EasyChair Preprint

№ 8317

Classification of Firewall Logs Actions Using Machine Learning Techniques and Deep Neural Network

Batool Al-Tarawneh and Hani Bani-Salameh

June 19, 2022

Classification of Firewall Logs Actions Using Machine Learning Techniques and Deep Neural Network

Batool A. AL-Tarawneh[1, a)] and Hani Bani-Salameh[1, b)]
[1)]*Department of Software Engineering, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology*

[a)] 2070379@std.hu.edu.jo
[b)] hani@hu.edu.jo

Abstract. The analysis of firewall logs is one of the most significant practices considered while monitoring network traffic to assess their impact. The log records of the Turkish Firat University's firewall device were analyzed using K-Nearest Neighbor (KNN), Random Forest (RF), and Deep Neural Network (DNN) classifiers. A comparison was conducted to measure the performance of the classifier in terms of accuracy, recall, precision, and F1 score. 65,532 records were examined using 12 attributes, where the action was identified as a label of these attributes because it handles the packets based on their features either allowing them to pass, blocking them, blocking their activity, or blocking the request itself. The result of this analysis indicated that the best algorithm that selects the best features according to appropriate action is Random Forest.

INTRODUCTION

The firewall is a network security device that can be a hardware or software and is responsible for monitoring the packet traffic that enters and exits the network to determine whether or not specific traffic will be allowed. The firewall logs network access attempts, including the source and destination IP addresses, port numbers, and other information. The logs explain how the firewall handles incoming and outgoing network connections based on the source and destination IPs, ports, protocols, and many other parameters. Many researchers analyze firewall logs for unnecessary or too tolerant rules to eliminate them, which improves the identification of possibly harmful behavior [1] [2].

Machine learning approaches must use a huge amount of historical data, which may lead to difficulties such as variation caused by thousands or millions of variables. It is a trade-off, the larger and more comprehensive the sample size, the more trustworthy the results. If the data samples are too little, not all patterns will be caught or considered, resulting in inaccurate results due to insufficient detection of existing patterns. In this paper, a data set consisting of 65,532 instances was obtained from Kaggle of Firat University Firewall logs.

Machine learning also requires algorithms that are appropriate for the given dataset, therefore three algorithms have been used, including KNN, which employs all input variables and training samples for each new observation that is classified. In addition to the RF since it handles non-linear relationships effectively which is the situationin the used dataset because it has numerous decision trees. Finally, the DNN is used to find its effects on multiple features-single classes in the numerical datasets.

Due to the significance of data quality, it was necessary to process it in a manner suitable to the form of the data fed into machine learning algorithms, by dropping features that do not add value to the classification process or result in waste of processing time without benefit, in addition to dealing with anomalies and variance in values.

This paper examines the effect of KNN, RF, and DNN on data and compares their results with the support vector classifiers used in [3]. After the classification process, the accuracy, F1 score, precision, and recall of the classifiers are analyzed, with the F1 score representing a major metric for classifier performance.

The (second section) presents some previous works on machine learning in the field of network and logs analysis. The (third section) presents an overview of machine learning and deep learning. The (fourth section) briefly explains how the firewall works. The (fifth section) presents the methodology used in this work. The (sixth section) lays down how to prepare the data for classification. The (seventh section) explains the used algorithms. The (eighth section) presents the results. Finally, in the (ninth section) the conclusion is drawn.

