absence) of any other attribute when the class variable is given [21].

Logistic regression: is an algorithm that has the same predictive analysis capability as any other regression analyses. The description of the data is one of the tasks of logistic regression as well explain and illustrates the resulting relationship between classes and attributes [22].

Decision tree (DT): data mining decision tree is one of the important techniques and works in measurements, calculations and machine learning. A decision tree (as an insight model) is used to transform from a specific concept (represented as branches) to decisions about the object's purpose and value (represented as leaves). "Leaves are called class labels and branches are called conjunctions of climax which represents to those class labels. Decision trees where the objective variable can take enduring qualities (ordinarily genuine numbers) are regression trees". Decision trees are considered as one of the most common techniques among data mining techniques because of their intelligibility and clarity [23].

3. EXPERIMENTAL ANALYSES

Our assessment was carried out on 1,048,560 samples from the CICIDS 2019 dataset which divided them by 20% for testing and 80% for training then classify DDoS attack and BENIGN. All experiments were carried out by python language. This contains libraries of different data mining algorithms. Confusion matrix is an important algorithm to summarize classification algorithm performance and behavior. It gives true and false classification results. Calculating confusion, one of the most important ways to correctly explain the idea in the form of a two-dimensional and multi-dimensional matrix of the intent of your classification model and the types of errors it makes. The here are the probabilities of classifying events as in the Table 2.

Table 2. Confusion matrix

Parameters		Predicated class	
		Normal	Attack
Actual class	Normal	TP	FP
	Attack	FN	TN

Through Table 2 shows that "true positives (TP) and true negatives (TN) are valid classifications. False positive (FP) occurs when a result is incorrectly predicted as a yes (or positive) when it "is actually no (negative). False negative (FN) it occurs when the outcome is incorrectly predicted as negative when it is actually positive." The accuracy and performance were verified depending on the selected features as in the table I with three data mining algorithms using 10-fold cross-validation to improve the results. Table 3 shows the confusion matrix results for RF, LR, NB and DT algorithms.

Table 3. The confusionmatrix for the four algorithms

NB		Predicated class		RF		Predicated class	
Actual class	BENGIN	182887	93	Actual class	BENGIN	203053	17
	DDoS	20185	481		DDoS	19	557
DT		Predicated class		LR		Predicated class	
Actual class	BENGIN	203061	29	Actual class	BENGIN	203051	35
	DDoS	11	545		DDoS	21	539

4. PERFORMANCE EVALUATION

Several types of errors came from classification in one way or another affect its strength. It is summarized in the following confusion matrix (Table 3). More specifically, we need to look at the costs of errors [24]. So, we used a set of common metrics to evaluate the information received: positive predictive value or precision (Pr) It is the ratio of properly attack classified flows (TP), in front of whole attack classified flows (TP+FP). Sensitivity or recall (Rc) It is the ratio of properly attack classified flows (TP), in front of whole attack classified flows (TP+FN). Detection rate is the rate of real events that can be predicted to be events. The accuracy, precision, recall, F1 were calculated by applying the set of equations [25].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Precision (Pr) = \frac{TP}{TP+FP} \tag{3}$$

$$Recall(R) = \frac{TP}{TP+FN} \tag{4}$$

$$F - Score = 2 * \frac{P*R}{P+R} \tag{5}$$

Table 4 shows the performance test results of our evaluation metrics for the four most common algorithms selected for data mining are NB, RF, DT and logistic regression. These results depend on the confusion matrices (Table 3) in addition to the performance measurement as shown in (2), (3), (4) and (5).

Among the four classification techniques for dealing with numeric data especially that were evaluated, DT and RF classifiers are ahead of the others with an accuracy rate of 99.98% and 99.98% respectively and for them there is a possibility success (Precision) is 99%. The F1 score to DT and RF are 98.19% and 98.42%

respectively which means they indicate that this experimental method is more ideal. Table 5 shows comparisons with previous research.

