Mansoura University
Faculty of Engineering
Mansoura Engineering Journal

Development Machine Learning Techniques to Enhance Cyber Security Algorithms

Ghada Mohamed, *Ehab H. Abdelhay*, *Ibrahim Yasser* and Mohamed A. Mohamed

KEYWORDS:
Cloud computing, DDoS attacks, Machine Learning, Random forest , and Weka.

***Abstract*— Nowadays, Cyber security threats are a growing global problem. As technology evolves, cyber threats, including cyber-hacking threats, and cybercrime organizing groups, are on the rise. Distributed Denial of Service (DDoS) is one of the most serious attacks faced by Cloud computing. This attack aims to make cloud services unavailable to end-users by exhausting system resources, resulting in heavy losses that pose a threat to national security and information security assets, and thus making the development of defensive solutions against such attacks necessary to expand the use of Cloud computing technology. Machine learning (ML) has promising results in detecting cyber-attacks including DDoS when applied to intrusion detection systems. In this research, the proposed system was built using Random forest (RF) is supervised machine learning algorithm, which is an ensemble learning method that operates by constructing a multitude of decision trees at training time. The experiments conducted using the most common and standard data sets, NSL-KDD, and CICIDS 2017, achieved a detection accuracy of up to 99.09% for the first dataset and 99.97% for the second dataset respectively. The proposed system performs well when compared to other methods in terms of accuracy, detection rate, and low false-positive rate.**

I. INTRODUCTION

IN the era of transformation, and because of the COVID-2019 epidemic, life has almost completely turned to the Internet since the beginning of 2020. As a result, cyber security needs to find innovative ways to improve and develop its capabilities.

Received: (29 June, 2021) - Revised: (17 October, 2021) - Accepted: (17 November, 2021)

***Corresponding Author:** Ghada Mohamed Amer, Communications Engineer at North Delta Electricity Distribution Company, Faculty of Engineering, Mansoura University. (E-mail: g.a.amer@outlook.com).*

Ehab Hany Abd El Hay, Assistant Prof., Faculty of Engineering, Mansoura University. (E-mail: ehababdelhay@mans.edu.eg).

Ibrahim Yasser, Assistant Lecturer, Nile Higher Institute for Engineering and Technology. Faculty of Engineering, Mansoura University. (E-mail: ibrahim_yasser@mans.edu.eg).

Mohamed Abd El Azim Mohamed, chief, Dean of the College and Chairman of the Board of Directors, Faculty of Engineering, Mansoura University. (E-mail:mazim12@mans.edu.eg).

Cloud computing provides technical services, platforms, and IT software such as Internet services [1]. Its main objective is to allow users to use what they want and to pay for promising on-demand services to meet their software or infrastructure needs, and they are gradually included by organizations as private, public, or hybrid clouds [2]. The attractive features of Cloud computing continue to integrate into many sectors including industry, government, education, and entertainment, to name a few.

Although Cloud computing is an important and positive shift, some security issues hinder the use of this technology. Distributed denial of service (DDoS) attack is one of the main attacks in cloud services [3]. Traditionally, DDoS attackers target a server that serves its customers. The attackers, acting as real customers, try to flood the server in such a way that the service is unavailable due to frequent data requests and a busy service queue [4].

The global digital transformation will continue to have a significant impact on Cloud computing, and DDoS attacks will be one of the main concerns of this period. According to

Cisco, as shown in ***Fig. 1***, DDoS attacks will double by 15.4 million by 2023 worldwide compared to 2018 [5]. DDOS attacks increased significantly last year after the digital transformation created by the Coronavirus, according to Kaspersky's DDOS report, DDoS attacks in the second quarter (Q2) of 2020 were 217% higher than the same period last year [6]. DDoS attacks in (Q4) 2020 also increased by only 10% compared to the fourth quarter of 2019. Compared to (Q3) 2020, the number of attacks in (Q4) 2020 decreased by 31%, while (Q3) of 2020 decreased [7].

The most used mechanisms that identify a DDoS attack consist of several stages: preventing, detecting, and reacting to the attack. Intrusion detection has become a necessary component for building network security to detect abnormal use of the system by monitoring and analyzing network behavior to detect an attack. Though there are many methods to fight DDoS attacks, the best ones are the proactive and reactive methods. Proactive mode provides the highest accuracy detection capabilities by constantly searching for potential attackers. This mode uses a built-in tool that has very high visibility through packet analysis, thus checking every bit of the traffic received using pre-defined information and behavioral indicators. It then determines what bots or attacks are and then blocks them. Since the system is always on, proactive mode tends to be costly, especially in the case of a large network. In reactive mode takes advantage of the flow data available from routers and peripheral keys and analyzes metadata for anomalies. When this analysis leads to the discovery of something potentially dangerous. It is interacted with by inserting a dilution device. So, it's interactive in nature, which means the mitigation device is activated only when a risk is detected, so this method is cost-effective but actual response time is sacrificed [8].

It remains a difficult task to detect increasingly complex network attacks. Machine learning (ML) has promising results in all technologies including cyber security and provides us with intelligence when applied to intrusion detection systems. ML techniques have been the best solution for a quick and accurate prediction of a DDoS attack to combine computer science with statistics [9]. Machine learning can be classified into two main types as follows: supervised learning and unsupervised learning. Supervised machine learning relies on labeled data that is trained to teach models to yield the desired output. The dataset is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification. There are some problems in the performance of systems that rely on Machine learning techniques, such as low detection accuracy, high training time, and a high rate of false alarms. To overcome these problems, the Random forest algorithm was used in this proposed system.

The main contributions of this paper can be summarized as follows:

A. Random forest classification algorithm based on machine learning is proposed for intrusion detection DDoS attacks in Cloud computing.

- The performance of the proposed system is being evaluated using two datasets, the NSL-KDD dataset and the ISCX intrusion detection dataset.
- The performance of the proposed system is compared with other algorithms Adaboost, Bayes Network learning, multi-layer perceptron(backpropagation), support vector machine (SMO), and K-nearest neighbors (IBK).
- The proposed system can be investigated for various parameter values.

This paper is structured as follows. Relevant work is described in Section 2. Section 3 describes how DDoS attack detection techniques are classified. Section 4 describes the proposed attack detection system. Experimental results and various analyzes of the results are given in section 5. Finally, Section 6 concludes the treatise.

