Chapter 15
Machine Learning-Based DDoS Attack Detection Using Support Vector Machine

V. Kathiresan, Vamsidhar Yendapalli, J. Bhuvana, and Esther Daniel

Introduction

In this information age plenty of data generated every second, storing, managing and retrieving the data is a big challengeable task in this era. About 463 Exabyte's of data is going to be generated per day in the year 2025 as per the prediction. In addition to this securing the data from the intrusion and information leakage is the highly essential and most challengeable task.

The cybercrime and data theft increased more than 600% during the pandemic [1, 2]. Ransomware attack increased to the large extend. The companies are investing a lot in cyber security and intrusion deduction. Malware attack, phishing attack, password attack, man in the middle attack and ransomware attack are quiet common attacks.

Especially the distributed denial of service attack is the one which makes the service inactive. It makes the particular server or resource inactive or inoperative. DDoS attack effects the organisation to the large extend [3, 4]. The cost of DDoS attack is maximum when compared to the other attacks. This book chapter focuses the DDoS attack since it is taking a significant role in the cyber-attacks.

Machine learning systems are the one which is learning from the data and doing classification or prediction according to the need [5, 6]. This book chapter focuses on

V. Kathiresan (✉) · V. Yendapalli
Department of Computer Science and Engineering, School of Technology (GST), GITAM University (Deemed to be University), Bengaluru, Karnataka, India
e-mail: xyzkathir@gmail.com

J. Bhuvana
Department of CS and IT, Jain (Deemed to be University), Bangalore, India
e-mail: J.bhuvana@jainuniversity.ac.in

E. Daniel
Karunya Institute of Technology and Sciences, Coimbatore, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

V. Sarveshwaran et al. (eds.), *Artificial Intelligence and Cyber Security in Industry 4.0*, Advanced Technologies and Societal Change,
https://doi.org/10.1007/978-981-99-2115-7_15

utilising the machine learning techniques in the cyber security. Intrusion detection is the one, by which the traffic coming to the internal network from outside the world can be classified as genuine or intrusion, based on the characteristics of the traffic includes destination port, flow duration, total forward packet, total backward packets, flow packets, etc.

Support vector machine is the machine learning technique. It is the machine learning-based classification model which classifies the given input based on the hyper plan. The data points that lie one side of the hyperplane belong to one class. The data points that lie other side of the hyperplane belong to another class [7, 8].

In this work machine learning's support vector machine is used to detect the distributed denial of service attack. The machine learning model is created and trained with the network traffic-based intrusion deduction data which is taken from Canadian Institute for Cyber Security. The dataset contains 79 attributes including destination port, flow duration, total forward packet, total backward packets, etc.

The model is getting trained with the labelled data which is having DDoS attack—yes as one label and DDoS attack—no as another label. Once the model got trained with the training data whenever the new traffic is coming it will detect whether it is an authorised traffic or intrusion. The dataset is divided into training set and testing set. Training set has been used for training purpose. The testing set is used to assess the quality of the model. The metrics precision, recall, $F1$-score and support are used in order to assess the quality of the SVM-based machine learning model.

Previous Work

Mihoub et al. in the year 2022 [9] proposed a method to detect the distributed denial of service attack using the machine learning classification model random forest. Random forest is a machine learning classification method which is commonly used for classification. Regression operation also can be performed by using random forest, but random forest performs well with classification. This proposed method is functioning based on looking back enabled method.

Liu et al. in the year 2020 proposed a method to detect the intrusion in the wireless sensor network [10]. Modified and improved KNN is used to detect the distributed denial of service in the wireless sensor networks. KNN is the well-known classification algorithm in machine learning.

Mahajan et al. in the year 2022 proposed a method to detect distributed denial of service attack using deep learning [11]. The 5G technology is enabling high data rate. In high data rate communication the chances for distributed denial of service attack also more. Here deep learning is utilised to solve the problem of distributed denial of service attack detection. Once the attack is detected then it is handled by mitigation policies. Model's performance is measured by the metrics.