LITRETURE REVIEW

Several works have analyzed network security and how to predict cyber threats that networks may encounter using machine learning and its techniques over the decades. [4] proposed classification models that were utilized in firewall systems to generate the appropriate action by analyzing packet features with Shallow Neural Network (SNN) and Optimizable Decision Tree (ODT). Allow, Deny, and Drop/Reset were the three labels. For ODT and SNN, the findings revealed the accuracy of 99.8% and 98.5 %. [5] employed logs to discover a web shell connection and employed LSTM, which had a 95.97 % accuracy rate and a 96.15% recall rate. In [6] attack patterns were analyzed insoftware-defined networks (SDN) using machine learning, 32 points and 17 million login attempts were constructedfor 112 separate locations and applied to over 6000 source IP addresses. Software-defined networking can handle large numbers of attacks by restricting network connections at the switch level. On the other side, historical networkattack data may be utilized to automatically detect and approve or deny risky connections. As a result, it has been recommended that previous network attack data be used to train machine learning algorithms such as C4.5, Bayesian(BayesNet), DT, and Naive-Bayes to predict which host would be attacked next. The best technique was found to beBayesian networks. Furthermore, in [7] Machine learning techniques were used in to learn network characteristics. Italso discussed the issues and challenges associated with dealing with big data classification, as well as the problems associated with combining supervised, unsupervised, and lifelong learning techniques. Web application firewalls were employed [8], as well as signature-based and anomaly-based approaches. the study adopted natural language processing techniques and a linear support vector machine learning algorithm, with an overall detection accuracy rateof 99.53%.

MACHINE LEARNING AND DEEP LEARNING: A BRIEF OVERVIEW

Machine learning is a sub-field of Artificial Intelligence (AI) that focuses on using data and algorithms to learnand enhance accuracy in a way that is close to how humans learn. Machine learning is a critical aspect of data science. Algorithms are built to make decisions in Internet-based applications, corporations, businesses, and devices that depend on the Internet of Things (IoT).

Machine learning algorithms are mathematical methods that use input data to accomplish such a project or a task without periodic programming. Using iteration, these algorithms automatically modify or adjust their structure, improving their ability to perform a desired task. During the training phase, samples of input data are presented alongside the intended outputs. The algorithm then optimizes itself to create not just the desired output, but also to make generalizations to obtain the best possible result from new data. Machine learning's "training" element refers to the training set[9].

Machine learning models generate algorithms that are activated by acquiring as much input data as possible and producing an accurate output to predict the proper output for new data. There are four major types of machine learning algorithms: supervised, unsupervised, semi-supervised, and reinforcement learning [10].

Machine learning is divided into four categories:

- Supervised machine learning is used for structured datasets, it is used to train algorithms that classify data or reliably predict outcomes. As the model receives input data, it improves the weights until the structure is properly fitted. Supervised learning using methods such as Neural Networks (NN), Linear Regression (LR), Random Forests (RF), support vector machine (SVM), and others are used to solve many real-world problems.

- Unsupervised machine learning analyses unclassified data sets via machine learning algorithms. Without the need for human interference, these algorithms uncover invisible patterns or data sets. It is the perfect solution for exploratory data analysis, images, and pattern recognition because of its ability to discover similarities and differences in data.

- Semi-supervised learning is a type of learning that falls between supervised and unsupervised learning. A smaller, structured dataset is used to direct classification and extract features from a larger, unlabeled dataset during training. It may address the issue of missing classified data.

- Reinforcement learning is a type of machine learning similar to supervised learning but without the use of sample data to train the algorithm. Rather, it uses a trial-and-error approach to learn as it works.

DNN is a sort of Artificial Neural Network (ANN) that is trained using algorithms to learn representations from data sets. It is one of the types of deep learning. Because it leverages more complicated and non-linear functions. These algorithms have a greater and deeper number of processing layers. The dependence on and growth of machine learning, as well as deep learning's fly-by of the big data boom of the previous decade [11].

FIREWALL MECHANISM

Firewalls protect devices and networks against a wide range of security threats, including unauthorized access from outside the firewall network. They also create reports on multiple threat attempts and analyze incoming packets for attacks or viruses that may be considered dangerous. If these packets are found to be malicious, the firewall preventsthem from reaching the internal network and, as a result, from reaching users' devices.

