Received January 23, 2022, accepted February 15, 2022, date of publication February 17, 2022, date of current version March 2, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152577

A Machine Learning-Based Classification and Prediction Technique for DDoS Attacks

ISMAIL[1], MUHAMMAD ISMAIL MOHMAND[1], HAMEED HUSSAIN[2], AYAZ ALI KHAN[3], UBAID ULLAH[1], MUHAMMAD ZAKARYA[4], (Senior Member, IEEE), AFTAB AHMED[4], MUSHTAQ RAZA[4], IZAZ UR RAHMAN[4], AND MUHAMMAD HALEEM[5]

[1]Department of Computer Science, Brains Institute Peshawar, Peshawar 47301, Pakistan
[2]Department of Computer Science, University of Buner, Buner 19281, Pakistan
[3]Department of Computer Science, University of Lakki Marwat, Lakki Marwat 28420, Pakistan
[4]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan
[5]Department of Computer Science, Kardan University, Kabul 1001, Afghanistan

Corresponding author: Muhammad Haleem (m.haleem@kardan.edu.af)

This work was supported in part by the University of Peshawar, Pakistan; and in part by Kardan University, Kabul.

Close to my reseach!
- 2 classifier used but don't say why?
- Can use different data!

ABSTRACT Distributed network attacks are referred to, usually, as Distributed Denial of Service (DDoS) attacks. These attacks take advantage of specific limitations that apply to any arrangement asset, such as the framework of the authorized organization's site. In the existing research study, the author worked on an old KDD dataset. It is necessary to work with the latest dataset to identify the current state of DDoS attacks. This paper, used a machine learning approach for DDoS attack types classification and prediction. For this purpose, used Random Forest and XGBoost classification algorithms. To access the research proposed a complete framework for DDoS attacks prediction. For the proposed work, the UNWS-np-15 dataset was extracted from the GitHub repository and Python was used as a simulator. After applying the machine learning models, we generated a confusion matrix for identification of the model performance. In the first classification, the results showed that both Precision (PR) and Recall (RE) are ~89% for the Random Forest algorithm. The average Accuracy (AC) of our proposed model is ~89% which is superb and enough good. In the second classification, the results showed that both Precision (PR) and Recall (RE) are approximately 90% for the XGBoost algorithm. The average Accuracy (AC) of our suggested model is ~90%. By comparing our work to the existing research works, the accuracy of the defect determination was significantly improved which is approximately 85% and 79%, respectively.

INDEX TERMS DDoS attacks, machine learning, random forest, XGBoost, prediction.

I. INTRODUCTION

Distributed network attacks are referred to, usually, as Distributed Denial of Service (DDoS) attacks. These attacks take advantage of specific limitations that apply to any arrangement asset, such as the framework of the authorized organization's website. A DDoS attack sends different requests (with IP spoofing) to the target web assets to exceed the site's ability to handle various requests, at a given time, and make the site unable to operate effectively and efficiently – even for the legitimate users of the network. Typically, the target of various DDoS attacks are web applications and business websites; and the attacker may have different goals [1], [2]. Some common types of the DDoS attacks are shown in Figure 1. We give brief description of each attack in Section I-A.

The Internet of Things (IoT) implies the arrangement of interconnected, web-related objects that can collect and interchange information through remote organizations without manual intervention [3]. The "Things" can simply be related clinical tools, bio-chip transponders, solar panels, and related vehicles with sensors that can warn the driver of numerous potential problems [4], or any article with sensors that can collect and move information in the organization. Artificial intelligence (AI) is a small tool that transforms information into data. In the past 50 years (approximately), information has had an impact on users privacy and security. Except for the possibility of researching it and finding the examples hidden in it, the amount of information is negligible. Artificial intelligence technology is usually used to find important

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

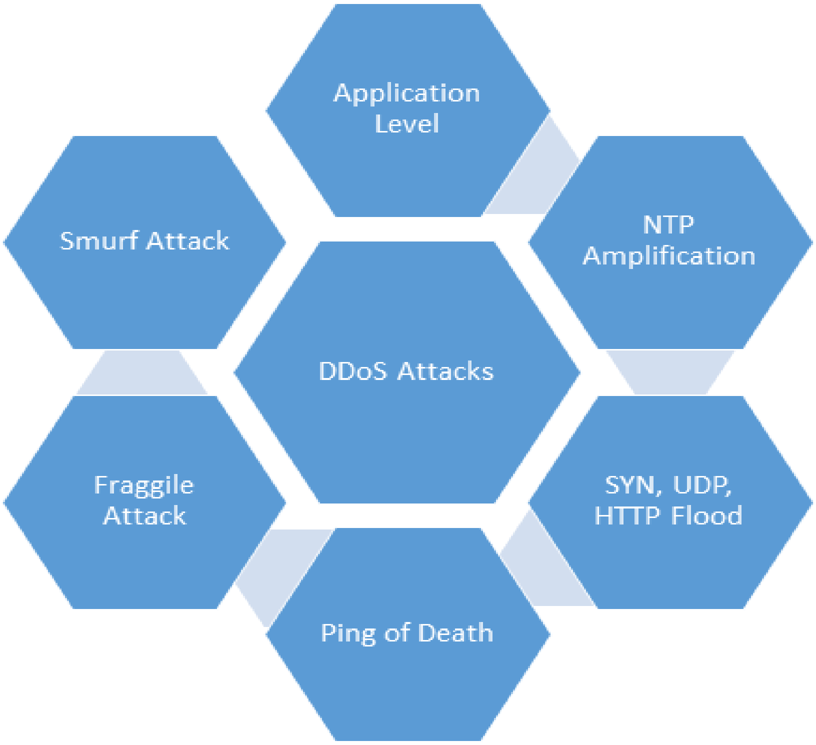


FIGURE 1. Various types of DDoS attacks.

secret examples in complex information, and this work will try to find them in some way. Mysterious examples and data about a problem can be used to predict future events and play a wide range of complex dynamics.

There were different approaches proposed for DDoS attack classification and prevention. In [4] deep learning models are proposed for intrusion detection. The dataset was UNSW-nb15 and the models were Convention neural network (CNN), BAT-MC, BAT, and Recurrent neural network. The overall model's performance was very good. They found CNN best for the proposal. The average accuracy was 79%. In paper [5] authors proposed a hybrid model deep learning model for intrusion detection. They combined two deep learning for the classification of CNN and LSTM from the RNN model. The dataset was used in this work is KDD. They found an 85.14% average accuracy for the proposed. However, up to our knowledge different deep learning models are used for DDoS attacks. Similarly, they used the same KDD dataset from the UCI repository in research. In finally all authors found the same results 85%.

A. TYPES OF THE DDoS ATTACKS

The SYN Flood abuses the shortcomings in TCP association packets, which is called a three-way handshake. The host obtains a synchronization (SYN) message to initiate a "handshake". The user recognizes the message by sending an acknowledgment (ACK) [1] banner to the underlying host, and the association will be closed at this time. Nevertheless, in the SYN flood, absurd messages are still sent, and the association will not be closed, thus turning off the help [2]. The UDP flood is a kind of denial-of-service attacks in which numerous User Datagram Protocol (UDP) packets are forwarded to a computer server (targeted) in order to exhaust that server's capability to execute and reply requests. Moreover, the firewall that is used to protect the server (targeted) may also become overwhelmed as a consequence of the UDP flooding attacks, which subsequently results in a denial of service (DoS) to legal and legitimate traffic flows and users. The HTTP flood is an attack type in which the attacker seemingly exploits even the legitimate HTTP GET or POST requests in order to attack a web application or a web server. The HTTP flood attacks frequently use a botnet – a group of Internet-connected computers.