Tonkal et al. in the year 2021 proposed a method to detect the distributed denial of service attack in software-defined network (SDN) [12]. The dataset which is used in this approach contains 23 features. The approach is using KNN, decision tree and artificial neural network approach too. The dataset contains the traffic details of Transfer Control Protocol (TCP), User Datagram Protocol (UDP) and Internet Control Message Protocol (ICMP). It contains both normal traffic and attack traffic. After using the algorithms KNN, decision tree and artificial neural network, it has been proved that decision tree is good in performance basis.

Kumar et al. proposed a method in the year 2011 to detect distributed denial of service attack [13]. Neural classifier is the machine learning-based classification method which is used for detection. KDD Cup, DARPA 1999, DHRPA 2000 or the datasets which are used for the training of the model. In this method the falls positive and the false negative are taken into the consideration. Method has been proved that it is having less false positive and false negative.

Zekri et al. [14] proposed a method using machine learning-based C4.5 algorithm to detect the distributed denial of service (DDoS) attack. C4.5 is the decision tree-based machine learning classification algorithm. The DDoS attack detection is done for cloud computing environment in order to secure the cloud environment from intrusion. Cloud resources can be utilised effectively when security is enabled properly.

He et al. proposed denial of service (DoS) (2017) [15] attack detection method in network using naïve Bayes method. Naïve Bayes method is a machine learning-based classification algorithm. Traditionally DoS attack detection happened based on threshold value. In order to improve the efficiency in detection machine learning-based naïve Bayes classification algorithm proposed. It enhances the security in cloud-based environment.

De Miranda et al. [16] proposed fuzzy logic and machine learning-based algorithm for distributed denial of service (DDoS) attack. In this proposed method reduction of quality (RoQ) attack is targeted. The K-nearest neighbour (KNN) is the classification algorithm based on machine learning which supports DDoS attack detection in the proposed method. Algorithm's performance is proved based on $F1$-score metrics.

Aamir and Zaidi proposed a clustering-based semi-supervised ML method for DDoS attack detection (2021) [17]. This proposed method uses clustering techniques rather than classification method. Since it is a clustering method unlabelled data and partially labelled data can be used. Agglomerative clustering is the clustering method which is used to do the clustering with respect to DDoS detection. Accuracy is measured to assess the quality of the model.

Aysa et al. [18] proposed a method to detect DDoS attack detection for IoT or wireless sensor network (WSN). The method is based on machine learning. Machine learning-based decision tree approach is used to detect the DDoS attack. Decision tree is a classification method which provides solution for DDoS detection. Model's performance is measured by the accuracy metrics of the classification model. It prevents the abnormal traffic in the wireless sensor networks.

Distributed Denial of Service Attack

In the year 2021 the distributed denial of service attack grown 31%. DDoS attack affects the organisation to the large extend. The cost or loss involved in the DDoS attack is high when compared to the other attacks. The service or server which undergone the DDoS attack will become inoperative. It will slowly move operative state to inoperative state. The attack affects all the resources including software and hardware resources. Edge network devices are the target for the DDoS attackers. Monitoring the network traffic and detecting the attack is the best way of detecting the intrusion. Even after that attack DDoS attack makes the system more vulnerable.

Application Layer Attack

The main objective of application layer attack is making the application inoperative. After identifying the vulnerabilities in the application. The attack happened against the application and make application inoperative. It makes the application unable to provide service to its uses. It is done by sending millions of requests with exception. Keep on sending the handshake message even after dialogue over. This kind of attempts makes the server irresponsive to the original user query.

Protocol Attack

Protocol-based distributed denial of service attack is different from application layer attack. Protocol attack is hard to find. There are lot of complications involved in identifying the protocol-based DDoS attack [19, 20]. The vulnerability in the network protocol is identified and utilised in this attack. It is hard to identify based on the complexity in the protocol. The close monitoring of the network traffic and analysis of streams in depth can increase the probability of identifying the protocol attack. Border Gateway Protocol is the one which undergone the protocol attack. Protocol attack is not the one which is frequently happened, but the impact of the attack is high.

Syn Flood

Syn flood attack takes significant role in the DDoS attack. The attacker sends repeated Syn message or packet to the server and makes the server irresponsive. Making the server busy in replying is the objective of the attack. Once the server is receiving frequent more number of Syn messages or Syn packets in order to process all the requests the entire server resources became busy. Making server resources busy and making it irresponsive from the user is the objective of the attack [21].