Table 4. The performance examination results

Model	Accuracy	Recall	Precision	F1 score	Time consumer
RF	0.9998	0.9851	0.9835	0.9842	4.48 m
NB	0.9004	0.8693	0.5114	0.6438	18 S
DT	0.9998	0.9747	0.9900	0.9819	35 S
LR	0.9997	0.9695	0.9812	0.9761	1.51 m

Table 5. Comparison with previous studies

Ref.	Dataset	Algorithm	Accuracy	
			Previous studies	Our proposed approach
Elsayed <i>et al.</i> [9]	CICDDoS2019	NB	57%	90.04%
		RF	86%	99.98%
		DT	77%	99.98%
		LR	95%	99.97%
Patel and Shukla [21]	CICIDS 2017	C4.5	99.96%	99.98%
		LR	92.49%	99.97%

5. CONCLUSION

This work is based on the use of the CICDDoS2019 dataset of the latest update and also includes the latest DDoS attack. Analyzes and experiments were conducted applying a set of supervised and basic classification algorithms to accurately classify the attack from legitimate streams. When comparing the results with other algorithms we find the best classifiers are, decision trees, random forests, and Naïve Bayes. The main contribution to this work is to achieve the best accuracy and reduce the time consumed relative to previous research by selecting features that contain high weights. In addition, we dealt with all samples in the data set. For future work, we suggest choosing other features and using other data mining algorithms to build an efficient IDS system capable of detecting DDoS attacks.





REFERENCES

- [1] A. Bhati, A. Bouras, U. A. Qidwai, and A. Belhi, "Deep learning based identification of DDoS attacks in industrial application," in *Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020*, Jul. 2020, pp. 190–196, doi: 10.1109/WorldS450073.2020.9210320.
- [2] M. Gohil and S. Kumar, "Evaluation of classification algorithms for distributed denial of service attack detection," in *Proceedings - 2020 IEEE 3rd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2020*, Dec. 2020, pp. 138–141, doi: 10.1109/AIKE48582.2020.00028.
- [3] J. Ye, X. Cheng, J. Zhu, L. Feng, and L. Song, "A DDoS attack detection method based on SVM in software defined network," *Security and Communication Networks*, vol. 2018, pp. 1–8, 2018, doi: 10.1155/2018/9804061.
- [4] M. M. Oo, S. Kamolphiwong, and T. Kamolphiwong, "The design of SDN based detection for distributed denial of service (DDoS) Attack," in *2017 21st International Computer Science and Engineering Conference (ICSEC)*, Nov. 2017, pp. 1–5, doi: 10.1109/ICSEC.2017.8443939.
- [5] L. F. Rahman, T. Ozcelebi, and J. Lukkien, "Understanding IoT systems: a life cycle approach," *Procedia Comput. Sci.*, vol. 130, pp. 1057–1062, 2018.
- [6] F. A. Fernandes Silveira, F. Lima-Filho, F. S. Dantas Silva, A. de Medeiros Brito Junior, and L. F. Silveira, "Smart detection-IoT: A DDoS sensor system for internet of things," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jul. 2020, vol. 2020-July, pp. 343–348, doi: 10.1109/IWSSIP48289.2020.9145265.
- [7] S. S. Mohammed *et al.*, "A new machine learning-based collaborative DDoS mitigation mechanism in software-defined network," in *2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Oct. 2018, vol. 2018-October, pp. 1–8, doi: 10.1109/WiMOB.2018.8589104.
- [8] S. Yadav and S. Subramanian, "Detection of application layer DDoS attack by feature learning using Stacked AutoEncoder," in *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, Mar. 2016, pp. 361–366, doi: 10.1109/ICCTICT.2016.7514608.
- [9] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *2019 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2019, vol. 2019-October, pp. 1–8, doi: 10.1109/CCST.2019.8888419.
- [10] M. S. Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "DDoSNet: a deep-learning model for detecting network attacks," in *2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Aug. 2020, pp. 391–396, doi: 10.1109/WoWMoM49955.2020.00072.
- [11] Canadian Institute for Cybersecurity, "DDoS 2019," 2019, Accessed: Apr. 26, 2022 [Online]. Available: <https://www.unb.ca/cic/datasets/ddos-2019.html>.
- [12] M. Shurman, R. Khrais, and A. Yateem, "DoS and DDoS attack detection using deep learning and IDS," *The International Arab Journal of Information Technology*, vol. 17, no. 4A, pp. 655–661, Jul. 2020, doi: 10.34028/iajit/17/4A/10.