To prevent attacks, firewalls analyze incoming traffic based on specified policies and filter traffic from unsecured or suspicious sources. Data traffic is protected by firewalls at the data entry point of a computer, known as the port, where data is exchanged with external devices like saying that the source address 172.169.1.1 is allowed or denied toreach the destination in the internal network.

Filtering packets is one of the firewall's major functions. A packet is a small group of data that travels together, when the firewall implements this function, the packets attempting to enter the network are passed to a series of filters. These filters filter out packets that meet predefined threats while allowing others to proceed to their destination [12].

Firewalls avoid a lot of attacks like Backdoors, so it helps in protecting users from exploiting vulnerabilities in applications and devices that allow remote access, and denial of service which is a frequent kind of cyber attacks thatslows down or kills a server. In the Open Systems Interconnection (OSI) model, the firewall operates at the networklayer, where it bases its defensive mechanisms on the IP and header data to determine whether or not to allow a packetto pass. Figure 1 depicts an example of traffic generated from the Internet and being directed towards devices withinthe network (green arrows), where the firewall prevented malicious packets from reaching them (red curvy arrow).

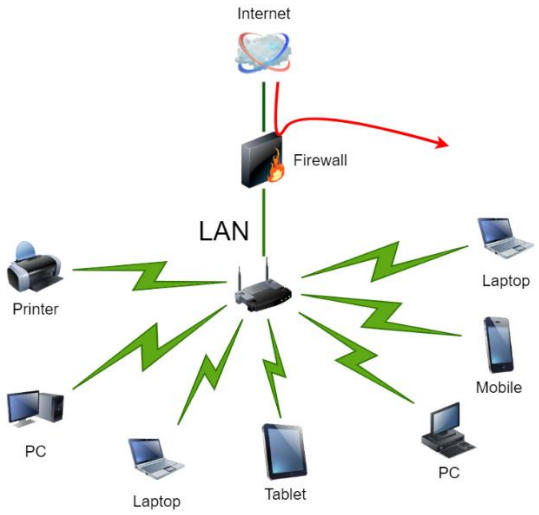

FIGURE 1. Firewall protecting a harmful packet to access the network

METHODOLOGY

Figure 2 describes the proposed methodology in an illustrative diagram, with information packets coming at the firewall device of Firat University that provided the logs. Additionally, data was processed by removing outliers and splitting it into features and labels, followed by machine learning models that were used to determine the appropriateaction for each packet, and finally, the accuracy was determined and the results were displayed using the classification report.