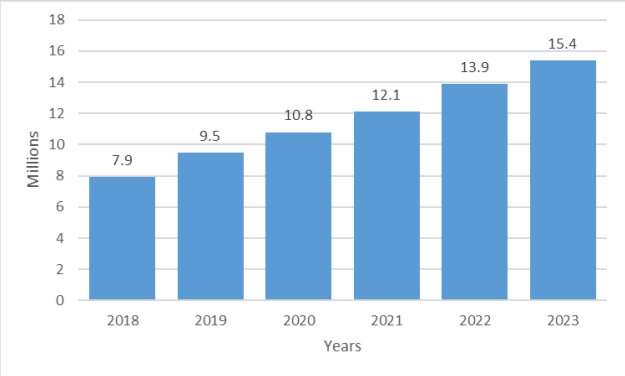

Fig 1. Global DDoS attacks, 2018-2022.

II. RELATED WORK

This section provides some previous work devoted to improving the performance of DDoS attacks for intrusion detection in Cloud computing.

Khalaf et al. [10] Provided a comprehensive and detailed review of statistical approach and artificial intelligence using Bayesian networks, fuzzy logic, genetic algorithms, K-NN, neural networks, software factors, and support transmission machines in detecting and preventing DDoS attacks. He also divided DDoS attacks according to the vulnerability, degree of automation, impact, and dynamics category.

Hosseini and Azizi. [11] Proposed the hybrid framework to detect the DDoS attack. They used naïve bays, Random forests, resolution trees, multilayered cognition (MLP), and K-NN to discover high-speed DDoS.

Wani et al. [12] Used machine learning algorithms to detect high-priced DDoS attacks in a cloud environment. Using various machine learning algorithms such as vector machine supports, naïve bays, Random forest classification, and total accuracy 99.7%, 97.6%, and 98.0% of support carrier machine, Random forests, and naïve bays respectively.

Alsirhani et al. [13] Proposed a DDoS detection system using a set of classification algorithms: Naive Bayes, Decision Tree (Entropy), Decision Tree (Gini), Random forest) controlled by a fuzzy logic system in Apache Spark.

Sharma, Verma, and Sharma. [14] Using isolated forest anomaly detection technology, they analyzed and proposed

various ML algorithms for detecting DDoS attacks in a Cloud computing environment.

Shamsolmoali and Zareapoor. [15] A statistical technique has been introduced to discover and filter DDoS attacks and can ease most TCP attacks accurately revealing up to 97%.

Xiao et al. [16] An effective detection approach depends on CKNN (closest neighbors' traffic classification to K with link analysis to discover DDoS attacks. The link information is used for training data to improve the accuracy of the classification and reduce the overall expenditure resulting from the density of training data. The method of the network is called The training data included in the account.

Kuang et al. [17] A method based on a support vector device has been proposed. Analysis of key nucleus components is being used to reduce the advantage and improve the family improvement of chaotic particles used to improve different parameters.

Zekri et al. [18] Suggested hybrid technology. Use the Snort tool-based tool to detect known attacks and unknown attacks, use the resolution tree workbook (C4.5).

Kushwah and Ali. [19] Proposed a model to detect DDoS attacks according to ANN. ANN training uses black hole optimization algorithms.

Kushwah and Ranga. [20] They have proposed a new system to detect DDoS attacks in the Cloud computing environment. This system was built using V-ELM (extreme voting learning machine) and compared to other ML algorithms. Experiments were also conducted to analyze the performance of the proposed system with other parameter values.

Sofi et al. [21] used weka tool to detect anomaly in the network traffic and conclude that an efficient detection algorithm to detect DDoS attack.

III. DDoS Attack Classification Techniques

The Intrusion detection system is one of the most common employment solutions with DDoS problems and privacy and privacy, integrity, and availability of web services and computer networks. Intrusion Detection is the process of examining actions that occur in computer systems or networks and analyze them for signs of possible events that suffer or imminent threats to contain the computer security policies are adopted the guidelines for the use of policies or standard security practices [22].

There are three intrusion detection techniques. These are signature-based, anomaly-based, and hybrid-based Intrusion Detection Systems. DDoS Attack Classification Techniques are indicated in ***Fig. 2.***

A. Signature-Based Techniques:

Signature-based detection is achieved by comparing the information collected from a network or a database system. This technique also knows as an abuse discovery. In the cloud environment, the sneak detection method can be used to sign in the front end of the cloud to discover known attacks from the external network. It can also discover internal and external interventions if organized in the back-end cloud. Bakshi and Yogesh [23] suggest a solution to detect DDOS attacks based on the Signing Slot Detection System. IDS is installed on the default adapter to monitor traffic in both directions, incoming and outgoing. Lo, Huang and Ku. [24] The proposed system reduces the effect of DDOS attacks. The IDSS in Cloud computing areas is alerted with each other. In the system, each IDSS contains a useful factor that is used to calculate and control the alerts sent from other IDSs or not. The problem is to sign the signature in that with the recognition of new attack plans, the IDS signatures database must be updated frequently.

B. Anomaly-Based Techniques:

During the normal period, a network profile is created using these techniques. Deviation from the normal profile is used to detect attacks. These techniques can detect previously unknown attacks [25]. They are divided into three sub-divided are statistical, machine learning, and SDN-based models:

1) Statistical techniques:

Statistics-based techniques create general profiles using statistical attributes such as general contract averages and changes. Statistical tests strive to see if the observed transactions differ from the normal profile. IDS assign a score to transactions whose profile is not normal. When the score reaches the threshold, the alarm goes up. Wu et al. [26] proposed a unique real-time DDoS detection scheme in the SDN environment by using the principal component analysis (PCA) method to analyze the traffic packet data network state and reduce the total computational cost. The problem with statistics-based methods is that they need an accurate statistical distribution. The learning process of statistical-based techniques takes a long time to be accurate and effective.

2) Machine learning -based techniques:

Through Machine learning, systems can be individually selected without external assistance. These choices are made when the machine can learn from the data and understand the basic patterns contained therein. It then returns the results, classifications and predictions via pattern matching , and additional analysis [27].

IDS machine learning models mainly include artificial neural network (ANN), support vector machine (SVM), K nearest neighbor methods (KNN), Bayes NET, and Random forest for supervised learning. In unsupervised learning use clustering and combined and hybrid methods [21].

- Artificial Neural Networks are imitating the way human minds work. ANN consists of several hidden layers, an input layer, and an output layer. Units in the adjacent layers are completely connected. Contains many ANN units and can theoretically approximate arbitrary functions; As a result, it has excellent capacity appropriate, especially for non-linear functions. ANNs training takes time because of the complex model structure. It should be noted that ANN models are trained using a rear shackle algorithm, which cannot be used to train deep neural networks [19].