In the normal three-way handshake of the TCP. The client is sending synchronisation message to the server. The server is replying back with the Syn and acknowledgement message, then again client is sending acknowledgement message. After the three-way handshake the packet will start getting transfer. During Syn flood attack the attacker will send the spoofed Syn packet continuously to the server. The server will became busy by replying the acknowledgement for the packets received.

Volumetric Attacks

Volumetric attack targets the internal network. The attacker targets the internal network and creates the malicious traffic inside the internal network. Due to the artificially created traffic the service which is being delivered to the user or client will get interrupted. The main objective of volumetric attack is consuming the bandwidth of the internal network and makes the server or distributed system inoperative. The volumetric attack finds the vulnerability in DNS in order to occupy the traffic by sending the malicious code.

SVM

Support vector machine is the machine learning-based classification model. SVM mainly focuses binary classification. It supports multiclass classification also. SVM classifies the classes based on support vectors and hyperplane. Hyperplane is the one which classifies the given data into multiclasses. It is a supervised learning algorithm. SVM supports both classification and regression. It is more famous for classification [22].

Hyperplane divides the data environment into two classes. The objective of hyperplane is to have the maximum margin. Positive side of the hyperplane contains 1 class, and negative side of the hyperplane contains another class. Support vector machine is getting trained based on hinge loss function.

$$\text{hingeloss} = \arg\min \sum_{i=0}^{n-1} \max(0, 1 - y_i(w^T x_i + w_0))$$

Hyperplane

Hyperplane divides the data object into classes. Hyperplane is having margin. SVM's objective is to increase the size of margin or having the maximum margin size. When the data environment is one-dimensional dot is the hyperplane which divides the data into two. When the distribution is two-dimensional a line can divide the entire distribution into two classes. In three-dimensional environment hyperplane is a plane. In multi-dimensional, it is called as hyperplane. The data points which are closed to the hyperplane is called support vectors. Support vectors decides the size of the margin. The objective of SVM is having bigger margin.

Support Vectors

Support vectors are the data points which are very close to the hyperplane. Margin of the hyperplane is decided by the support vectors in the environment. Every data point in the environment is represented by vector. If the data environment is three-dimensional environment then each vector contains three elements. The values in the vector are directly proportional to the dimension of the data environment.

Linear SVM

Linear SVM is used to do the classification on linearly separable data. The data is categorised into two types, the first one is linearly separable, and the second one is linearly inseparable. Based on the nature of the data distribution it is classified into linearly separable and linearly inseparable. Linear SVM is applied on the data which is linearly separable. In this the hyperplane divides the data linearly into two classes with respect to binary classification. If dimension is two then hyperplane is the line. If the number of dimension is three the hyperplane is plane. In the case of multi-dimensional environment the hyperplane is segregating the classes [23].

Nonlinear SVM

All the time the data is not convenient to separate it linearly. If the data is not ready to linearly separable, it should be converted into linearly separable data. In order to apply the SVM the linearly inseparable data should be converted into linearly separable data [24].

Increasing the dimension is the one way to make linearly inseparable data into linearly separable data. Kernel trick in SVM is used to make linearly inseparable data into linearly separable data. x is considered as the original independent variable. $\emptyset(x)$ is the independent variable after applying the kernel tricks [25].

Deduction of DDoS Through SVM: (SVMBD)

In the proposed method SVMBD is detecting the DDoS attack based on the SVM algorithm. The dataset which is used for the training and model creation has been downloaded from Canadian Institute for Cyber Security. The Canadian Institute for Cyber Security (CIC) is providing the data in order to enhance the research in cyber security. The data, which is used, contains more than 10 k tuples including both the classes. The dataset contains two classes that are normal traffic and intrusion traffic. The dataset contains 78 attributes; it is a high-dimensional dataset. The data undergone the pre-processing techniques removes the missing values and changes the values into finite values.