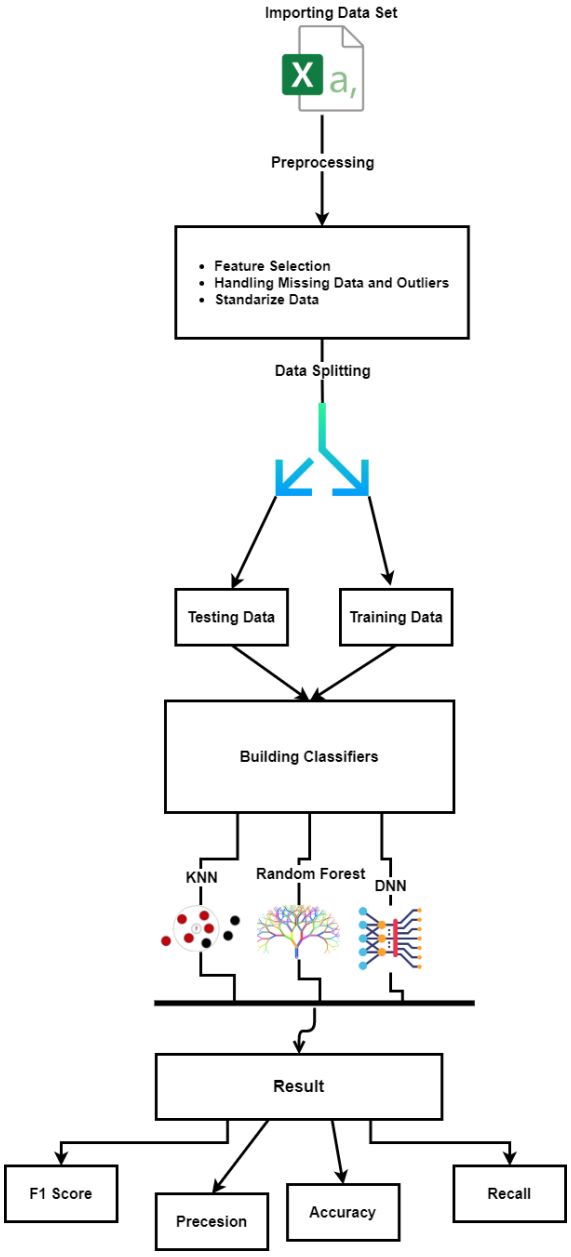

FIGURE 2. The proposed approach for machine learning classifiers

Preprocessing data, or the act of preparing raw data and making it acceptable for machine learning models is the first and most important stage in developing a machine learning model. The data used in a machine learning project is rarely clean. As a result, it is critical to clean and prepare data before beginning a process of machine learning.

Preprocessing data in the real world is important because it avoids outliers and missing values that make it impossible to implement machine learning models directly, thus increasing the accuracy and efficiency of the model. Preprocessing includes several important steps to get the data ready for use in machine learning models, starting from obtaining data set or importing it (which was adopted in this paper), importing libraries, searching for missing data, encoding the data, standardizing it, and finally, the feature selection [13].

After that, dividing the dataset into training and testing sets is a must. A training set is used for the model training process, while a testing set is used to test the model that has been already trained. Cross-validation with 10 k-fold is also used to avoid overfitting. The test group must be big enough to produce statistically meaningful outcomes, and the entire data collection must be representative. The test set is a representative sample of new results. In this work, the data is trained and tested with an 80:20 ratio.

KNN, RF, and DNN were used to classify the data in various ways to see which one performed best in evaluating the required action and then calculating the accuracy, F1 score, recall, and precision based on the confusion matrix. The data were collected, processed, tested, and trained in this study to investigate which classifier could bestuse the features to pick the best action for the packets.

PREDICTION EVALUATION

This section covers the classification process, like how to utilize DNN, KNN, and RF classification algorithms, as well as the hyperparameters that are compatible with the data.

Dataset Analysis and Preprocessing

The dataset was imported from Kaggle, it includes 12 attributes and 65,532 elements. Table 1 shows the attributesin the first 11 rows as features and the label is in the last row. The features are used to describe the dependent data. Labels are used to represent the independent data.

TABLE 1. Dataset description; The first eleven rows are the feature and the last row is the label

Parameter	Description	Data Type
Source Port	The port number where the packet has been sent from the network	Numeric
Destination Port	The port number where the packet has received inside the network	Numeric
NAT Source Port	The port number where the packet has been sent from the Internet	Numeric
NAT Destination Port	The port number where the packet has been received by the Internet	Numeric
Bytes Sent	Number of Sent Bytes	Numeric
Bytes Received	Number of Received Bytes	Numeric
Bytes	Summation of the sent and received bytes	Numeric
pakts sent	Number of Sent Packets	Numeric
pakts Received	Number of Received Packets	Numeric
Packets	Total number of the sent and received packets	Numeric
Elapsed Time (sec)	The time elapsed for a packet from the leaving the source until it arrived at the destination	Numeric
Action	describe the action had taken for each packet	String

After importing the data set, pakts Received, pakts sent, bytes received and bytes sent features have been dropped

due to their lack of great influence since their summation is found in the packets' and bytes' columns respectively. Thedependent and independent variables are then extracted, with one independent variable in this work, action, indicatingthe mechanism for which the firewall will deal with the packets. The null was also checked in the dataset after dropping the columns that do not provide any additional information crucial in the classification process. There was no missing information. The number of features for each label with their number and percentage in the dataset was extracted as in Table 2.