- Support Vector Machines (SVM) are strong classifiers used to classify the binary dataset into two categories with superior aircraft. It can be an effective way to discover vulgarities in the case of limited data samples, where dimensions will not change accuracy [28].
- K-nearest neighbors are one of the greatest basic yet important classification algorithms in machine learning. These KNNs are used in real-life situations where non-parametric algorithms are required. These algorithms do not make any molds about how the data is dispersed. When we are given previous data, the KNN classifies the organizes into groups that are identified by a specific attribute [29].
- A Bayesian Network (BN) is an obvious cyclic graph. It refers to JPD on a set of V random variables. By using a directed graphical model, Bayesian Network labels random variables and conditional dependencies Bayesian networks are appropriate to represent probability and predictability of potential causes and contributing factors.
- Random forest is an ensemble learning method for execution classification, regression, and other tasks by providing the output as a class that is the default individual tree method or mean for building decision trees. The idea I have put off this way is to disassociate some trees. An ensemble technique called bagging is like a Random forest. It is generated from various bootstrap samples from the training data. And by averaging them, we reduce the change in trees. Therefore, this approach produces many decision trees. During training, Random forest ensemble learning methods can be categorized, and thus many decision trees can be constructed and operated [30].

3) SDN-Based Techniques:

Software-Defined A software-defined network (SDN) provides a starting point for the data plane and control plane. The controller centrally controls the entire network. SDN provides the ability to program the network and enables the dynamic formation of flow policies. The console is vulnerable to distributed denial of service (DDoS) attacks, because of which resources are fatigued, and the services provided by the console are inaccessible. DDoS detection requires an adaptive accurate classifier to make decisions from uncertain information. Early detection of an attack on the controller is risky. The implementation of SDN consists of three layers: the data plane and the SDN controller application layer. These technologies are only available when SDN is used in cloud networks [31].

C) Hybrid Techniques:

Hybrid detection technology The efficiency of IDS can be greatly improved by combining signature-based and anomaly-based techniques. The catalyst behind this combination is the ability to detect both known and unknown attacks using signature-based and anomaly-based detection techniques. The problem with these technologies is that resource consumption is extremely high [32].

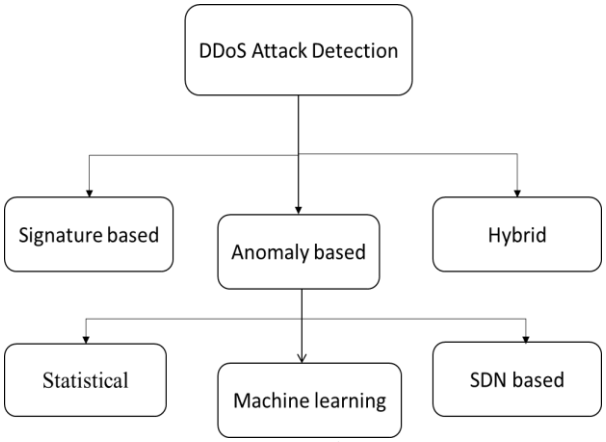

Fig 2. DDoS Attack Classification Techniques

IV. PROPOSED SYSTEM

In Fig. 3, the proposed DDoS attack detection with the cloud is shown. The detector is connected to a cloud network, which monitors all traffic flowing to and from the cloud. The internal structure of the detector is shown in Fig. 4. It contains 3 modules: training database module, preprocessing module, and classifier module.

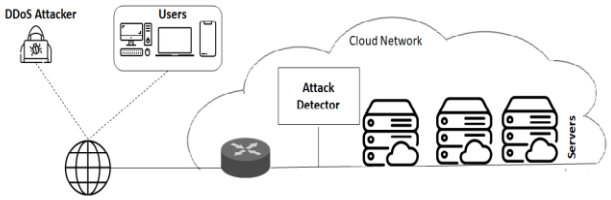

Fig 3. The proposed DDoS attack detection system is a cloud network.

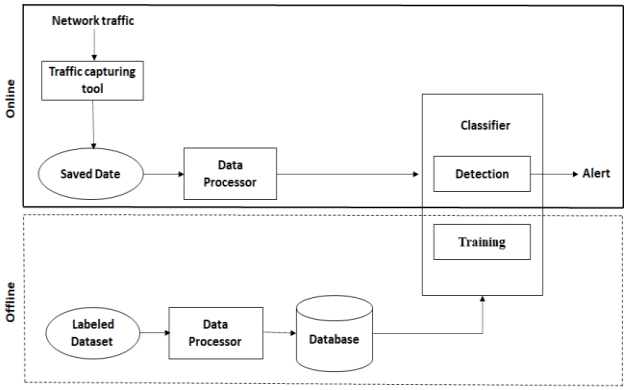

Fig 4. DDoS Attack Detector.

A. Training Database Module:

The detector relies on a supervised classifier, which means the classifier must be trained before it is used to detect attacks. To train the classifier, an NSL-KDD data set is used. To create the training database, the data of the previous network flows are taken. Network traffic data consists of various parameters, such as flow, time, content, and basic features, and each feature contains information about both types of properties of

normal and anomalous packets in the original data set. Features are generated using both transaction connection times and transaction flow identifiers to mathematically represent potential features of network observations. First, the network traffic features are extracted for each packet with the type of packet. These features are represented in symbolic and/or, numeric datasets. Next, the symbol attribute value is then replaced with a number. Then normalization is done. Data normalization is the process of readjusting attributes in the range 0 to 1. Normalization is important for learning because it eliminates bias in instances of the raw network without losing statistical attributes.

The resulting values are [0, 1] for the data used to compute the normalization intervals. These features are used to classify samples as normal samples or anomaly samples. Since it is a binary classification, normal samples are classified as 0, and the anomaly samples are classified as 1.

B. Preprocessor Module:

The pre-processing module works always captures network traffic and working samples to use by the classifier. Samples are made in sets consists of decision trees. Network traffic is captured during each period using the traffic capture tool. The captured data is saved in a separate node. For each node, features are specified in the same features that were used in the training database samples. Test sample takes and uses the random-created decision tree rules for predicting a classification. The results of the final classification by voting for these trees.

C. Classifier Module:

In this work, Random forest has been used as a classifier. Because It reduces the risk of over fitting and easily determines the importance of the feature. In addition, it maintains accuracy when a portion of the data is missing because the bagging feature makes Random forest classification an effective tool for estimating missing values. RF is one of ensemble classification methods that uses a bagging approach to builds decision trees on different samples to classification the result of RF is acquired by majority vote. RF consists of many individual decision trees that operate as an ensemble at training time to output the class for classification. An RF algorithm is a combination of a training phase and a testing phase. The training phase uses a bootstrap sampling method to generate various subsets of the training data. When using the bootstrapping technique, about one-third of the samples are not present in the [InBag]. These samples are known as Out-Of-Bag Data [OBB] [33]. OOB data is used to obtain an unbiased estimate of the prediction error as trees are added to the forest during the construction phase. Because OOB data are compared with predicted values at each step, this data plays an important role in the growth of the tree. Trees are created in the forest in a way that has a lower error rate than OOB data retrieval values. Then, by training these subsets, a decision tree is built. Finally, every trained decision tree is made up of RF. The test phase uses each randomly generated decision tree rule to get the test function to predict the outcome and store the predicted values. Calculate the votes for each predicted target. The final prediction obtained by considering the majority vote is classification trees. Fig. 5 shows the main structure of the RF algorithm.

