PREDICTIO OF DDO'S ATTACKS BY USING MACHINE LEARNING

G. Sathwika[1] , E. Raju Kumar[2] , Dr. A. Prathapa Reddy, Professor[3]

[1,2.3] SVS GROUP OF INSTITUTIONS, Hasanparthy, Bheemaram, Hanamkonda, Telangana 506015

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 classication 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.

Keywords : ddos, kdd, xgboost, RE, PR, RE, Random Forest algorithm

1. 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 spoo_ng) 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].

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 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 _ood, absurd messages are still sent, and the association will not be closed, thus turning off the help [2]. The UDP _ood 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 _rewall that is used to protect the server (targeted) may also become overwhelmed as a consequence of the UDP ooding 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 food 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 passing 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 laws 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 XG Boost 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 XG Boost 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.

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 preprocessing 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. PROBLEM STATEMENT

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 CNNC 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.

3.1 LIMITATIONS

The system doesn’t have the accuracy and effectiveness. There is no real-world datasets to evaluate OFDPI's exhibitions on the Ryu SDN regulator and Mininet stage.

IV. PROPOSED SYSTEM

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.

The first step involves the selection of dataset for utilization. The second step involves the selection of tools and language. The third step involves data pre-processing techniques to handle irrelevant data from the dataset. In the fourth step feature extraction and label. Encoding is performed to convert symbolical data into numerical data. 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.

4.1 LIMITATION

The system is designed and developed an approach using supervised machine learning classifiers for DDoS attack detection based on different techniques. The proposed system is designed a step-by-step framework for data utilization.

V. METHODOLOGY

5.1 Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results in Line Chart, View Prediction Of DDOS Attack Type, Find View Prediction DDOS Attack Type Ratio, Download Predicted Datasets, View DDOS Attack Type Ratio Results, View All Remote Users.

5.2 View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

5.3 Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like PREDICT DDOS ATTACK TYPE, VIEW YOUR PROFILE.

VI. ALGORITHMS USED

6.1 Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

6.2 Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

6.3 K-Nearest Neighbors (KNN)

Simple, but a very powerful classification algorithm. Classifies based on a similarity measure. Non-parametric. Lazy learning. Does not “learn” until the test example is given. Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

6.3 Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

6.4 Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

6.5 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

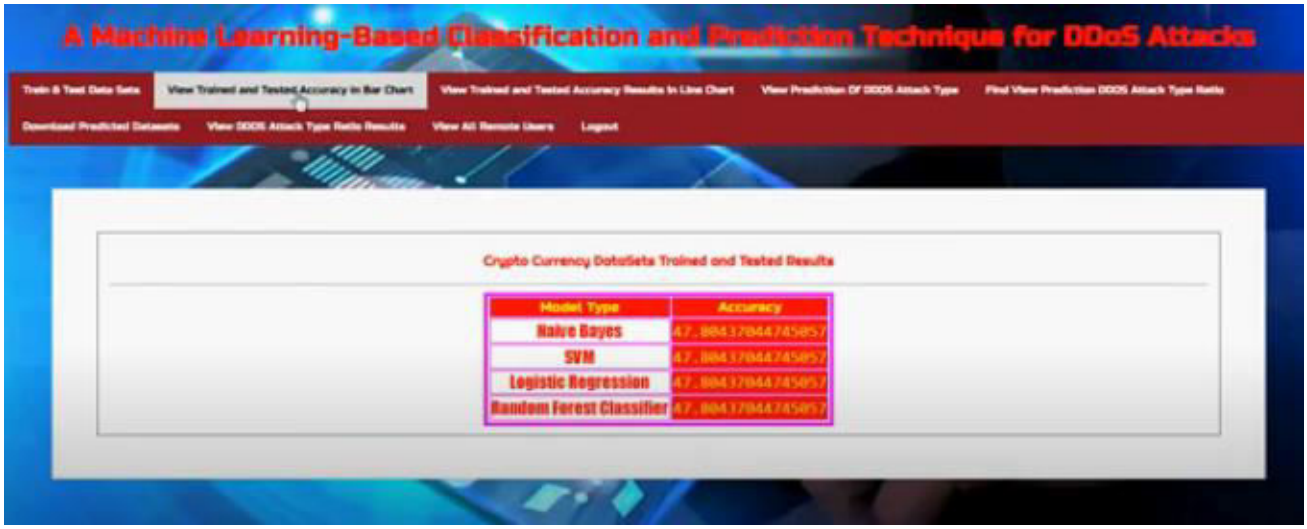
An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