TABLE 2. Labels distribution in the firewall 's logs

Label	Frequencies	Percentages
Allow	37640	57.4376%
deny	14987	22.8697%
drop	12851	19.6103%
reset-both	54	0.0824%

Standardization is performed at the end of the preprocessing phase, which is a statistical technique in which values are oriented around the mean with a unit standard deviation. This implies that the attribute's mean becomes between zero and one. The standardization formula (Equation 1) where the standard deviation of the feature values is σ, X isthe original value, and the mean of the feature values is μ, and is extracted from the main standard deviation equation [14].

$$\acute{X} = \frac{X - \mu}{\sigma} \tag{1}$$

X in (Equation 1) is calculated in (Equation 2) while X' is the new value and X is the original value:

$$\acute{X} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2}$$

To analyze the data in-depth to understand it and how it is distributed, it was also needed to rely on the distribution of features according to their percentile, as shown in Table 3:

TABLE 3. Dataset features' description

	Source Port	Destination Port	Nat Source Port	Nat Destination Port	Bytes	Packets	Elapsed Time (sec)
count	65532	65532	65532	65532	65532	65532	65532
mean	49391.96	10577.38	19282.97	2671.04	97123.95	102.86	65.83
std	15255.71	18466.02	21970.68	9739.16	5618439	5133	302.46
min	0	0	0	0	60	1	0
25%	49183	80	0	0	66	1	0
50%	53776.5	445	8820.5	53	168	2	15
75%	58638	15000	38366.25	443	752.25	6	30
max	65534	65535	65535	65535	1269359000	1036116	10824

Splitting Data

In the dataset, the data were split into the training data and the testing data in an 80:20 ratio [15]. This indicates that 80% of the experiences and observations are used for learning and the remaining 20% are used for testing. When there is a lack of training data, parameter estimates could vary greatly. Fewer test results, on the other hand, lead to greater consistency in performance measures. In addition to the use of cross-validation for a more accurate result.

INITIATE CLASSIFICATION PROCESS

Based on the previous section, data analysis and distribution were used for the RF, KNN, and DNN classifiers for the following reasons:

- The RF was selected because it gives better accuracy with cross-validation, as well as the ability to handle missing data. Furthermore, when there are a large number of trees, the accuracy of the results will not be affected by the over-fitting [16].

- The KNN algorithm was used, although it does not perform well with data that have high variance in the values, to notice its effect when the values are limited to a specific range (0 and 1), as was done during the preprocessing data by "StandardScaler" function.

- DNN algorithm has been used for its enormous processing power, especially in non-linear data, as in the datasetused in this work.

Random Forest Classifier

Random forest is a quick and simple machine learning algorithm that consistently produces excellent results. Because of its simplicity and versatility, it is the most commonly used algorithm for both classification and regression tasks. The RF looks for the best feature among a random subset of features as they expand. As a consequence, there is a lot of variation, which leads to a better model overall. As a result, the node splitting algorithm in RF only needs to take into account a random subset of features.

A grid of hyperparameter ranges is constructed using Scikit-Learn's "GridSearchCV", and K-Fold cross-validation is performed with each set of values to determine the algorithm that best matches the available data. Insidethe random forest, 33 decision trees are used in the built model. The n_estimators hyperparameter is the number of trees that the algorithm creates before considering the overall result or integrating the prediction. Getting more trees, on the other hand, improves efficiency and performance, and makes predictions more consistent, but it slows down the calculations. The hyperparameters used are shown in Table 4:

TABLE 4. Random Forest algorithm hyperparameter

hyperparameter	description	value
bootstrap	In each decision tree, this is the maximum number of levels	False
max_depth	In each decision tree, this is the maximum number of levels	4
max_features	The feature used when looking for the optimum split	auto
min_samples_leaf	the smallest amount of data points that may be maintained on a leaf node	2
min_samples_split	the smallest amount of data points that can be placed in a node before it is split	5
n_estimators	number of trees in the algorithm	33

When bootstrap is false; all data is used to fit the model, so a random subsample of features can be selected in eachsplit. The number of estimators is 33, which means the number of trees in the forest is proportional to the

depth, because the deeper one tree is, the more information is captured in the data. The max feature is auto, meaning it willtake all the features in each tree which will lead to a good result.