The attributes of the dataset include destination port, flow duration, total forward packets, total length of forward packets, total length of the backward packet, etc. The model is getting trained with the dataset and tested with the unknown data. The quality and the performance of the model is tested based on the well-known metrics accuracy, precision, recall and $F1$-score. Figure 15.1 represents the flow diagram of the proposed SVMBD.

SVM-Based Model Creation

After pre-processing the dataset with 10,741 tuples and 78 attributes, it is splitted into X and Y. X is stated as independent variable, and Y is stated as dependent variable. Here the number of attributes in the independent variable is 77. Y is stated as dependent variable. Here the attribute ‘Label’ is considered as dependent variable. Based on the list of X, the Y is going to be identified.

Training the Model

Once the entire dataset is divided into independent variable and dependent variable and Y. Both X and Y are splitted into training data and testing data. The entire X is splitted into X-train and X-test. The same way entire Y is splitted into Y-train, Y-test. The purpose of dividing the entire data into training set and testing set is to evaluate the model.

The model should be evaluated based on the unforeseen data. The 15% of testing data taken into the consideration for evaluating the model. This 15% of data is unknown to the model because the data is not present in the training of the model.

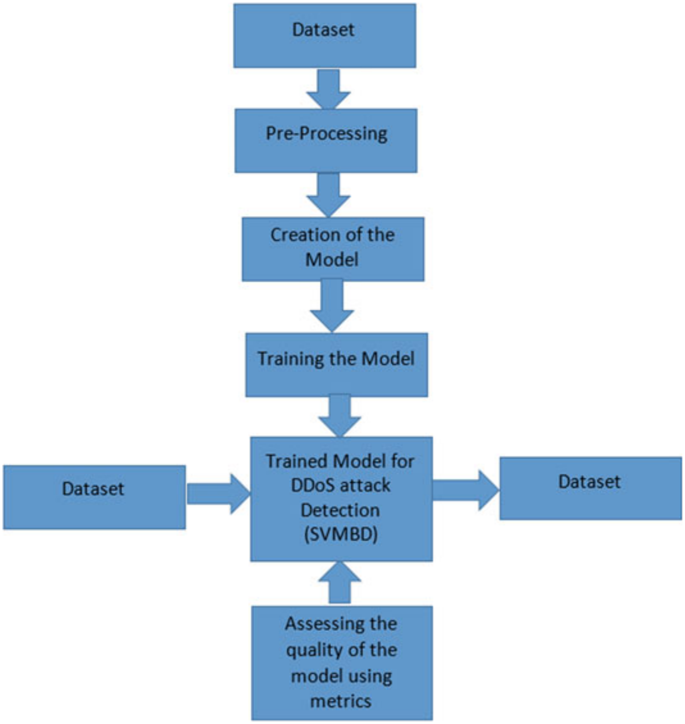

Dataset
Pre-Processing
Creation of the Model
Training the Model
Dataset
Trained Model for DDoS attack Detection (SVMBD)
Dataset
Assessing the quality of the model using metrics

Fig. 15.1 SVM-based DDoS attack detection scheme (SVMBD)

The model is getting trained with X-train and Y-train datasets. X-train and Y-train datasets is 85% of the original dataset.

Testing the Model

About 85% of the original data is taken into the consideration for training purpose. Remaining 15% of the data is allocated for testing. It is used to assess the quality of the model. Testing data is used to validate the model [26].

If the model is tested by the training data itself then it is not an effective way of testing. In order to enable the quality testing unforeseen data should be used. The model's performance is measured in an effective way by validating the model using unforeseen test data.

Table 15.1 Performance evaluation before scaling

Accuracy	Precision	Recall	$F1$-score
0.85	0.86	0.86	0.86

Assessing the Quality of the Model Using Metrics

The model's performance is measured by comparing the Y-test data and Y-pred data. The metrics accuracy, precision, recall and $f1$-score are used to measure the quality of the model.

Accuracy is the main metric to access the quality of the classification model.

$$\text{Accuracy} = \frac{\text{True positive} + \text{True Negative}}{\text{Total number of Predictions}}$$

Accuracy of our proposed SVMBD algorithm is 85% (before scaling). It is stated in Table 15.1. Precision is another one metric which is used to access the quality of the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

TP true positive
FP false positive.