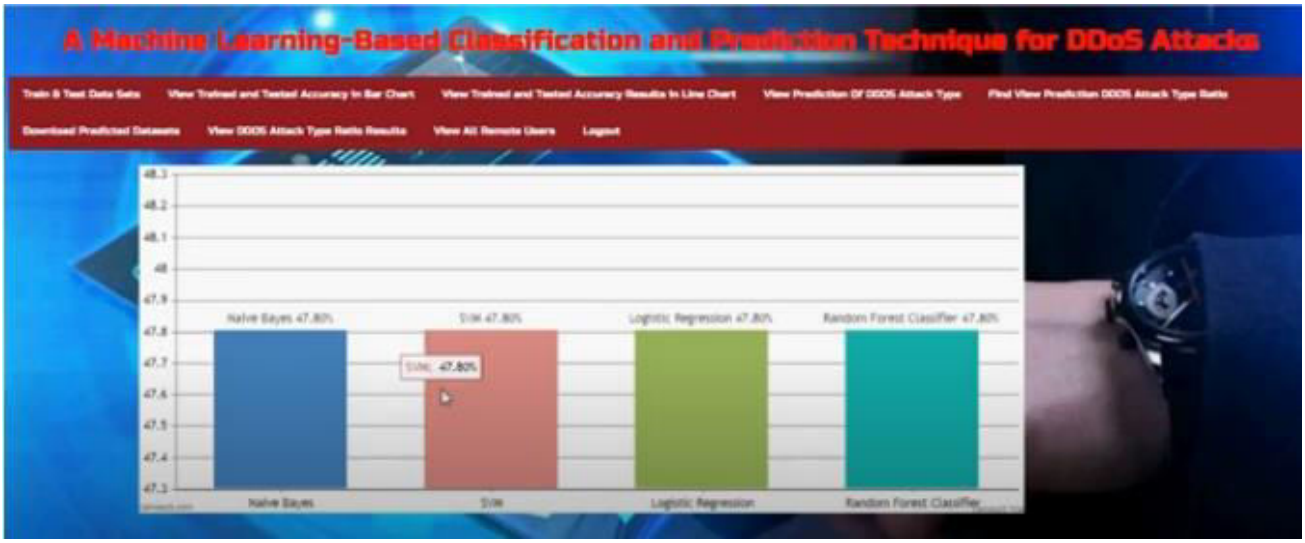
Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

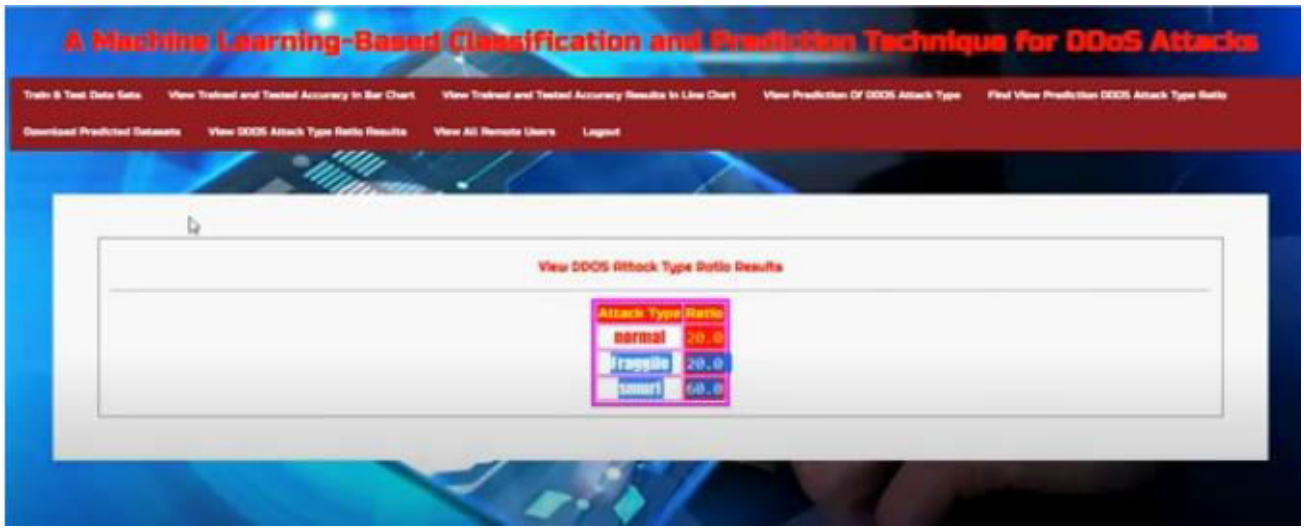
6.6 SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

VII. RESULTS







VIII. CONCLUSION

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 XG Boost 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 XG Boost Precision (PR) and Recall (RE) are approximately 90% accurate. We noted approximately 90% average Accuracy (AC) of 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.

IX. 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, ``Classi_cation 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 classi_cation 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. Com- mun.*, 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 shallowlearning," *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 identi_cation in softwarede _ned 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 de_ned 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.