Similarly, a Death Ping controls IP conventions by sending malicious pings to the framework. This is a famous DDoS attack in last two decades, but now this attack is not much popular. The Smurf attack uses a malware program called smurf to abuse the Internet Protocol (IP) and Internet Control Message Protocol (ICMP). It will imitate the IP address and use ICMP to ping the IP address of the specified organization. The Fraggle attack is a type of DDoS attack which uses a large amount of UDP traffic to transmit to the transmission organization of the switch. This is like a Smurf attack using UDP instead of ICMP [6]. Besides these, application-level attacks intentionally exploit weaknesses in an application. The target of this attack is to gain control of the application by bypassing normal access controls. In an NTP amplification attack, the attacker abuses a functionality of the Network Time Protocol (NTP) server in order to devastate a targeted server or network with a large quantity of User Datagram Protocol (UDP) traffic; and as a result this rendering the destination infrastructure unreachable to regular legitimate users traffic [7].

B. MOTIVATION FOR MACHINE LEARNING

In paper [2] authors proposed different algorithms for classification because the current algorithms have a lot of flaws and drawbacks. First, they cannot work with irrelevant values and feature engineering because the confusion matrix results are not accurate. Some labeled results are zero that means algorithms do not work well. So, this is important to train the model precisely. Another problem is that some results show (Null) that means missing values also included in data that was not computed. Similarly, we need to justify existing algorithms with an advanced algorithm to find out the fastest and sufficient model. They also showed that random forest is not better than the KNN model because the result is less for the KNN model.

In [5], CNN and RNN both are two different algorithms that can be used for different purposes. For example, CNN is used for feature extraction and RNN is used for regression in time series data utilization. The authors used the CNN and RNN [4] model for intrusion detection. However, this is a very long and time-consuming process. Therefore, it is very important to perform advanced machine learning techniques to model optimization that train the best model for highly accurate work. Here, in this paper, intrusion detection is a classification problem. Therefore, it is a very serious problem to handle these implemented algorithms. In the last one, no such methodology is used for data mining to improve the quality of data.

Among the machine learning techniques, random forest and XGBoost both are powerful supervised learning models. Both are applicable and used for classification problems. The random forest algorithm is approximately 100 times faster than other algorithms and best working for classification problems. This should be noted that the XGBoost is the ideal algorithm of machine learning because it is approximately 100 times faster than the random forest and best for forbid data analysis. Both are simple and faster than other algorithm in terms of execution times.

C. CONTRIBUTIONS

To further improve the accuracy and effectiveness, we propose an approach using different machine learning classifiers with model optimization. Also, it is important to perform machine learning data mining techniques to improve data quality. There are many research works being proposed for DDoS attacks detection and prevention; however, the main problem is that all the researcher worked with old datasets, in particular, KDDCUP [1]. Therefore, this is very important to work with the latest datasets where we can examine the current state of the DDoS attacks detection and prevention. The main contributions of the research conducted in this paper are three-fold.

- To design a step-by-step framework for data utilization.
- To design and develop an approach using supervised machine learning classifiers for DDoS attack detection based on different techniques.
- To evaluate and validate the proposed work and then compare it with existing studies in the literature.

The remainder of this paper is organized as follows. In Section II, we introduced the related work. In Section III, we present the proposed methodology. In Section IV, we conduct experiments on real-world datasets and compare performance with some existing baselines. Finally, we conclude the paper along with directions for future research and investigation in Section V.

II. RELATED WORKS

In the literature review section we briefly explained all the related model and the closest rival to our proposed study. We studied the latest research papers of the past two years for this research work and also Gozde Karatas *et al.* [2] proposed a machine learning approach for attacks classification. They used different machine learning algorithms and found that the KNN model is best for classification as compared to other research work. Nuno Martins *et al.* [1] proposed intrusion detection using machine learning approaches. They used the KDD dataset which is available on the UCI repository. They performed different supervised models to balance un classification algorithm for better performance. In this work, a comparative study was proposed by the use of different classification algorithms and found good results in their work. Laurens D'hooge *et al.* [6] proposed a systematic review for malware detection using machine learning models. They compared different malware datasets from online resources as well as approaches for the dataset. They found that machine learning supervised models are very effective for malware detection to make a better decision in less time.

Xianwei Gao *et al.* [7] proposed a comparative work for network traffic classification. They used machine learning classifiers for intrusion detection. The dataset is taken is CICIDS and KDD from the UCI repository. They found support vector machine SVM one of the best algorithms as compare to others. Tongtong Su *et al.* [3] proposed adaptive learning for intrusion detection. They used the KDD dataset from an online repository. These models are Dtree, R-forest, and KNN classifiers. In this study, the authors found that Dtree and ensemble models are good for classification results. The overall accuracy of the proposed work is 85%.

Kaiyuan Jiang *et al.* [4] proposed deep learning models for intrusion detection. The dataset is KDD and the models are Convention neural network (CNN), BAT-MC, BAT, and Recurrent neural network. The overall model's performance was very good. They found CNN as best for learning. The accuracy is improved from 82% to 85%.

Arun Nagaraja *et al.* [5] proposed a hybrid model deep learning model for intrusion detection. They combined two deep learning models for the classification of CNN+ LSTM from the RNN model. The dataset was used in this work is KDD. They found an 85.14% average accuracy for the proposed. Yanqing Yang *et al.* [8] proposed a similarity-based approach for anomaly detection using machine learning. They used k mean cluster model for feature similarity detection and naïve Bayes model used for classification.

Hui Jiang *et al.* [4] used an auto-encoder for labels and performed deep learning classification models on the KDD dataset. They found an 85% average accuracy for the proposed model [9]. SANA ULLAH JAN *et al.* [10] proposed a PSO-Xgboost model because it is higher than the overall classification accuracy alternative models, e.g. Xgboost, Random-Forest, Bagging, and Adaboost. First, establish a classification model based on Xgboost, and then use the adaptive search PSO optimal structure Xgboost. NSL-KDD, reference dataset used for the proposed model evaluation. Our results show that, PSO-Xgboost model of precision, recall, and macro-average average accuracy, especially in determining the U2R and R2L attacks. This work also provides an experimental basis for the application group NIDS in intelligence.

Maede Zolanvari *et al.* [11] proposed a recurrent neural network model for classification intrusion detection. They compared other deep learning models with RNN. Finally, they found RNN is the best model for intrusion detection by using the KDD dataset. Yijing Chen *et al.* [12] proposed a domain that generates an algorithm for botnet classification. It was a multiple classification problem. They used advanced deep learning LSTM for multiple classification problems. They found good results with 89% average accuracy for the proposed work.

Larriva-Novo *et al.* [13] proposed two benchmark datasets, especially UGR16 and UNSW-NB15, and the most used dataset KDD99 were used for evaluation. The pre-processing strategy is evaluated based on scalar and standardization capabilities. These pre-processing models are applied through various attribute arrangements. These attributes depend on the classification of the four sets of highlights: basic associated highlights, content quality, fact attributes, and finally the creation of highlights based on traffic and traffic quality based on associated titles Collection. The goal of this inspection is to evaluate this arrangement by using different information pre-processing methods to obtain the most accurate model. Our proposition shows that by applying the order of organizing traffic and some pre-processing strategies, the accuracy can be improved by up to 45%. The pre-processing of a specific quality set takes into account more prominent accuracy, allowing AI calculations to effectively group these boundaries identified as potential attacks.