- [13] F. Alotaibi and A. Lisitsa, "Matrix profile for DDoS attacks detection," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems, FedCSIS 2021*, Sep. 2021, pp. 357–361, doi: 10.15439/2021F114.
- [14] B. S. H. Jaddoa and R. S. Al-Hamdani, "Design of detection and defense system against cloud attacks using data mining techniques," Thesis, Iraqi Commission for Computers and Informatics, Baghdad, Iraq, 2018.
- [15] R. E. Noonan and S. Fegock, "Feature selection using particle swarm optimization in intrusion detection," in *Proceedings of the 1979 SIGPLAN symposium on Compiler construction*, 1979, vol. 14, no. 8.
- [16] S.-J. Yang and H.-L. Huang, "Design a hybrid flooding attack defense scheme under the cloud computing environment," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Jun. 2019, pp. 41–46, doi: 10.1109/ICIS46139.2019.8940313.
- [17] P. N. Tan *et al.*, *Introduction to Data Mining*, First Edition, Boston: Pearson, 2005.
- [18] S. Agarwal, "Data mining: data mining concepts and techniques," in *2013 International Conference on Machine Intelligence and Research Advancement*, Dec. 2013, pp. 203–207, doi: 10.1109/ICMIRA.2013.45.
- [19] B. E. Lowe, "The random forest algorithm with application to multispectral image analysis," Thesis, University of Texas at Tyler, 2015.
- [20] O. Mbaabu, "Introduction to random forest in machine learning," section.io, Dec. 11, 2020, Accessed: Jun. 06, 2022, [Online]. Available: <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>.
- [21] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Oriental Journal of Computer Science and Technology*, vol. 8, no. 1, pp. 13–19, 2015, doi: 10.21884/ijmter.2017.4211.vxayk.
- [22] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, "Logistic regression for machine learning in process tomography," *Sensors*, vol. 19, no. 15, p. 3400, Aug. 2019, doi: 10.3390/s19153400.
- [23] A. Sudugala, W. Chanuka, A. M. Eshan, U. C. Bandara, and K. Abeywardena, "WANHEDA: a machine learning based DDoS detection system," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, Dec. 2020, pp. 380–385, doi: 10.1109/ICAC51239.2020.9357130.
- [24] A. Saber, M. Abbas, and B. Fergani, "A DDoS attack detection system: applying a hybrid genetic algorithm to optimal feature subset selection," in *ISIA 2020 - Proceedings, 4th International Symposium on Informatics and its Applications*, Dec. 2020, pp. 1–6, doi: 10.1109/ISIA51297.2020.9416558.
- [25] A. K. Santra and C. J. Christy, "Genetic algorithm and confusion matrix for document clustering," *International Journal of Computer Science*, vol. 9, no. 1, pp. 322–328, 2012.

BIOGRAPHIES OF AUTHORS



Dhurgham Kareem Gharkan     Master's student at the Iraqi Commission for Computers and Informatics, Institute of Informatics for Graduate Studies, his fields of research are intrusion detection system and data mining. He can be contacted by e-mail: ms202020630@iips.icci.edu.iq.



Amer A. Abdulrahman     is Assistant Professor at college of Sciences, University of Baghdad, Iraq. He Holds a Ph.D. degree in Computer Sciences with specialization in network security. His research areas are intrusion detection system, big data analysis and online machine learning. He can be contacted at email: amer.abdulrahman@sc.uobaghdad.edu.iq.