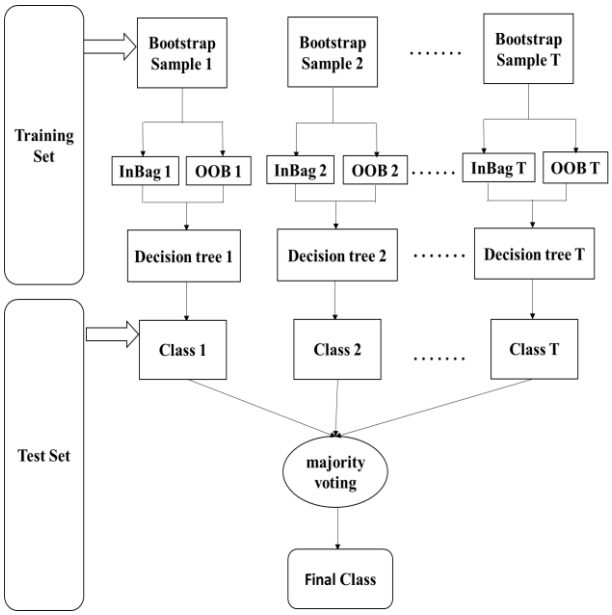

Fig 5. RF Algorithm Structure

Assuming the training samples for T = [T_1, T_2,....., T_n], i=1,….,n with T_i (x_i , y_i) where x ∈ R^d contains d characteristics and y_i ∈ [0,1] is the class of x_i. The main process of the RF algorithm is shown as follow:

1. Replace training samples C to generate bootstrap resamples B1, B2,….., BM.
2. For each resamples Bm, grow a decision tree DTm.
3. At each split, only predictors in a randomly selected subset of DDoS sample or normal sample.
4. Each tree is grown until all nodes contain notes no more than the maximal terminal node size.
5. For predicting the test case, the predicted value by the total RF is obtained by combining the results given by single trees.
6. The final prediction of the Random forest algorithm has a majority vote of all classification trees.

Since it is a binary classification, the normal traffic label is defined as 0, and the DDoS attack passes are referenced as 1. If the DDoS attack sample is detected, an alert will be created for cloud network administrators. ***Fig. 6.*** illustrates the proposed system flow scheme.

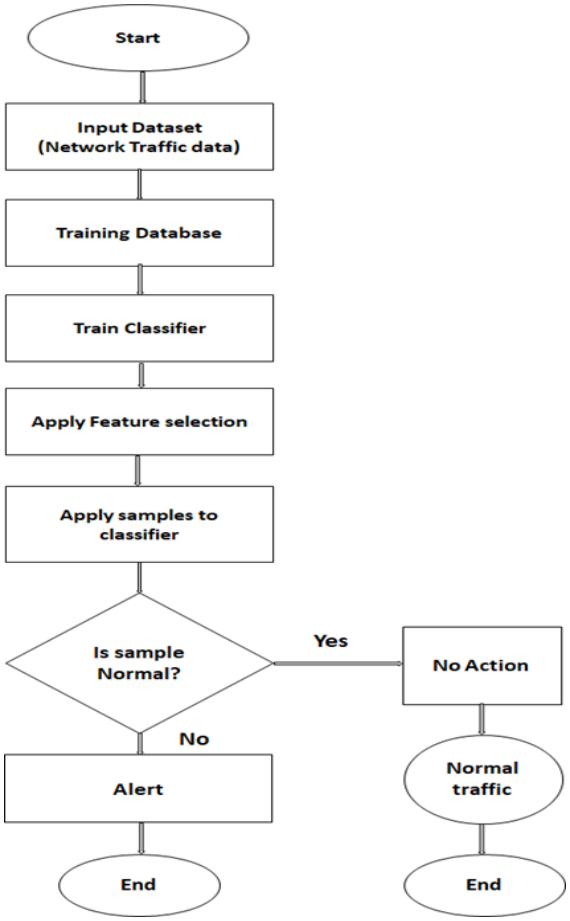

Fig 6. The proposed system flow scheme

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiments were performed to evaluate the proposed system performance used by Waikato Environment for Knowledge Analysis [34]. Weka was developed at Waikato University in New Zealand as a tool used to analyze data and predictive modeling and consists of visualization as tools and algorithms.

TABLE 1
DATASET INFORMATION

Dataset	Number of features	Benign Traffic	Attacks
NSL-KDD	41	22395	23530
ISCX -IDS	75	18150	18370

A. Datasets

The performance of the proposed system was evaluated using two benchmark datasets: the NSL-KDD dataset [35] and the ISCX intrusion detection dataset [36]. For more details about the used datasets give in ***Table1***.

B. Data Pre-processing

Data pre-processing is necessary since it allows for the enhancement of experimental data. Because the algorithms learn from the data, and the learning outcome for issue solving is largely dependent on the right data needed to solve a particular problem – which is termed features.

The process of data pre-processing is carried out using Weka's Filter Classifier, which consists of data cleaning, and transforming the data into the desired format for data extraction. In addition, a class balancer to be reweighted the instances in the data so that each class has the same total weight. The total sum of weights across all instances will be maintained.

C. Training

The model was trained for both datasets separately using the Random forest classifier algorithm by Weka. 32,145 training samples were used for the NSL-KDD dataset and 25,565 training samples for the ISCX dataset.

D. Testing

After training, tests are run. For testing, 13,780 test samples from the NSL-KDD dataset and 10,955 test samples from the ISCX dataset were used.

E. Performance Evaluation and Discussion of Results

For evaluating performance, metrics such as accuracy, Precision, Recall, and total false prediction have been calculated according to the confusion matrix given in Table6. Accuracy is the proportion of correct positive classifications over the total classifications. Detection rate (DR) is the proportion of the total number of assaults detected by the system to the total number of attacks in the dataset. False alarm rate (FAR) is the number of false alarms per the total number of warnings or alerts in each study or situation. Precision is the proportion of correct positive classifications of all cases that are expected as positive. The recall is the correct positive correct of all positive cases. The following equations are defined for evaluating and discussions:

$$Accuracy = \frac{tp+fn}{tp+fp+tn+fn} \times 100 \quad (1)$$

$$Detection\ rate = \frac{tp}{tp+tn} \times 100 \quad (2)$$

$$False\ alarm\ rate = \frac{fp}{tp+tn} \times 100 \quad (3)$$

$$Precision = \frac{tp}{tp+fp} \times 100 \quad (4)$$

$$Recall = \frac{tp}{tp+fn} \times 100 \quad (5)$$

Where,

- True Positive (tp) =The number of DDoS attacks identified as attacks.
- True Negative (tn) = The number of samples which that defined as belonging to normal (benign).