Precision of proposed SVMBD algorithm is 86% (before scaling).

Recall is another metric which is conveying the proportion of correctly classified positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall of our proposed SVMBD algorithm is 86% (before scaling).

$F1$-score is also quite famous metrics which is assessing the quality of the machine learning model.

$$F1\text{- score} = 2*\left(\frac{\text{Precision*Recall}}{\text{Precision} + \text{Recall}}\right)$$

$F1$-score of the proposed model SVMBD this 86% (before scaling).

Scaling is the mechanism which is related to pre-processing, which enhances the quality of data. Scaling makes the data more convenient to training of the machine learning model. Scaling operation performed on the data which is used to detect DDoS. After the scaling, performance of the model is assessed again using the metrics accuracy, precision, recall and $F1$-score. Improved performance achieved with 97% accuracy. It is stated in Table 15.2. Table 15.3 and Fig. 15.5 stated the comparison

Table 15.2 Performance evaluation after scaling

Accuracy	Precision	Recall	$F1$-score
0.97	0.95	0.98	0.96

Table 15.3 Performance before scaling versus after scaling

	Before scaling	After scaling
Accuracy	0.85	0.97
Precision	0.86	0.95
Recall	0.86	0.98
$F1$-score	0.86	0.96

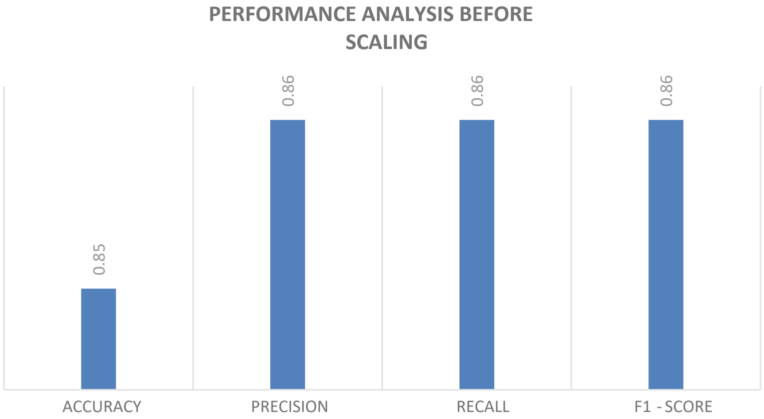


Fig. 15.2 Performance analysis before scaling

before scaling and after scaling. Figure 15.4 represents the detailed performance analysis after scaling. Figure 15.2 analysis chart represents the performance analysis associated with various metrics accuracy, precision, recall and score before scaling. Figure 15.3 analysis chart represents the performance analysis associated with various metrics accuracy, precision, recall and score after scaling.

Conclusion

Establishing and enhancing the cyber security in all the field is highly essential in today's digital era. This book chapter focused the distributed denial of service attack detection. The impact of distributed denial of service attack is more, the attack is creating more data loss. It creates more reputation problem for the organisation since the service given by the organisation stops due to the attack. The machine learning technique has been utilised here in order to detect the DDoS attack.

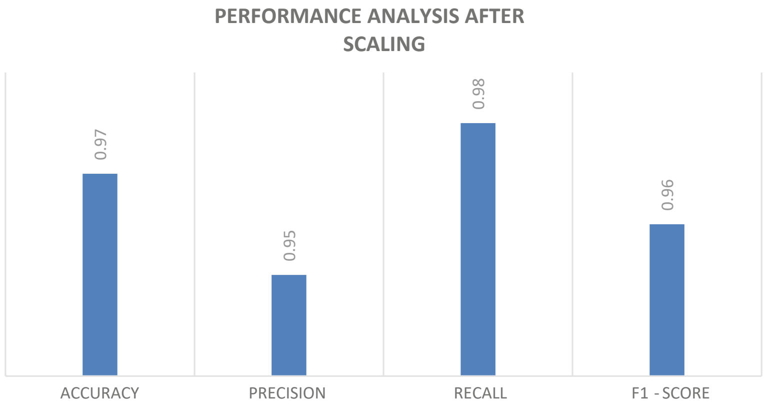
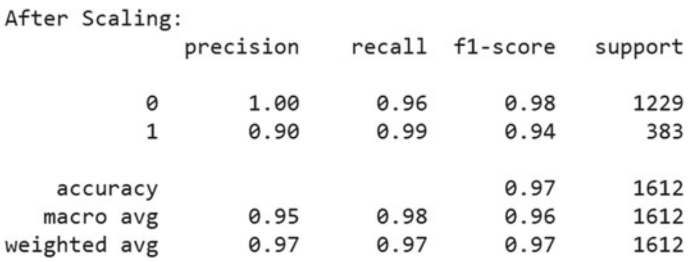