Zeeshan Ahmad *et al.* [14] proposed a scientific classification approach, which depends on the well-known ML and DL processes included in the planning network-based IDS (NIDS) framework. By examining the quality and certain limitations of the proposed arrangements, an extensive review of the new clauses based on NIDS was conducted. By then, regarding the proposed technology, evaluation measurement, and dataset selection, the ongoing patterns and progress of NIDS based on ML and DL are given. Taking advantage of the deficiencies of the proposed technology, in this paper, we put forward different exploration challenges and give suggestions.

Muhammad Aamir *et al.* [15] proposed AI calculations were prepared and tried on the latest distributed benchmark dataset (CICIDS2017) to distinguish the best performance calculations on information, which contains the latest vectors of port checks and DDoS attacks. The permutation results show that every variation of isolation check and support vector machine (SVM) can provide high test accuracy, for example, more than 90%. According to the abstract scoring criteria cited in this article, 9 calculations from a bunch of AI tests received the most noteworthy score (highest) because they gave more than 85% representation (test) accuracy in 22 absolute calculations. In addition, this related investigation was also conducted to note that through the k-fold cross approval, the area under the curve (AUC) check of the receiver operating characteristic (ROC) curve, and the use of principal component analysis (PCA) for size reduction in preparation for AI execution model. When considering such late attacks, it was found that many checks on different AI calculations of the CICIDS2017 datasets were not sufficient for port checks and DDoS attacks.

Kwak *et al.* [16], proposed a video steganography botnet model. In addition, they plan to use another video steganography technology based on the payload method (DECM: Frequency Division Embedded Component Method), which can use two open devices VirtualDub and Stegano to implant significantly more privileges than existing tools information. They show that proposed model can be performed in the Telegram SNS courier, and compared proposed model and DECM with the current image steganography-based botnets and methods in terms of the effectiveness and imperceptibility [17].

Zahid Akhtar *et al.* [18] proposed a concise overview of malware, followed by a summary of different inspection challenges. This is a hypothetical point of view article that needs to be improved. Duy-Cat and Can. et al [19] became familiar with a model that can identify and arrange distributed denial of service attacks that rely on the use of the proposed program including selected segments of neural tissue. The experimental results of the CIC-DDoS 2019 dataset show that our proposed model beats other AI-based models to a large extent. We also studied the selection of weighted misfortune and the choice of pivotal misfortune in taking care of class embarrassment [20].

Qiumei Cheng *et al.* [21] proposed a novel in-depth binding review (OFDPI) method with OpenFlow function in SDN using AI computing. OFDPI supports in-depth bundling inspection of the two decoded packages. The method of traffic and scrambled traffic is to prepare two dual classifiers respectively. In addition, OFDPI can test suspicious packages using bundling windows that depend on immediate expectations. We use real-world datasets to evaluate OFDPI's exhibitions on the Ryu SDN regulator and Mininet stage. As with sufficient overhead, OFDPI achieves a fairly high recognition accuracy for encoding traffic and decoding traffic. Stephen Kahara Wanjau *et al.* [22] a complete SSH-Brute power network attack discovery system is proposed, which relies on a standardized deep learning calculation, that is, a convolutional neural network. The model representations were compared, and experimental results were obtained from five old-style AI calculations, including logistic regression (LR), decision trees (DT), naive Bayes (NB), k-nearest neighbours (KNN), and support vector machines (SVM). In particular, four standard measurements metrics are often used, namely: (i) accuracy, (ii) precision, (iii) recall, and (iv) F measurement. The results demonstrate that model based on the CNN approach is better than the conventional AI technology. The accuracy is 94.3%, the accuracy is 92.5%, the review speed is 97.8%, and the F1 score is 91.8%. This is our ability to recognize the powerful features of SSH-Brute attacks [23], [24].

III. PROPOSED MODEL

In this research, we design a framework for the DDoS attack classification and prediction based on the existing dataset that used machine learning methods. This framework involves the following main steps.

1) The first step involves the selection of dataset for utilization.
2) The second step involves the selection of tools and language.
3) The third step involves data pre-processing techniques to handle irrelevant data from the dataset. In the fourth step feature extraction and label.
4) Encoding is performed to convert symbolical data into numerical data.
5) In the fifth step, the data splitting is performed into a train and test set for the model. In this step, we build and train our proposed model. However, model optimization is also performed on the trained model in terms of kernel scaling and kernel hyper-parameter tuning to improve model efficiency. When the model optimizes then we will generate output results from the model.

The main contribution is to generate the best model for data utilization, as well as, model optimization; and which performs best for model learning. After getting the results, we performed performance measures in terms of precision, recall, and f1 score. In this research work, we used two well-known supervised learning models which are: (i) Random Forest Classifier; and (ii) XGBoost Classifier. The architecture and data flow diagram of the proposed method is shown in Figure 2.

TABLE 1. UNSW-nb15 dataset.

Total Rows	Total Columns
82,332	45

IV. RESULTS AND DISCUSSION

This section contains all the obtained results of our proposed models. All the results are shown step by step in the form of figures, as well as, results explanation. In Section IV-N, we briefly describe and evaluate the performance of our suggested model with several closest rivals and existing research studies.

A. DATASET

We selected the UNSW-nb15 dataset from GitHub[1] that contains features' data about the DDoS attacks. This dataset is provided by the Australian Centre for Cyber Security (ACCS) [25]. Table 1 shows the total numbers of rows and columns in the dataset. The dataset consists of different features about the DDoS attacks including an ID number, Proto which presents medium of the network, label of the attacks, and attacks' cat which presents the severity of the DDoS attacks.

B. LANGUAGE AND TOOL

Python language is considered a suitable programming language both for simulations and real-world programming. It is considered the most powerful high level language for model learning [25]. Moreover, Python is also open-source, portable, and simple to use [25]. We used a jupyter notebook as a tool. This tool is open-source and browser-based which has evolved to become a robust tool for researchers to share documentation and code. This tool functions as a virtual lab notebook [26].

C. IMPORT LIBRARIES

It is the first step to import some important function for reading information in tabular in our language. In order to import the data, we used different Python functions and procedures which are built-in in this language. Moreover, this is very important in data reading from a specific directory to the programming language [27].

D. DATA PRE-PROCESSING

It is very important and time-consuming part of data analysis. Where we are going to clean the information from irrelevant data and convert it to quality information. For this step we are using statistical techniques to clean data and replace those values which are not important in our experimental analysis. This is essential of every data analysis for the initial phase examination. After it, we will be able to convert information into reliable form. To investigate the value and information graphical form. Here, we used the heat-map for illustrating