- False Positive (fp)= The number of samples which that defined as belonging to normal but incorrectly identified as an attack.
- False Negative (fn)= The number of samples which that defined as belonging to the attack but incorrectly identified as normal.

TABLE 2
2X2 CONFUSION MATRIX

	Predicted Value	
Actual Value	True Positive	False Negative
	False Positive	True Negative

1) NSL-KDD dataset results and performance evaluation

To verify the effectiveness of the proposed system. The Random forest algorithm is compared with other ML algorithms such as Artificial Neural Networks, Support Vector Machine, K-nearest neighbors, and Bayesian Network. The NSL-KDD dataset results appear in ***Table 3***. And found that RF and KNN algorithms outperform performance on other comparative algorithms. They obtain the accuracy of 99.09% and 97.49% for the NSL-KDD dataset, respectively. Moreover, it can be seen the Bayes Net has a low execution time for two NSL-KDD datasets.

TABLE 3
THE PROPOSED SYSTEM PERFORMANCE OF NSL-KDD DATASET.

Dataset	**Algorithm**	**Accuracy %**	**Precision %**	**Recall %**	**Execution time (Sec)**
NSL-KDD	Adaboost	92.1	92.1	92.1	8.36
	Bayes Net	95.3	95.3	95.3	2.86
	KNN (IBK)	97.49	97.5	97.5	54.95
	ANN (backpropa--gation)	94.62	94.6	94.6	98.43
	Random forest	**99.09**	**99.1**	**99.1**	**28.98**
	SVM (SMO)	94.36	94.4	94.4	603.47

There are two possible false classifications. The first one is the classification algorithm predicts the traffic is benign, but it is attacked, this is a false positive prediction. Second is the classification algorithm predicts the traffic is attacking but it is benign, this is a false negative prediction. In a real application, the true classification of the attack traffic may be more important than the true classification of the benign traffic. From ***Table 4*** can be seen that the Adaboost has a high False Alarm Rate (FAR) on datasets because the number of incorrectly classified DDoS is higher than that in other algorithms.

TABLE 4
CONFUSION MATRIX ON NSL-KDD DATASET.

Adaboost	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6389	616
Normal	472	6301

ANN	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6636	369
Normal	372	6401

Bayes Net	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6652	353
Normal	295	6475

RF	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6948	57
Normal	69	6704

KNN	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6858	147
Normal	198	6575

SVM	**Predicted class**	
Actual Class	DDoS	Normal
Anomaly	6656	349
Normal	428	6345

2) Result and Performance Evaluation for ISCX -IDS dataset

The result of the ISCX -IDS dataset appears in table 5. The RF is obviously better than other classification algorithms followed by KNN (IBK), Bayes Net, Adaboost, SVM (SMO), and ANN. From Table 6 can be seen that the SVM (SMO) followed by ANN has a high False Alarm Rate for ISCX datasets.

TABLE 5
THE PROPOSED SYSTEM PERFORMANCE OF THE ISCX DATASET

Dataset	**Algorithm**	**Accuracy %**	**Precision %**	**Recall %**	**Execution time (Sec)**
ISCX -IDS	Adaboost	99.84	99.8	99.9	12.69
	Bayes Net	99.92	99.9	99.99	4.15
	KNN (IBK)	99.95	99.9	99.9	41.24
	ANN (backpropag-ation)	97.96	98	98	50.93
	Random forest	**99.97**	**100**	**100**	**23.12**
	SVM (SMO)	98.03	98.1	98	28.75

TABLE 6
CONFUSION MATRIX ON ISCX DATASET.

Adaboost	**Predicted class**	
Actual Class	DDos	Normal
DDos	5470	9
Normal	9	5469

ANN	**Predicted class**	
Actual Class	DDos	Normal
DDos	5478	1
Normal	222	5256

Bayes Net	**Predicted class**	
Actual Class	DDos	Normal
DDos	5470	9
Normal	0	5478

RF	**Predicted class**	
Actual Class	DDos	Normal
DDos	5476	3
Normal	0	5478

KNN	**Predicted class**	
Actual Class	DDos	Normal
DDos	5476	3
Normal	3	5475

SVM	**Predicted class**	
Actual Class	DDos	Normal
DDos	5479	0
Normal	216	5262

3) *Comparison of the Results of the two Datasets*

A good classifier must have a high detection rate and a low false alarm rate to detect attacks. As shown in Tables 3 and 4, as well as Fig. 7, and Fig. 8, the proposed model provides the highest percentage of the detection rate and the lowest percentage of the false alarm rate. Referring to Tables 5 and 6 above, the RF algorithm was shown to produce better results for performance metrics that detect DDoS attacks by combining two sets of performance data for different rating algorithms. Thus, the proposed system is the maximum accuracy and detection rate because runs efficiently on a large database, so produces highly accurate predictions. RF can maintain accuracy when a large proportion of data is missing.

As well as a very low false rate with less training time, because they tend to tightly match all samples within training data, decision trees reduce the risk of overfitting. This indicates the success of the system and the overcoming of some problems that appeared in the literature review.

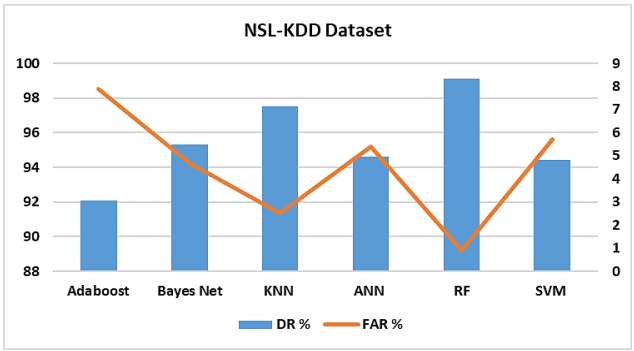

Fig 7. DR and FAR ratios for the NSL-KDD dataset.