Fig. 15.3 Performance analysis after scaling



```
After Scaling:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      1229
           1       0.90      0.99      0.94       383

    accuracy                           0.97      1612
   macro avg       0.95      0.98      0.96      1612
weighted avg       0.97      0.97      0.97      1612
```

Fig. 15.4 Detailed performance analysis after scaling

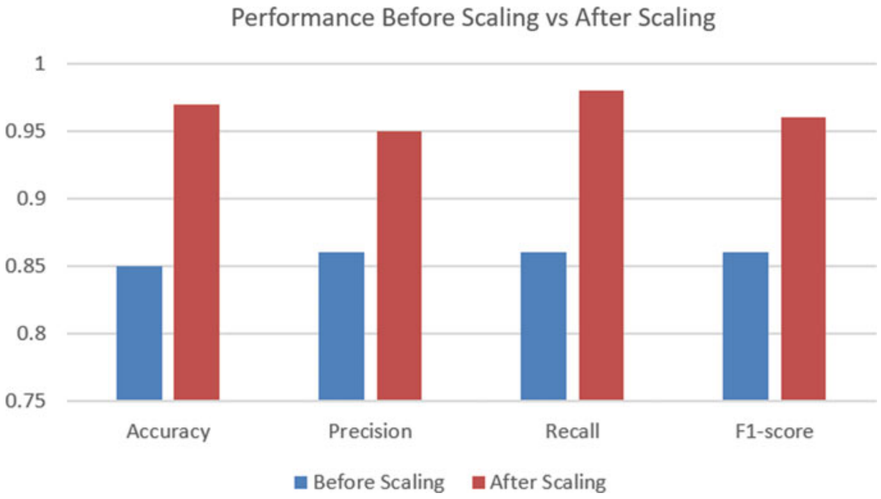


Fig. 15.5 Performance before scaling versus after scaling

Support vector machine-based machine learning model has been created and trained with the dataset downloaded from Canadian Institute for Cyber Security. The machine got trained with more the 10 K tuples which are having both the labels intrusion and genuine traffic. Once the model got trained the performance of the model is measured using the well-known classification metrics. Performance is measured before scaling the data as well as after the scaling. The model's performance has been proved with 97% of accuracy. The model is now ready to receive the traffic and classify whether it is an intuition or normal traffic.