[1] https://github.com/naviprem/ids-deep-learning/tree/master/datasets

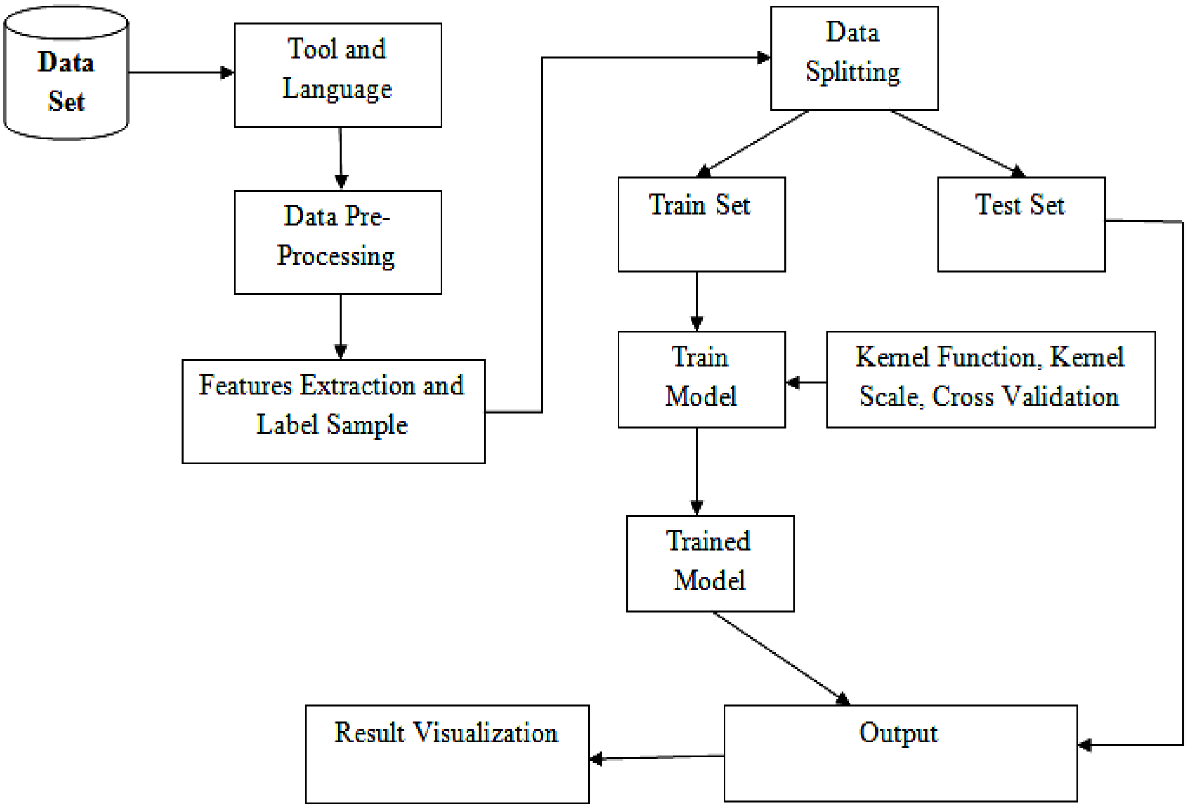
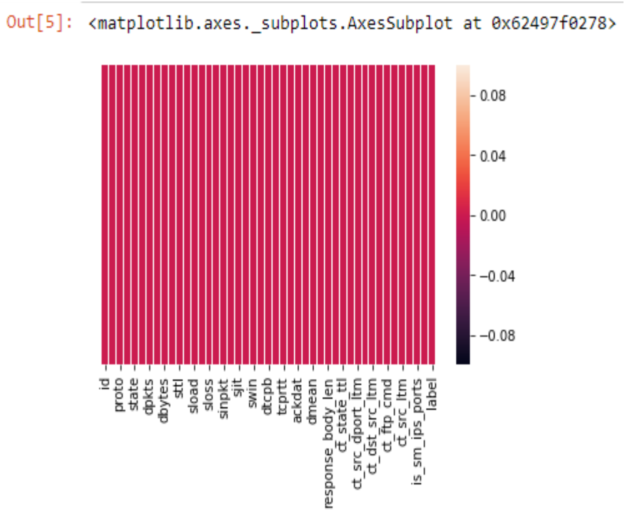
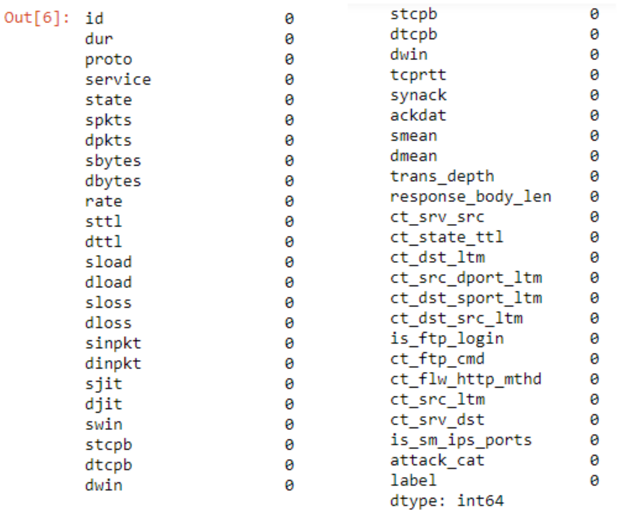


FIGURE 2. Data flow chart for the proposed machine learning based DDoD attack prediction technique.



```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x62497f0278>
```

FIGURE 3. Heat-map for missing values.



```
Out[6]: id                   0        stcpb                0
        dur                  0        dtcpb                0
        proto                0        dwin                 0
        service              0        tcprtt               0
        state                0        synack               0
        spkts                0        ackdat               0
        dpkts                0        smean                0
        sbytes               0        dmean                0
        dbytes               0        trans_depth          0
        rate                 0        response_body_len    0
        sttl                 0        ct_srv_src           0
        dttl                 0        ct_state_ttl         0
        sload                0        ct_dst_ltm           0
        dload                0        ct_src_dport_ltm     0
        sloss                0        ct_dst_sport_ltm     0
        dloss                0        ct_dst_src_ltm       0
        sinpkt               0        is_ftp_login         0
        dinpkt               0        ct_ftp_cmd           0
        sjit                 0        ct_flw_http_mthd     0
        djit                 0        ct_src_ltm           0
        swin                 0        ct_srv_dst           0
        stcpb                0        is_sm_ips_ports      0
        dtcpb                0        attack_cat           0
        dwin                 0        label                0
                                      dtype: int64
```

FIGURE 4. Heat-map missing values report.

the missing values, graphically. The below Figure 3 shows results of missing values, graphically. The results show that there are no irrelevant values that needs to remove. The below Figure 4 shows the complete results and outcomes of the experiments [28].

Figure 4 shows the results when all datasets are clean. In data pre-processing phase, we also observed and identified that our datasets are almost clean.

E. LABEL ENCODING

Not computer works with letter information, because computers can understand on and off. Also, in this case, our computer algorithms cannot understand the letter form of our information. Therefore, it is important to convert this information into digital form so that our proposed model can understand it. The tag encoder is a machine learning process,

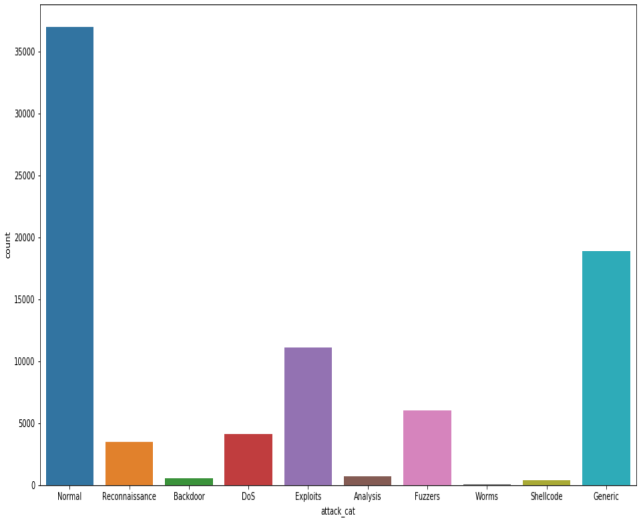


FIGURE 5. Attacks.

and we can transform it into the form we expect. The image which given below is full presentation of our dataset which are converted to numerical form.

F. DATA VISUALIZATION

The present of data where the information will understandable in the form of image or diagram. It is important to understand easily the information. Here we will be applying advance library for data visualization. This is the initial step where we are selecting our target for the proposed algorithm. Also, this step is used for selecting the test class. This step is very important to understand data in a much better way. Through this method we were able to select our target class for classification.