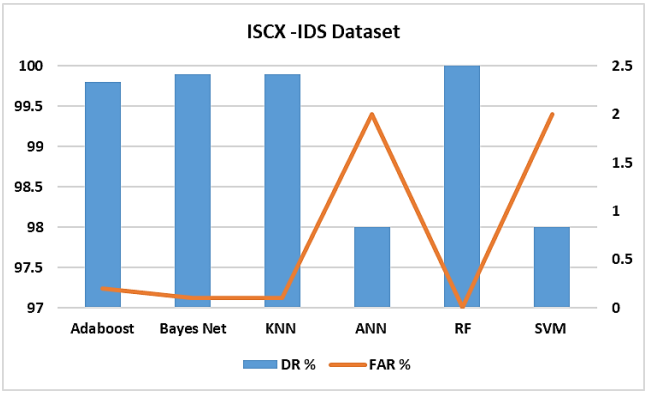

Fig 8. DR and FAR ratios for the ISCX -IDS dataset.

4) *The efficiency of different parameters in RF*

This part displays the effect maximum depth of trees and the number of RF trees. To discuss the effect of maximum depth of trees on accuracy and FAR in the FR algorithm. The maximum depth of trees was set from 1 to 30 and the number of trees is 90100 and 110.

Accuracy and FAR to NSL-KDD are given in Fig. 9, and Fig. 10, respectively. It is also illustrated by numbers when increasing trees and maximum depth trees. The detection performance has been greatly improved in terms of accuracy and FAR. Until the maximum depth of the trees reaches 22, the accuracy, and FAR will be quite stable.

The ISCX accuracy and FAR results are shown in Fig. 11, and Fig. 12. The detection performance is increasing first and then settles when the maximum tree depth reaches 15. These results show that with increased trees, RF can achieve better results. However, increasing the number of trees will also increase training time.

Because as the max depth of the decision tree increases, the performance of the model over the training set increases continuously. As the maximum tree depth value increases, the performance over the test set increases initially but after a certain point, it starts in stability. Among the parameters of a decision tree, maximum tree depth works on the macro level by greatly reducing the growth of the decision tree.

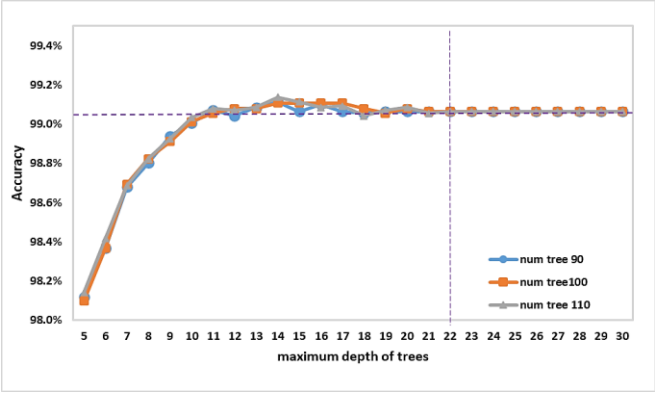

Fig 9. The effects of number of trees -Accuracy.

TABLE 13
COMPARISON OF THE PROPOSED ML RF-BASED ALGORITHM

Ref.	Dataset	Accuracy %	FAR%
[37]	NSL-KDD	98.23	0.33
[37]		97.4	0.45
[37]		96.38	**0.01**
[12]		98.8	0.05
Proposed System		**99.09**	**0.01**

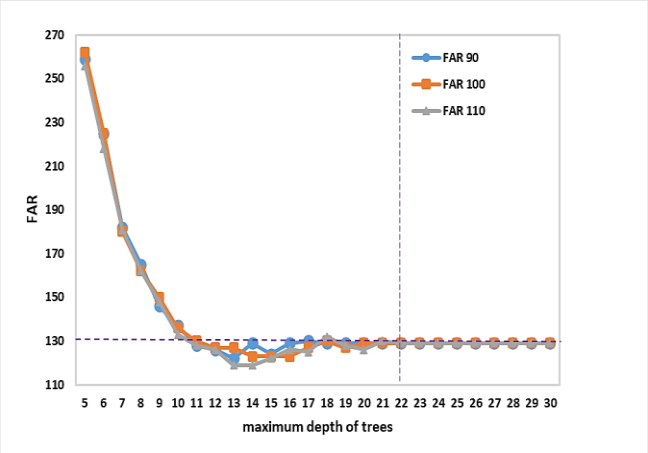

Fig 10. The effects of number of trees – FAR.

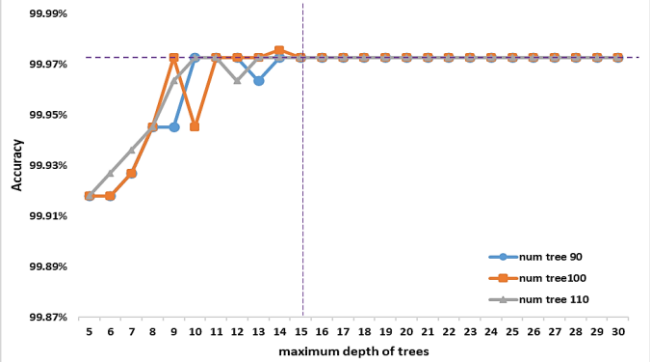

Fig 11. The effects of number of trees – Accuracy.

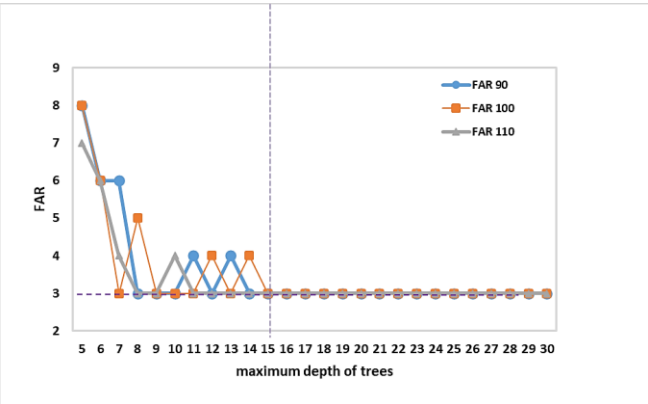

Fig 12. The effects of number of trees – FAR.

5) Comparison for performance metrics of the proposed ML RF Based algorithm with the state-of-the-art DDoS attack detection methods

To check the results obtained, they were compared with several DDoS attacks detection methods. All comparison results are summarized in Table 13. for the NSL-KDD because it is the most used benchmark dataset.

From Fig. 13, this paper has a higher accuracy of 99.09% and a low FAR of 0.01% some algorithms that can detect a DDoS attack. The experimental results of the proposed model were of higher accuracy with a lower false-positive rate (FPR) compared to the rest of the papers.