References

1. Buil-Gil, D., Miró-Llinares, F., Moneva, A., Kemp, S., Díaz-Castaño, N.: Cybercrime and shifts in opportunities during COVID-19: a preliminary analysis in the UK. Eur. Soc. **23**(sup1), S47–S59 (2021)
2. Monteith, S., Bauer, M., Alda, M., Geddes, J., Whybrow, P.C., Glenn, T.: Increasing cybercrime since the pandemic: concerns for psychiatry. Curr. Psychiatry Rep. **23**, 1–9 (2021)
3. Deshmukh, R.V., Devadkar, K.K.: Understanding DDoS attack and its effect in cloud environment. Procedia Comput. Sci. **49**, 202–210 (2015)
4. Sadre, R., Sperotto, A., Pras, A.: The effects of DDoS attacks on flow monitoring applications. In: IEEE Network Operations and Management Symposium, pp. 269–277. IEEE (2012)
5. Khanzode, K.C.A., Sarode, R.D.: Advantages and disadvantages of artificial intelligence and machine learning: a literature review. Int. J. Libr. Inf. Sci. (IJLIS) **9**(1), 3 (2020)
6. Attaran, M., Deb, P.: Machine learning: the new 'big thing' for competitive advantage. Int. J. Knowl. Eng. Data Min. **5**(4), 277–305 (2018)
7. Yuan, R., Li, Z., Guan, X., Xu, L.: An SVM-based machine learning method for accurate internet traffic classification. Inf. Syst. Front. **12**, 149–156 (2010)
8. Shetty, S., Rao, Y.S.: SVM based machine learning approach to identify Parkinson's disease using gait analysis. In: International Conference on Inventive Computation Technologies (ICICT), vol. 2, pp. 1–5. IEEE (2016)
9. Mihoub, A., Fredj, O.B., Cheikhrouhou, O., Derhab, A., Krichen, M.: Denial of service attack detection and mitigation for internet of things using looking-back-enabled machine learning techniques. Comput. Electr. Eng. **98**, 107716 (2022)
10. Liu, G., Zhao, H., Fan, F., Liu, G., Xu, Q., Nazir, S.: An enhanced intrusion detection model based on improved kNN in WSNs. Sensors **22**(4), 1407 (2022)
11. Mahajan, N., Chauhan, A., Kumar, H., Kaushal, S., Sangaiah, A.K.: A deep learning approach to detection and mitigation of distributed denial of service attacks in high availability intelligent transport systems. Mobile Netw. Appl. 1–21 (2022)
12. Tonkal, Ö., Polat, H., Başaran, E., Cömert, Z., Kocaoğlu, R.: Machine learning approach equipped with neighbourhood component analysis for DDoS attack detection in software-defined networking. Electronics **10**(11), 1227 (2021)
13. Kumar, P.A.R., Selvakumar, S.: Distributed denial of service attack detection using an ensemble of neural classifier. Comput. Commun. **34**(11), 1328–1341 (2011)
14. Zekri, M., El Kafhali, S., Aboutabit, N., Saadi, Y.: DDoS attack detection using machine learning techniques in cloud computing environments. In: 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), pp. 1–7. IEEE (2017)
15. He, Z., Zhang, T., Lee, R.B.: Machine learning based DDoS attack detection from source side in cloud. In: IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 114–120. IEEE (2017)

16. de Miranda Rios, V., Inácio, P.R., Magoni, D., Freire, M.M.: Detection of reduction-of-quality DDoS attacks using Fuzzy Logic and machine learning algorithms. Comput. Netw. **186**, 107792 (2021)
17. Aamir, M., Zaidi, S.M.A.: Clustering based semi-supervised machine learning for DDoS attack classification. J. King Saud Univ.-Comput. Inf. Sci. **33**(4), 436–446 (2021)
18. Aysa, M.H., Ibrahim, A.A., Mohammed, A.H.: IoT DDoS attack detection using machine learning. In: 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–7. IEEE (2020)
19. Yuan, J., Mills, K.: Monitoring the macroscopic effect of DDoS flooding attacks. IEEE Trans. Dependable Secure Comput. **2**(4), 324–335 (2005)
20. Srivastava, A., Gupta, B.B., Tyagi, A., Sharma, A., Mishra, A.: A recent survey on DDoS attacks and defense mechanisms. In: Advances in Parallel Distributed Computing: First International Conference on Parallel, Distributed Computing Technologies and Applications, PDCTA 2011, Tirunelveli, India, September 23–25, 2011. Proceedings, pp. 570–580. Springer Berlin Heidelberg (2011)
21. Bogdanoski, M., Suminoski, T., Risteski, A.: Analysis of the SYN flood DoS attack. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) **5**(8), 1–11 (2013)
22. Noble, W.S.: What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006)
23. Joachims, T.: Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226 (2006)
24. Suykens, J.A.: Nonlinear modelling and support vector machines. In IMTC 2001 Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No. 01CH 37188), vol. 1, pp. 287–294. IEEE (2001)
25. Hofmann, M.: Support vector machines-kernels and the kernel trick. Notes **26**(3), 1–16 (2006)
26. Erickson, B.J., Kitamura, F.: Magician's corner: 9. Performance metrics for machine learning models. Radiol.: Artif. Intell. **3**(3) (2021)