The visualization of showed total number of Normal = 37,000, Generic = 18871, Exploits = 11,132, Fuzzers = 6,062, DoS = 4,089, Reconnaissance = 3,496, Analysis = 677, Backdoor = 583, Shellcode = 378, and Worms = 44 attacks. If we see then it is a multiple classification problem, and we used supervised machine learning model for this classification problem.

G. DATA SPLITTING

We divide the dataset into two different classes: (i) dependent; and (ii) independent. The dependent class is also called the target class. The independent classes are those classes which do not depend on other classes. Therefore, we split the dataset herein in training and testing datasets, for our proposed model. For data splitting, we can use the sklearn model selection library in order to train and test the dataset for evaluation.

H. FEATURE SCALING

All algorithms in artificial intelligence and machine learning employ input data to generate output results. The characteristics and features in this input dataset are in the form of structural columns. To operate well with the algorithm, all algorithms require data characteristics with certain characters. The basic aims of feature engineering are to provide an input dataset that is compatible with the criteria of machine learning and artificial intelligence models. As a result, we begin by converting all classified attributes into equivalent numerical labels. The second goal ad objective is to improve the performance of machine learning and artificial intelligence models.

1) DATA NORMALIZATION

Feature Element scaling is a method of standardizing the existence of autonomous elements in the information within an appropriate range. The scaling is performed in the process of information pre-processing to deal with the magnitude or value or unit of height changes. If the component scaling is not completed, then the AI calculation will weigh greater mass, greater magnitude, and treat the more general quality as a lower value, and rarely consider units with important values. There are two most ideal ways to apply the highlight zoom.

a: NORMALIZATION

The first is normalization, and the second is standardization. In normalization, your perception is taken away through all perceptual methods, and when the parts are separated by the standard deviation, at this point, the perception is scaled. The attached recipe is used for the normalization strategy in AI. This is a very effective strategy to readjust the quality to achieve nothing but the same difference with one.

$$Xnew = \frac{(Xi - Xmean)}{(standarddeviation)} \tag{1}$$

b: STANDARDIZATION

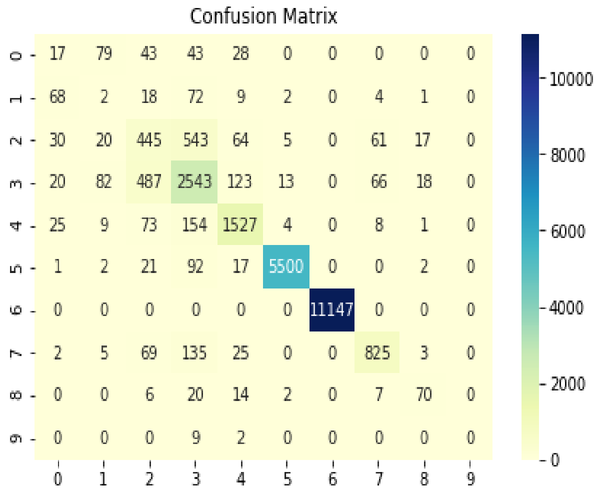
In standardization, divide your perception by the basis of all perceptions, and then, at this point, subtract the smallest perception from the most extreme perception, and then perform highlight scaling at that point. This process re-adjusts the components or perceptions, and spreads the value somewhere within the scope of nothingness.

$$Xnew = \frac{(X_i - min(X))}{(max(x) - min(x))} \tag{2}$$

In our proposed work, we use the standard scalar element scaling method for element scaling. This is due to the fact that it is the best strategy to use, most of the time, in including zooming.

I. SUPERVISED MODELS

Artificial intelligence (AI) is the use of computer reasoning and logic, which enables structures to recognize and further develop reality without explicit customization. Artificial intelligence revolves around the improvement of computer programs, which can acquire data and learn new information from it. Supervision is a group of calculation, which uses existing experiences, information, data [29], [30] to characterize and expect all the information indicators of the



	0	1	2	3	4	5	6	7	8	9
0	17	79	43	43	28	0	0	0	0	0
1	68	2	18	72	9	2	0	4	1	0
2	30	20	445	543	64	5	0	61	17	0
3	20	82	487	2543	123	13	0	66	18	0
4	25	9	73	154	1527	4	0	8	1	0
5	1	2	21	92	17	5500	0	0	2	0
6	0	0	0	0	0	0	11147	0	0	0
7	2	5	69	135	25	0	0	825	3	0
8	0	0	6	20	14	2	0	7	70	0
9	0	0	0	9	2	0	0	0	0	0

FIGURE 6. First confusion matrix of the random forest.

errand. In next section, we discuss our proposed model and the obtained results.

J. RANDOM FOREST CLASSIFIER

A random forest algorithm is a combination of the decision tree. It is very fast compared to other classifiers. Now after feature scaling the next step is the machine learning classification model. In our proposed work we used a random forest classification algorithm. The random forest, which is one of the most popular and powerful machine learning classification algorithms, is used for reaching a lot of decisions in the proposed model.

1) FIRST CONFUSION MATRIX

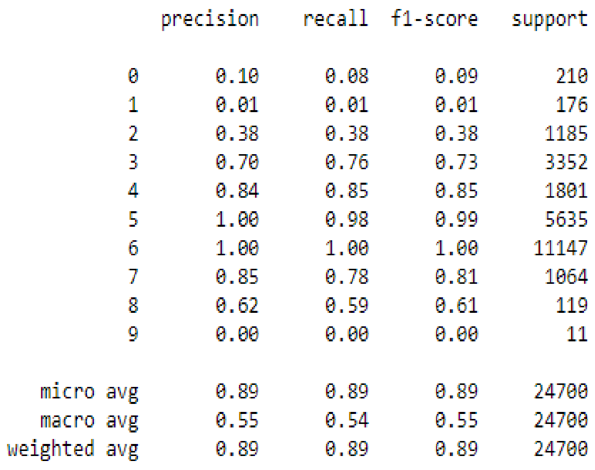
This method is used in the outline of AI group execution. Calculating the chaotic grid makes it easier for us to understand the correctness of the representation model and the types of errors it causes. It is used to calculate the accuracy of the representation, just like arranging true and prescient marks. They graphically display the classifier and its representation. The attached Figure 6 shows the disordered grid of our model.

Results: The give image is our metric from our model.

The confusion matrix denotes the overall number of actual and predicted labels for a particular algorithm. Similarly, the disordered dot matrix deals with the absolute number of actual marks and the expected names for arrangement. These real and expected names are a mixture of true positives, true negatives, false positives, and false negatives. Through these qualities, we will determine the accuracy of our model arrangements and expectations.

- TN solves the true negative: it is all the advantages of the precise anticipation of a negative case.
- FP resolves false positives: it is the sum of deviations from the basic expectations that have occurred as a positive.
- FN solves the false negatives: it is the sum of deviations from the basic expectations that appear negative.
- TP solves True-Positive: it is the sum of the exact expectation that an event is positive.

Therefore, this chaotic grid has a complete sixth mark, which is true certainty, true bad, false certainty, and false negative. After that, we used the above-mentioned chaotic grid to distinguish the proposed model exhibition. We use this chaotic dot matrix to determine the accuracy of the proposed model, thereby determining the accuracy of order reports and forecast results.



```
              precision    recall  f1-score   support

           0       0.10      0.08      0.09       210
           1       0.01      0.01      0.01       176
           2       0.38      0.38      0.38      1185
           3       0.70      0.76      0.73      3352
           4       0.84      0.85      0.85      1801
           5       1.00      0.98      0.99      5635
           6       1.00      1.00      1.00     11147
           7       0.85      0.78      0.81      1064
           8       0.62      0.59      0.61       119
           9       0.00      0.00      0.00        11

   micro avg       0.89      0.89      0.89     24700
   macro avg       0.55      0.54      0.55     24700
weighted avg       0.89      0.89      0.89     24700
```

FIGURE 7. First classification report of random forest.