K-Nearest Neighbor Classifier (KNN)

K-Nearest Neighbour is a supervised machine learning algorithm that assumes the new data and its states are equivalent to the existing data and sets the new state in the group that is the most compatible with the existing groups. The K-NN algorithm stores all existing data and classifies new data points based on similarity; that is, it is simple to groupnew data into a good group when it comes. In the K-NN algorithm, the data set is only stored during the training process[17].

K-NN algorithm initially works by determining the number of K neighbors and then calculating the number of datapoints in each class between these K neighbors to complete the model. The value of cross-validation k-fold was alsochosen as 10, as is common, and it was demonstrated in [18] that there is no definite best value for cross-validation and that it relies on experience. The number of best neighbors was determined using a code to estimate the appropriatenumber of neighbors.

Three hyperparameters were selected in the KNN algorithm where n_neighbors is set to 1; which means 1 neighbor is required to rank each point. The measure of distance we use is Minkowski, and the value p = 2 is defined as the Euclidean distance.

Deep Neural Network (DNN)

Artificial Neural Networks (ANN) are advancements of conventional and classical machine learning techniques that employthe hidden layer(s) to store and assess the significance of a single input to all outputs. The hidden layer takes account of the significance of inputs and integrates the importance of groupings of inputs. The DNN is an Artificial Neural Network with hidden processing layers, and each has many neurons with activations as its outputs [19].

Initially, the input layer is checked for the right amount of input features. The input variables are set to 7 for the seven input data (features). The optimal network structure was discovered by trial and error since there is no standard or an optimum number of layers. The first hidden layer was set to 12 nodes and the relu activation function, as well as the second layer, while the output layer contained one node using the sigmoid activation function.

The optimizer is defined as “Adam” and is known for its strength of gradual descent because of its capacity to constantly tune itself and produce good results in a variety of situations. Each epoch is split into batches, Where batch is the number of data points processed just before the model is updated, and epoch is the number of full passesover the training data set. The work was done on a modest number of epochs (100) with a batch size of (10), these configurations were chosen practically by experimentation to train the model to be good enough (at least).

RESULT

After using the RF, KNN, and DNN classification algorithms, the accuracy was calculated and the classification report was extracted to observe the performance of each classifier and its effect on the available dataset. The classification report shows Recall, F1 Score, and Precision according to (Equations 3,4,5) respectively. These metrics depend on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score = 2 \times \frac{P \,.\, R}{P + R} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Based on the preprocessing and evaluation in the previous two sections, the results were as in Table 5, which shows the classification report micro average results of the used classifiers in comparison with the results of the original work, taking into account the F1 Score as a key comparison factor. The macro average has been taken because it ensures that all samples in the dataset contribute equally to the final average scale.

TABLE 5. Classification Report Summary

Classifier	Precision	Recall	F1-Score
Random Forest	1.00	0.85	0.89
K-Nearest Neighbour	0.81	0.82	0.81
Deep Neural Network	0.25	0.50	0.33
	Original Paper Result [3]		
SVM Linear	0.675	0.853	0.754
SVM Polynomial	0.618	0.474	0.536
SVM RBF	0.63	0.971	0.764
SVM Sigmoid	0.603	0.987	0.748

The accuracy for RF is 99.7%, for KNN it is 99.3%, and for DNN it is 49.47%, so it is obvious that RF is the best algorithm to predict the best action for the firewall input values and has the best F1-Score or so-called Harmonic mean which identifies how the prediction is good and complete.