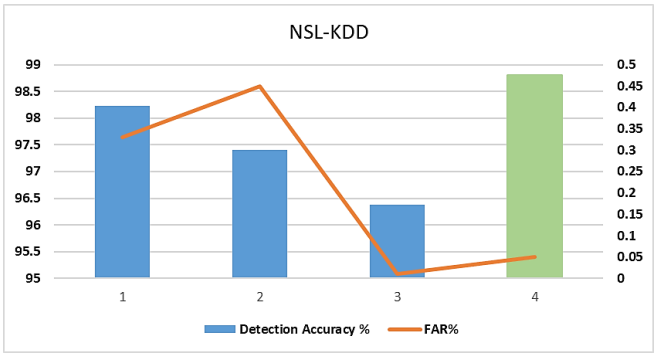

Fig 13. Comparison of the proposed ML RF-Based algorithm

VI. CONCLUSION

The purpose of this paper is to illustrate the capabilities of machine learning techniques for developing Cloud computing cybersecurity. Cloud computing technologies have now become indispensable in everyday life. But there are some challenges that hinder Cloud computing, and security is one of them.

In this paper, the Random forest algorithm was used to analyze and detect DDoS attacks. Performance evaluation was performed based on accurate detection accuracy, false alarm rate, accuracy, and recall measurements. The model was implemented by the Weka ML tool. To experiment with the proposed model, the NSL-KDD and ISCX datasets were used.

AUTHORS CONTRIBUTION

We encourage authors to submit an author statement outlining their individual contributions to the paper using the relevant roles:

1- Conception or design of the work (25% for Ghada Mohamed, Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

2- Data collection and tools (100% for Ghada Mohamed)

3- Data analysis and interpretation (100% for Ghada Mohamed)

4- Investigation (25% for Ghada Mohamed, Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

5- Methodology (25% for Ghada Mohamed, Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

6- Project administration (25% for Ghada Mohamed, Ehab

H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

7- Software (100% for Ghada Mohamed)

8- Supervision (100% for Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

9- Drafting the article (100% for Ghada Mohamed)

10- Critical revision of the article. (25% for Ghada Mohamed, Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

11- Final approval of the version to be published (25% for Ghada Mohamed, Ehab H. Abdelhay, Ibrahim Yasser and Mohamed A. Mohamed)

The corresponding author is responsible for ensuring that the descriptions are accurate and agreed upon by all authors.

.

FUNDING STATEMENT:

The author did not receive any financial support of the research authorship and publication of this article.

DECLARATION OF CONFLICTING INTERESTS STATEMENT:

The author declared that there are no potential conflicts of interest with respect to the research authorship or publication of this article.

REFERENCES

[1] P. Mell and T. Grance, "The NIST Definition of Cloud computing Recommendations of the National Institute of Standards and Technology." doi: 10.6028/NIST.SP.800-145.

[2] A. P. Achilleos *et al.*, "The cloud application modelling and execution language," *Journal of Cloud computing*, vol. 8, no. 1, p. 20, Dec. 2019, doi: 10.1186/s13677-019-0138-7.

[3] T. D. Quilichini, E. Grienenberger, and C. J. Douglas, "The biosynthesis, composition and assembly of the outer pollen wall: A tough case to crack," *Phytochemistry*, vol. 113. Elsevier Ltd, pp. 170–182, May 17, 2015, doi: 10.1016/j.phytochem.2014.05.002.

[4] "Distributed denial of service attacks | IEEE Conference Publication | IEEE Xplore," Accessed: Mar. 25, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/886455.

[5] "Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper - Cisco." https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html (accessed Mar. 26, 2021).

[6] "No summer vacation: DDOS attacks tripled year-on-year in Q2 2020 | Kaspersky." https://www.kaspersky.com/about/press-releases/2020_no-summer-vacation-ddos-attacks-tripled-year-on-year-in-q2-2020 (accessed Mar. 26, 2021).

[7] "A matter of profit: DDoS attacks in Q4 2020 dropped by a third compared to Q3, as cryptomining is on the rise | Kaspersky." https://www.kaspersky.com/about/press-releases/2021_a-matter-of-profit-ddos-attacks-in-q4-2020-dropped-by-a-third-compared-to-q3-as-cryptomining-is-on-the-rise (accessed Mar. 26, 2021).

[8] "Proactive or Reactive, Which is the Better Method for DDoS Defence?" https://hackercombat.com/proactive-or-reactive-which-is-the-better-method-for-ddos-defence/ (accessed Oct. 12, 2021).

[9] A. Aleroud and G. Karabatis, "A contextual anomaly detection approach to discover zero-day attacks," in *Proceedings of the 2012 ASE International Conference on Cyber Security, CyberSecurity 2012*, 2012, pp. 40–45, doi: 10.1109/CyberSecurity.2012.12.

[10] B. A. Khalaf, S. A. Mostafa, A. Mustapha, M. A. Mohammed, and W. M. Abduallah, "Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods," *IEEE Access*, vol. 7, pp. 51691–51713, 2019, doi: 10.1109/ACCESS.2019.2908998.

[11] S. Hosseini and M. Azizi, "The hybrid technique for DDoS detection with supervised learning algorithms," *Computer Networks*, vol. 158, pp. 35–45, Jul. 2019, doi: 10.1016/J.COMNET.2019.04.027.

[12] A. R. Wani, Q. P. Rana, U. Saxena, and N. Pandey, "Analysis and Detection of DDoS Attacks on Cloud computing Environment using Machine Learning Techniques," *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, pp. 870–875, Apr. 2019, doi: 10.1109/AICAI.2019.8701238.

[13] A. Alsirhani, S. Sampalli, and P. Bodorik, "DDoS Detection System: Using a Set of Classification Algorithms Controlled by Fuzzy Logic System in Apache Spark," *IEEE Transactions on Network and Service Management*, 2019, doi: 10.1109/TNSM.2019.2929425.

[14] V. Sharma, V. Verma, and A. Sharma, "Detection of DDoS Attacks Using Machine Learning in Cloud computing," in *Communications in Computer and Information Science*, Jun. 2019, vol. 1076, pp. 260–273, doi: 10.1007/978-981-15-0111-1_24.

[15] P. Shamsolmoali and M. Zareapoor, "Statistical-based filtering system against DDOS attacks in Cloud computing," in *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, Nov. 2014, pp. 1234–1239, doi: 10.1109/ICACCI.2014.6968282.

[16] P. Xiao, W. Qu, H. Qi, and Z. Li, "Detecting DDoS attacks against data center with correlation analysis," *Computer Communications*, vol. 67, pp. 66–74, Aug. 2015, doi: 10.1016/j.comcom.2015.06.012.

[17] F. Kuang *et al.*, "A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection," *Soft Comput*, vol. 19, pp. 1187–1199, 2015, doi: 10.1007/s00500-014-1332-7.