TABLE 2. Performance measure.

AC (%)	PR (%)	RE (%)	F1 (%)
89	89	89	89

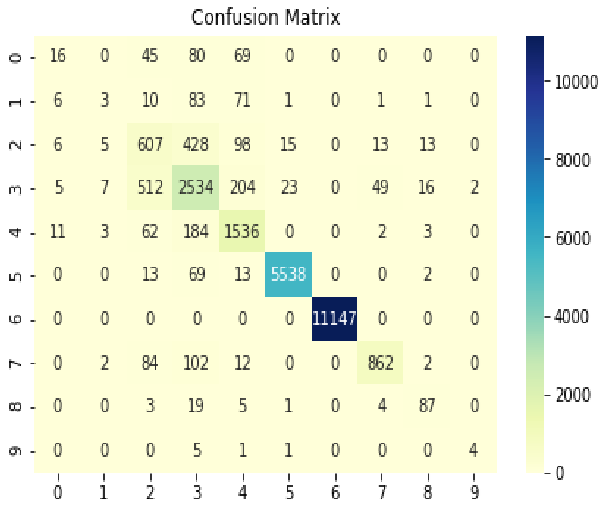
2) FIRST CLASSIFICATION RESULT

Currently we use the above disordered grid to complete our model exhibition. The following Figure 7 demonstrates that all representations of our suggested model and work rely on the factor of accuracy. Performance evaluation metrics, including the F metric (F1), average accuracy (AC), precision (PR), and recall (RE) rely on the chaotic network given above. Figure 7, as given below, illustrates the complete classification outcomes.

In the classification, our observation suggests that the precision (PR) factor is approximately 89% while the recall (RE) factor is also 89% accurate. Nevertheless, the average Accuracy (AC) of the suggested model is ∼89% that is believed wonderful and extremely awesome in the given setup. This should be noted that the average accuracy factor denotes the F1 score as ∼89%.

K. XGBOOST CLASSIFIER

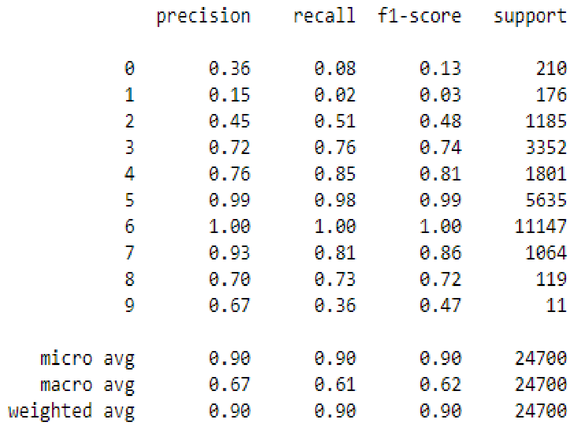
In the era of machine learning and artificial intelligence, the XGBoost algorithm is known as the queen by scientific and academic researchers. Most of the researchers considering as a weapon for big data utilization. This model also working



Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	16	0	45	80	69	0	0	0	0	0
1	6	3	10	83	71	1	0	1	1	0
2	6	5	607	428	98	15	0	13	13	0
3	5	7	512	2534	204	23	0	49	16	2
4	11	3	62	184	1536	0	0	2	3	0
5	0	0	13	69	13	5538	0	0	2	0
6	0	0	0	0	0	0	11147	0	0	0
7	0	2	84	102	12	0	0	862	2	0
8	0	0	3	19	5	1	0	4	87	0
9	0	0	0	5	1	1	0	0	0	4

FIGURE 8. Second confusion matrix of XGBoost.



```
              precision    recall  f1-score   support

           0       0.36      0.08      0.13       210
           1       0.15      0.02      0.03       176
           2       0.45      0.51      0.48      1185
           3       0.72      0.76      0.74      3352
           4       0.76      0.85      0.81      1801
           5       0.99      0.98      0.99      5635
           6       1.00      1.00      1.00     11147
           7       0.93      0.81      0.86      1064
           8       0.70      0.73      0.72       119
           9       0.67      0.36      0.47        11

   micro avg       0.90      0.90      0.90     24700
   macro avg       0.67      0.61      0.62     24700
weighted avg       0.90      0.90      0.90     24700
```

FIGURE 9. Second classification report of XGBoost.

on tree but 100 times faster than other models. The XGBoost learning model have very fast speed, scalability, efficiency and simplicity. This model is more reliable for big data. This model is working on probability. The confusion matrix and outcomes of the classification, are given below, for the XGBoost method.

1) SECOND CONFUSION MATRIX

Figure 8, as given below, illustrates the confusion matrix for the XGBoost model and evaluation of its performance.

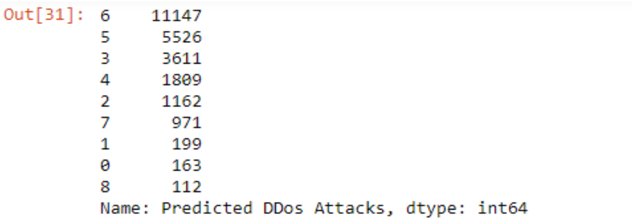
2) SECOND CLASSIFICATION RESULT

The algorithms' performance can be identified by the following results. Figure 9, as given below, demonstrates the complete classification outcomes.

In the classification, the obtained results demonstrated that the precision (PR) factor is approximately 90% while the recall (RE) is ∼90% accurate. Moreover, the average

TABLE 3. Performance measure.

AC (%)	PR (%)	RE (%)	F1 (%)
90	90	90	90



```
Out[31]: 6    11147
         5     5526
         3     3611
         4     1809
         2     1162
         7      971
         1      199
         0      163
         8      112
         Name: Predicted DDos Attacks, dtype: int64
```

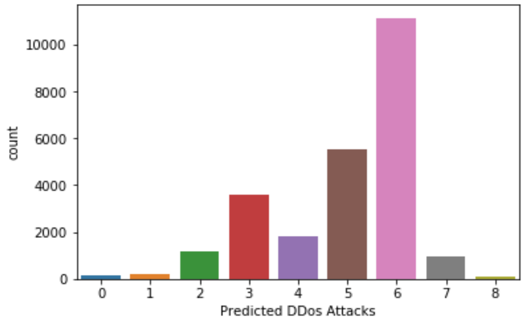


FIGURE 10. First prediction of random forest classifier.

Accuracy (AC) of our proposed approach is ∼90% which is wonderful and extremely awesome. This should be noted that the average accuracy denotes the F1 score 90%.

L. FIRST PREDICTION RESULT

In prediction, we used classification to generate the prediction results for future decisions. Then, we present the prediction results and outcomes, graphically. Figure 10, as shown below, demonstrates the prediction results of the random forest method.

This prediction showed Normal (7) = 11,147, Generic (6) = 5,526, Exploits (5) = 1,809, Fuzzers (4) = 1,162, DoS (3) = 971, Reconnaissance (2) = 199, Analysis (1) = 163, Backdoor (0) = 112, attacks for future decision. As evident from the results, this prediction, as compared to the actual data, is approximately 89% accurate.

M. SECOND PREDICTION RESULT

Figure 11, as shown below, demonstrates the prediction results and outcomes for the XGBoost machine learning algorithm.