CONCLUSION

A firewall is an advanced tool that has become more and more widespread, especially with the increasing number of devices that rely on the Internet, and thus it protects and secures the information stored on the devices. As a result, the task of the firewall is to protect these devices by recording all logs of transmitted data and filtering the data they have. In this work, 65532 packets from Firat University's firewall logs were processed and prepared to be suitable formachine learning models, Random Forest, K-Nearest Neighbor, and Deep Neural Network were used to analyze traffic packet values. The data training and testing were done with an 80:20 ratio. The hyperparameters of each algorithm were adjusted in proportion to the nature of the data. After completing the machine learning process, accuracy and other metrics such as Precision, F1-Score, and Recall were extracted. Random Forest was observed to be the best classifier in this experiment with an accuracy of 99.7%. And the F1-Score is better, reaching 85% , that is, it can predict the best data that leads the firewall to make the right decision "Action" to either allow, deny, drop, or reset both.

ACKNOWLEDGMENTS

I would like to express my gratitude to my primary supervisor, Dr. Hani Bani Salameh, who helped and directed mein this work.

REFERENCES

1. D. Jeon and B. Tak, "Blackeye: automatic ip blacklisting using machine learning from security logs," Wireless Networks , 1–12 (2019).
2. A. K. Meena and N. Hubballi, "Nviz: An interactive visualization of network security systems logs," in *2020 International Conference onCOMmunication Systems NETworkS (COMSNETS)* (2020) pp. 685–687.
3. F. Ertam and M. Kaya, "Classification of firewall log files with multiclass support vector machine," in *2018 6th International symposium ondigital forensic and security (ISDFS)* (IEEE, 2018) pp. 1–4.
4. Q. A. Al-Haijaa and A. Ishtaiwia, "Machine learning based model to identify firewall decisions to improve cyber-defense," InternationalJournal on Advanced Science, Engineering and Information Technology **11** (2021).
5. Y. Wu, Y. Sun, C. Huang, P. Jia, and L. Liu, "Session-based webshell detection using machine learning in web logs," Security and Communi-cation Networks **2019** (2019).
6. S. Nanda, F. Zafari, C. DeCusatis, E. Wedaa, and B. Yang, "Predicting network attack patterns in sdn using machine learning approach," in
   2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN) (IEEE, 2016) pp. 167–172.
7. S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," ACM SIGMETRICS Performance Evaluation Review 41, 70–73 (2014).
8. B. IŞ iker and SoĞ ukpinar, "Machine learning based web application firewall," in *2021 2nd International Informatics and Software Engineer-*
   ing Conference (IISEC) (2021) pp. 1–6.
9. I. El Naqa and M. J. Murphy, "What is machine learning? machine learning in radiation oncology," (2015).
10. C. G and J. M. Roogi, "A quick review of ml algorithms," in *2021 6th International Conference on Communication and Electronics Systems(ICCES)* (2021) pp. 1–5.
11. A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE access **7**, 53040–53065 (2019).
12. L. Durante, L. Seno, and A. Valenzano, "A formal model and technique to redistribute the packet filtering load in multiple firewall networks,"IEEE Transactions on Information Forensics and Security **16**, 2637–2651 (2021).
13. H. Bani-Salameh, M. Sallam, *et al.*, "A deep-learning-based bug priority prediction using rnn-lstm neural networks," e-Informatica SoftwareEngineering Journal **15** (2021).
14. P. Dangeti, *Statistics for machine learning* (Packt Publishing Ltd, 2017).
15. A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation,"(2018).
16. scikit learn. (n.d.). Ensemble methods. Retrieved June 15, 2022, from https://scikit-learn.org/stable/modules/ensemble.html?highlight=why+use+random+forest+classifier.
17. J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," Expert Systems withApplications **115**, 356–372 (2019).
18. I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of machine learning algorithms with different k values in k-fold cross-validation,"(2021).
19. T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," IEEE Access **7**, 47230–47244 (2019).