[18] M. Zekri, S. El Kafhali, N. Aboutabit, and Y. Saadi, "DDoS attack detection using machine learning techniques in Cloud computing environments," in *Proceedings of 2017 International Conference of Cloud computing Technologies and Applications, CloudTech 2017*, Feb. 2018, vol. 2018-January, pp. 1–7, doi: 10.1109/CloudTech.2017.8284731.

[19] G. S. Kushwah and S. T. Ali, "Detecting DDoS attacks in Cloud computing using ANN and black hole optimization," in *2nd International Conference on Telecommunication and Networks, TEL-NET 2017*, Apr. 2018, vol. 2018-January, pp. 1–5, doi: 10.1109/TEL-NET.2017.8343555.

[20] G. S. Kushwah and V. Ranga, "Voting extreme learning machine based distributed denial of service attack detection in Cloud computing," *Journal of Information Security and Applications*, vol. 53, 2020, doi: 10.1016/j.jisa.2020.102532.

[21] S. Behal, S. Bhagat, S. State, and T. Campus, "Detection of DDoS Attacks using Weka Tool : A Case Study Detection of DDoS Attacks using Weka Tool : A Case Study," no. December 2017, 2018.

[22] K. Scarfone and P. Mell, "Special Publication 800-94 Guide to Intrusion Detection and Prevention Systems (IDPS) Recommendations of the National Institute of Standards and Technology," 2002, doi: 10.6028/NIST.SP.800-94.

[23] A. Bakshi and B. Yogesh, "Securing cloud from DDOS attacks using intrusion detection system in virtual machine," in *2nd International Conference on Communication Software and Networks, ICCSN 2010*, 2010, pp. 260–264, doi: 10.1109/ICCSN.2010.56.

[24] C. C. Lo, C. C. Huang, and J. Ku, "A cooperative intrusion detection system framework for Cloud computing networks," in *Proceedings of the International Conference on Parallel Processing Workshops*, 2010, pp. 280–284, doi: 10.1109/ICPPW.2010.46.

[25] B. M. E. Oztemel and L. Baton Rouge, "DETECTION OF INTRUSION THROUGH," 2019.

[26] D. Wu, J. Li, S. K. Das, J. Wu, Y. Ji, and Z. Li, "A novel distributed denial-of-service attack detection scheme for software defined networking environments," in *IEEE International Conference on Communications*, Jul. 2018, vol. 2018-May, doi: 10.1109/ICC.2018.8422448.

[27] Z. He, T. Zhang, and R. B. Lee, "Machine Learning Based DDoS Attack Detection from Source Side in Cloud," in *Proceedings - 4th IEEE International Conference on Cyber Security and Cloud computing, CSCloud 2017 and 3rd IEEE International Conference of Scalable and Smart Cloud, SSC 2017*, Jul. 2017, pp. 114–120, doi: 10.1109/CSCloud.2017.58.

[28] M. Suresh and R. Anitha, "Evaluating machine learning algorithms for detecting DDoS attacks," in *Communications in Computer and Information Science*, 2011, vol. 196 CCIS, pp. 441–452, doi: 10.1007/978-3-642-22540-6_42.

[29] W. Meng, W. Li, and L. F. Kwok, "Design of intelligent KNN-based alarm filter using knowledge-based alert verification in intrusion detection," *Security and Communication Networks*, vol. 8, no. 18, pp. 3883–3895, Dec. 2015, doi: 10.1002/sec.1307.

[30] C. Vens, "Random forest," *Encyclopedia of Systems Biology*, pp. 1812–1813, 2013, doi: 10.1007/978-1-4419-9863-7_612.
[31] Q. Niyaz, W. Sun, and A. Y. Javaid, "A Deep Learning Based DDoS Detection System in Software-Defined Networking (SDN)," 2017, doi: 10.4108/eai.28-12-2017.153515.
[32] U. Ravale, N. Marathe, and P. Padiya, "Feature selection based hybrid anomaly intrusion detection system using K Means and RBF kernel function," in *Procedia Computer Science*, 2015, vol. 45, no. C, pp. 428–435, doi: 10.1016/j.procs.2015.03.174.
[33] L. Breiman, *Random forests*, vol. 45, no. 1. Springer, 2001.
[34] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." https://www.cs.waikato.ac.nz/ml/weka/ (accessed Mar. 31, 2021).
[35] "NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/nsl.html (accessed Apr. 01, 2021).
[36] "IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/ids-2017.html (accessed Apr. 01, 2021).
[37] M. Idhammad, K. Afdel, and M. Belouch, "Semi-supervised machine learning approach for DDoS detection," *Applied Intelligence 2018 48:10*, vol. 48, no. 10, pp. 3193–3208, Feb. 2018, doi: 10.1007/S10489-018-1141-2.

Title Arabic:

تطوير تقنيات التعلم الآلي لتعزيز خوارزميات الأمن السيبراني

Arabic Abstract:

في للوقت الحاضر يلعب الأمن السيبراني دورًا مهمًا في مجال تكنولوجيا المعلومات (IT)، هذا وقد أصبح تأمين المعلومات أحد أكبر التحديات التي تواجه مُجتمع المعلومات، خاصة مع التطورات الأخيرة التي تشهدها مجالات الحوسبة السحابية.

إذ أدت إلى اتجاه جديد متزايد للهجمات الإلكترونية، التي يُعد هجوم رفض الخدمة الموزعة (DDoS)، أحد أخطر ما تواجهه الحوسبة السحابية. إذ يجعل هذا الهجوم الخدمات السحابية غير قابلة مُتاحة للمستخدمين النهائيين من خلال استنفاد موارد النظام، مما يؤدي إلى خسائر فادحة، لذلك فإن تطوير حلول دفاعية ضد هذه الهجمات أصبح ضروري للتوسع في استخدام تكنولوجيا الحوسبة السحابية.

هذا ويُعتبر استخدام التعلم الآلي (ML) إحدى طرق تأمين الحواسيب السحابية. إذ يتم استخدام تقنيات ML بطرق مختلفة لاكتشاف الهجمات والثغرات الأمنية على السحابة.

يُحاول هذا البحث، اقتراح نظام لاكتشاف هجمات DDoS في بيئة الحوسبة السحابية. حيث جرى بناء النظام المقترح باستخدام خوارزمية Random forest (RF)، والتعلم الآلي الخاضع للإشراف. في هذا العمل، تم تقييم النظام المقترح باستخدام مجموعة بيانات NSL-KDD ومجموعة بيانات كشف التسلل ISCX (ID).

وقد أظهرت نتائج التجربة أن الطريقة المقترحة يمكن أن تحقق أداءً جيدًا، والذي يمتلك مزايا عند مقارنته بالطرق الأخرى الموجودة من حيث الدقة ومعدل الكشف والمعدل الإيجابي الخاطئ المنخفض.