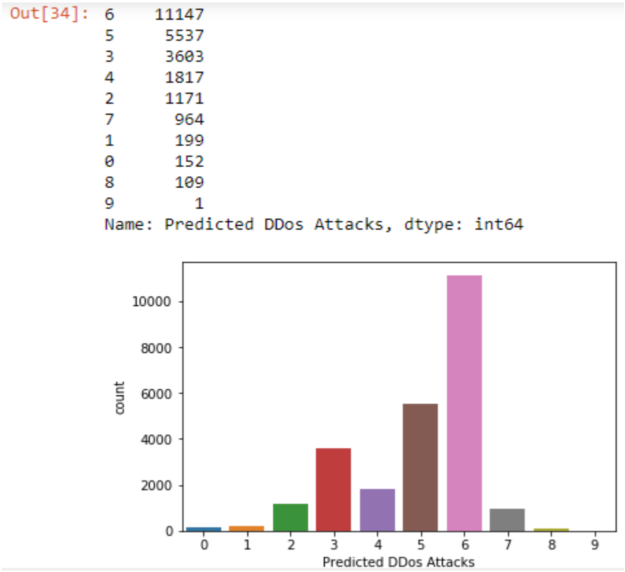
This prediction showed Normal (8) = 11,147, Generic (7) = 5,537, Exploits (6) = 3,603, Fuzzers (5) = 1,817, DoS (4) = 1,171, Reconnaissance (3) = 994, Analysis (2) = 199, Backdoor (1) = 152 and Shellcode (0) = 109 attacks for future decisions. Our evaluation and observations suggest that this prediction, as compared to actual data, is approximately 90% accurate.

N. WORK COMPARISON

In existing research, the [4] used UNSW-nb 15 dataset for the proposed work, and they performed the CNN model

TABLE 4. Work comparison of the proposed model against other closes rivals.

Research Work	Dataset	Algorithm	Average Accuracy Score
[4]	UNSW-nb15	(CNN), BAT-MC, BAT	79%
[3]	KDD	CNN + LSTM	85%
[5]	KDD, UCI	CNN + LSTM	85.14%
[7]	KDD, UCI	SVM	78.34%
This Research	UNSW-nb15	Random Forest	89%
		XGBoost	90%



```
Out[34]: 6    11147
         5     5537
         3     3603
         4     1817
         2     1171
         7      964
         1      199
         0      152
         8      109
         9        1
         Name: Predicted DDos Attacks, dtype: int64
```

FIGURE 11. Second prediction of XGBoost classifier.

for classification. The overall score of this work was 79%. Also, the [3] same work with the same algorithm as the LSTM attention method. They used the KDD dataset for the proposed work and found 85% average accuracy for his work. As compared to our proposed work, we used supervised learning models i.e. Random forest and XGBoost on UNSW-nb 15 datasets [31]. We also used hyper-parameters in this proposed model. We found very good accuracy from 89% to 90%, approximately. The comparative study of the proposed algorithm with other closest rivals such as CNN, SVM, using different datasets, is shown in Table 4. Based on our observations and results, we noted that the XGBoost machine learning model is more suitable for detecting the DDoS attacks. Furthermore, supervised models are also superior to the non-supervised techniques. However, these results are strongly dependent of the dataset being used for the training and testing phases.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a complete systematic approach for detection of the DDOS attack. First, we selected the UNSW-nb15 dataset from the GitHub repository that contains information about the DDoS attacks. This dataset was provided by the Australian Centre for Cyber Security (ACCS) [29], [30]. Then, Python and jupyter notebook were used to work on data wrangling. Secondly, we divided the dataset into two classes i.e. the dependent class and the independent class. Moreover, we normalized the dataset for the algorithm. After data normalization, we applied the proposed, supervised, machine learning approach. The model generated prediction and classification outcomes from the supervised algorithm. Then, we used Random Forest and XGBoost classification algorithms. In the first classification, we observed that both the Random Forest Precision (PR) and Recall (RE) are approximately 89% accurate. Furthermore, we noted approximately 89% average Accuracy (AC) for the proposed model that is enough good and extremely awesome. Note that the average Accuracy illustrates the F1 score as 89%. For the second classification, we noted that both the XGBoost Precision (PR) and Recall (RE) are approximately 90% accurate. We noted approximately 90% average Accuracy (AC) fo the suggested model that is wonderful and extremely brilliant. Again, the average Accuracy illustrates the F1 score as 90%. By comparing the proposal to existing research works, the defect determination accuracy of the existing research [4] which was 85% and 79% were also significantly improved.

Looking to the future, for functional applications, it is important to provide a more user-friendly, faster alternative to deep learning calculations, and produce better results with a shorter burning time. It is important to work on unsupervised learning toward supervised learning for unlabeled and labeled datasets. Moreover, we will investigate how non-supervised learning algorithms will affect the DDoS attacks detection, in particular, we non-labeled datasets are taken into account.

ACKNOWLEDGMENT

The implementation code of the proposed method will be provided for a research purpose if requested from the principal author through email.

REFERENCES

[1] N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, "Adversarial machine learning applied to intrusion and malware scenarios: A systematic review," *IEEE Access*, vol. 8, pp. 35403–35419, 2020.

[2] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020.

[3] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," *IEEE Access*, vol. 8, pp. 29575–29585, 2020.

[4] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-xgboost model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020.

[5] A. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty, and V. S. Kiran, "Similarity based feature transformation for network anomaly detection," *IEEE Access*, vol. 8, pp. 39184–39196, 2020.

[6] L. D'hooge, T. Wauters, B. Volckaert, and F. De Turck, "Classification hardness for supervised learners on 20 years of intrusion detection data," *IEEE Access*, vol. 7, pp. 167455–167469, 2019.

[7] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512–82521, 2019.

[8] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," *IEEE Access*, vol. 8, pp. 42169–42184, 2020.

[9] C. Liu, Y. Liu, Y. Yan, and J. Wang, "An intrusion detection model with hierarchical attention mechanism," *IEEE Access*, vol. 8, pp. 67542–67554, 2020.

[10] S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, "Toward a lightweight intrusion detection system for the Internet of Things," *IEEE Access*, vol. 7, pp. 42450–42471, 2019.

[11] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine learning-based network vulnerability analysis of industrial Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6822–6834, Aug. 2019.

[12] Y. Chen, B. Pang, G. Shao, G. Wen, and X. Chen, "DGA-based botnet detection toward imbalanced multiclass learning," *Tsinghua Sci. Technol.*, vol. 26, no. 4, pp. 387–402, Aug. 2021.

[13] X. Larriva-Novo, V. A. Villagrá, M. Vega-Barbas, D. Rivera, and M. S. Rodrigo, "An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets," *Sensors*, vol. 21, no. 2, p. 656, Jan. 2021.

[14] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, p. e4150, Jan. 2021.

[15] M. Aamir, S. S. H. Rizvi, M. A. Hashmani, M. Zubair, and J. A. Usman, "Machine learning classification of port scanning and DDoS attacks: A comparative analysis," *Mehran Univ. Res. J. Eng. Technol.*, vol. 40, no. 1, pp. 215–229, Jan. 2021.

[16] M. Kwak and Y. Cho, "A novel video steganography-based botnet communication model in telegram SNS messenger," *Symmetry*, vol. 13, no. 1, p. 84, Jan. 2021.

[17] A. Agarwal, M. Khari, and R. Singh, "Detection of DDOS attack using deep learning model in cloud storage application," *Wireless Pers. Commun.*, vol. 2, pp. 1–21, Mar. 2021.

[18] Z. Akhtar, "Malware detection and analysis: Challenges and research opportunities," 2021, *arXiv:2101.08429*.

[19] D. C. Can, H. Q. Le, and Q. T. Ha, "Detection of distributed denial of service attacks using automatic feature selection with enhancement for imbalance dataset," in *Proc. ACIIDS*, 2021, pp. 386–398, doi: 10.1007/978-3-030-73280-6_31.

[20] Q. Tian, J. Li, and H. Liu, "A method for guaranteeing wireless communication based on a combination of deep and shallow learning," *IEEE Access*, vol. 7, pp. 38688–38695, 2019.

[21] Q. Cheng, C. Wu, H. Zhou, D. Kong, D. Zhang, J. Xing, and W. Ruan, "Machine learning based malicious payload identification in software-defined networking," 2021, *arXiv:2101.00847*.

[22] S. K. Wanjau, G. M. Wambugu, and G. N. Kamau, "SSH-brute force attack detection model based on deep learning," Murang'a Univ. Technol., Murang'a, Kenya, Tech. Rep. 4504, 2021. [Online]. Available: http://repository.mut.ac.ke:8080/xmlui/handle/123456789/4504

[23] K. S. Sahoo, B. K. Tripathy, K. Naik, S. Ramasubbareddy, B. Balusamy, M. Khari, and D. Burgos, "An evolutionary SVM model for DDOS attack detection in software defined networks," *IEEE Access*, vol. 8, pp. 132502–132513, 2020.

[24] M. Khari, "Mobile ad hoc netwoks security attacks and secured routing protocols: A survey," in *Proc. 2nd Int. Conf. Comput. Sci. Inf. Technol. (CCSIT)*. Bengaluru, India: Springer, Jan. 2012, pp. 119–124.

[25] K. Srinath, "Python—The fastest growing programming language," *Int. Res. J. Eng. Technol.*, vol. 4, no. 12, pp. 354–357, 2017.

[26] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, "Using the jupyter notebook as a tool for open science: An empirical study," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, Jun. 2017, pp. 1–2.

[27] R. Saini and M. Khari, "An algorithm to detect attacks in mobile ad hoc network," in *Proc. 2nd Int. Conf. Softw. Eng. Comput. Syst. (ICSECS)*. Kuantan, Malaysia: Springer, Jun. 2011, pp. 336–341.

[28] P. Singh and M. Khari, "Empirical analysis of energy-efficient hybrid protocol under black hole attack in manets," in *Research in Intelligent and Computing in Engineering* (Advances in Intelligent Systems and Computing), vol. 1254. Singapore: Springer, 2021, pp. 725–734, doi: 10.1007/978-981-15-7527-3_68.

[29] M. Zakarya and A. A. Khan, "Cloud QoS, high availability & service security issues with solutions," *Int. J. Comput. Sci. Netw. Secur.*, vol. 12, no. 7, p. 71, 2012.

[30] M. Zakarya, "DDoS verification and attack packet dropping algorithm in cloud computing," *World Appl. Sci. J.*, vol. 23, no. 11, pp. 1418–1424, 2013.

[31] R. Saini, M. Wadhwa, and M. Khari, "Vulnerabilities and attacks in global system for mobile communication (GSM)," *Int. J. Adv. Res. Comput. Sci.*, vol. 2, no. 3, pp. 139–142, 2011.

ISMAIL received the B.S. degree in computer science from Abdul Wali Khan University Mardan, Pakistan. He is currently pursuing the M.S. degree in computer science with the Brains Institute, Peshawar, Pakistan. He is also a Research Scholar at the Brains Institute. He is also a Microsoft Certified System Engineer and a Cisco Certified Network Associate. His research interests include computer security, DDoS attack classification, detection, prevention, and prediction based on machine learning methods.

MUHAMMAD ISMAIL MOHMAND is currently working as an Assistant Professor at the Brains Institute, Peshawar, Pakistan. He has written original research articles in various international journals, and he is interested in academics and research. His research interests include computer networking, image processing, network traffic estimation, web security services, e-business, and network security.

HAMEED HUSSAIN received the bachelor's degree in information technology from Gomal University Dera Ismail Khan, Pakistan, in 2007, and the M.S. and Ph.D. degrees in computer science from the COMSATS Institute of Information Technology (CIIT), Pakistan, in 2009 and 2017, respectively. He is the author of several international publications. He is an Active Researcher. His research interests include optimization, machine learning, fog and edge computing, real-time systems, resource allocation, and load balancing in high-performance computing.

AYAZ ALI KHAN received the Ph.D. degree in computer science from Abdul Wali Khan University, Pakistan. He is currently an Assistant Professor with the Department of Computer Science, University of Lakki Marwat, Pakistan. He has deep understanding of the theoretical computer science and data analysis. Furthermore, he also owns deep understanding of various statistical techniques which are, largely, used in applied research. His research has appeared in several international conferences, journals, and transactions of repute. His research interests include cloud computing, mobile edge clouds, the Internet of Things (IoT), performance, energy efficiency, algorithms, and resource management.

UBAID ULLAH received the master's degree in software engineering from Riphah International University, Islamabad, Pakistan. He is currently working as a Senior Lecturer at the Brains Institute of Science and Technology, Peshawar. His research interests include aspect-oriented software development, reverse engineering, requirements engineering, and machine learning.

MUHAMMAD ZAKARYA (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Surrey, Guildford, U.K. He is currently a Lecturer with the Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Pakistan. He is the Program Director of the iFuture: a leading research group at AWKUM which has research collaboration with the CLOUDS Laboratory, The University of Melbourne, Australia, and the IoT Laboratory, Cardiff University, U.K. He has been listed in the world's top 2% scientists list for 2020. He has deep understanding of the theoretical computer science and data analysis. Furthermore, he also owns deep understanding of various statistical techniques which are, largely, used in applied research. His research has appeared in several international conferences, journals, and transactions of repute. His research interests include cloud computing, mobile edge clouds, the Internet of Things (IoT), performance, energy efficiency, algorithms, and resource management. He is a TPC member of few prestigious international conferences, including CCGrid, GECON, and UCC. He is also an Associate Editor of IEEE ACCESS journal and *Journal of Cloud Computing* (Springer). He is also a Guest Editor of *Cluster Computing* journal (Springer).

AFTAB AHMED received the Ph.D. degree in electronic engineering from the University of York, U.K., in 2019. He is currently a Lecturer with the Computer Science Department, Abdul Wali Khan University Mardan, Pakistan. His research work is related to improvement in performance in ultra-dense high-capacity networks. His research interests include radio resource management, topology management to improve system performance and overall energy efficiency in ultra-dense high-capacity wireless networks and machine learning.

MUSHTAQ RAZA received the Ph.D. degree in computer science from the Faculty of Sciences, University of Porto, Portugal, with a focus on software engineering. He is currently an Assistant Professor of computer science with Abdul Wali Khan University Mardan (AWKUM) and a Research Collaborator with the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto. Previously, he was a Researcher with INESC TEC and has published more than 20 papers in renowned journals and conferences in software engineering. His research interests include software process improvement, machine learning, big data analysis, software engineering, and the Internet of Things (IoT). He is also a Program Committee Member of ICSSP, top conference in software engineering, and a Focal Person of the National Technology Fund with AWKUM.

IZAZ UR RAHMAN received the Ph.D. degree in computer science from the Department of Electronic and Computer Engineering, Brunel University, U.K. He is currently an Assistant Professor with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. His research interests include power systems, optimization algorithms, the Internet of Things, and artificial intelligence.

MUHAMMAD HALEEM is currently an Assistant Professor with the Department of Computer Science, Faculty of Engineering, Kardan University, Kabul, Afghanistan. His research interests include the Internet of Things, machine learning, and data analytics.