Prediction of DDoS Attacks Using Machine Learning Algorithms Based on Classification Technique

Anupama Mishra[1(✉)] and Deepesh Rawat[2]

[1] Computer Science and Engineering, Himalayan School of Science and Technology, Swami Rama Himalayan University, Dehradun, India
anupamamishra@srhu.edu.in

[2] Electronics & Communication Engineering, Himalayan School of Science and Technology, Swami Rama Himalayan University, Dehradun, India

Abstract. Distributed denial of service attacks often know as network threat is a severe threat, and are a type of cyber-attack that are directed at a particular system or network in an effort to make that system or net-work out of reach and unusable for a period of time. The improved detection of a wide variety of dis-tributed denial-of-service (DDoS) cyber threats by utilizing advanced algorithms and a higher level of accuracy while maintaining a manageable level of computational cost has consequently emerged as the utmost essential part of detecting DDoS in today's world. The DDoS attack that has been launched against the targeted network or system must be determined in view of defending the machines in a net-work that has been targeted. In this paper, a number of ensemble classification techniques are dis-cussed, which combine the performance of various algorithms to improve overall performance. Using many performance metrics such as a receiver operating characteristics (ROC) curves, precision, accuracy, recall and F1 scores, we present and analyzed the performance of algorithms used in our proposed approach.

Keywords: Distributed denial of service attack · Machine learning · Random forest · Naïve Bayes · Decision tree

1 Introduction

A denial-of-service attack, according to the World Wide Web security question, is one that is "de-signed to prevent a computer or network from providing normal services [1]." The rapid expansion of Internet has resulted the specific type of DoS attack development that has proven to be extremely effective and difficult to defend against the distributed DoS attack. This attacks do not originate from a single source, but rather from a number of spoofed sources that use a variety of attack types in a coordinated effort. Fake Internet Protocol (IP) addresses, as opposed to real ones, are used to identify computers that are either unwitting accomplices or that the attacker has control over. Attackers are able to coordinate deadly attacks on multiple targets at the same time using their own resources

R. Mehra et al. (Eds.): ICCISC 2022, CCIS 1672, pp. 39–50, 2022.
https://doi.org/10.1007/978-3-031-22915-2_4

and the resources of their “zombies,” resulting in greater damage in a shorter amount of time than they could have done otherwise [2].

A distributed denial of service which is also a cyberattack can bring websites, servers, and other online services to a crawl. The perpetrator uses multiple computers and devices to send fraudulent requests to a server, making it appear that the server is being attacked by a large number of people. The term is some-times used interchangeably with the term “denial of service attack”, but “DDOS” refers specifically to an attack that uses multiple sources to flood a target with requests. Some DDOS attacks involve the use of botnets, which are networks of compromised computers and devices that have been malware-infected without the users’ knowledge [3].

DDoS is a form of cyberattack where a network is attacked with heavy traffic that’s create a problem for users to access a website or service. This type of attack is often used as a tactic to make a target site or service appear overwhelmed or unreliable to users. DDoS attacks can also be used to force a website or service to a user to a specific location, where it is under the control of the attacker. DDoS attacks are often used for online harassment, for example, when a website is under constant attack and can’t improve its performance or stay online [4]. DDoS is a method of disrupting a system in a network by sending a large amount of data to a server or system from multiple different sources. The result is that the system or network is not able to handle the load and ultimately crashes. This makes the system or net-work unavailable to its intended use [5].

DDoS is a form of cyberattack where multiple hosts are used to bombard a web server or other net-work target with data, often using a botnet or other network of malware-infected computers, until the target’s resources are consumed and it is rendered inaccessible to legitimate users. Unlike a traditional DoS attack, where a single host is used to flood a target’s resources and crash their services, a DDoS at-tack is much more complicated and powerful. DDoS attacks can be extremely difficult to stop due to their decentralized nature. DDoS attacks can be carried out with a handful of hosts or even a single host, making them much more difficult to detect and investigate [6]. A denial of service refers to a situation in which a service or resource is utilized to the point of being rendered unusable or inaccessible to other users. These are often seen in the context of online gaming servers, but can also affect online banking or shopping services.

A distributed denial of service occurs when a single device or user is able to bring down a server or network resource. This can be accomplished by distributing a request across a network, such as when a single computer is used to send an entire file to a website. Figure 1 depicts how this attack [3] had a significant impact on the telecommunications industry [7]. Current and former employees of several tech companies are accusing Amazon, Facebook, Google and Microsoft of failing to protect their employees from the COVID-19 pandemic, with some accusing the companies of endangering their workers’ safety. We’ve seen this story before. In the wake of a string of high-profile layoffs, employees have accused the companies of not doing enough to protect their workers’ health. But for companies as large as Amazon and Facebook, the risks of a pandemic are remote. Therefore, the developed technology helps in this field to defend our system and network from the threat. Machine learning is one of the technologies which is being used for defensive mechanisms in many applications.

Following are the sections that provides an outline for the paper: Sect. 2 presents the related work based on existing defensive mechanisms, our proposed work is discussed in Sect. 3. Section 4 evaluated the research work, and Sect. 5 brings the research to its conclusion.

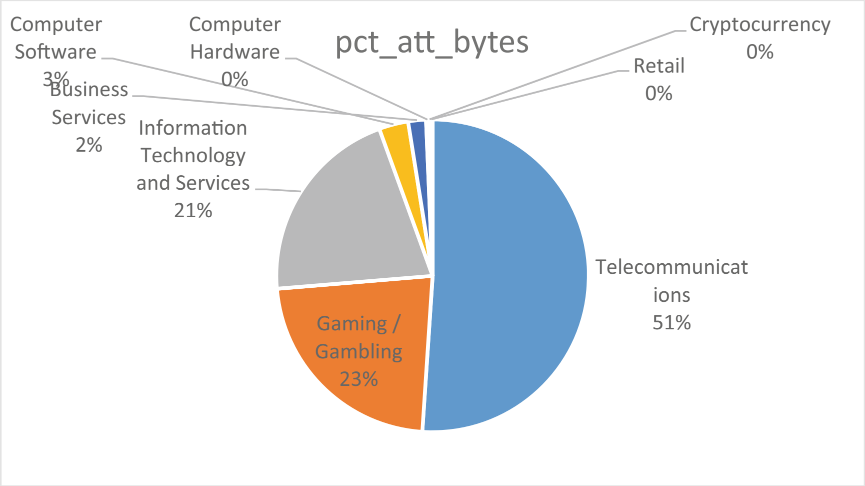

Fig. 1. Impact of distributed denial of service attacks

2 Related Work

Machine learning allows to compare models with different features, which can help to choose the one that is best for the use of applications. Specifically, we can use a machine learning model to identify which types of data are most predictive of the outcome of interest, such as cyberattack.

A defensive approach based on low-cost was proposed by the authors in [6, 7]and focused in calculation of the entropy between benign traffic and DDoS attacks. As an additional suggestion, the authors Intensity reduction strategy for dealing with attacks has the following characteristics: The following are three advantages of using this methodology over other current methods: The first is that it has a high level of detection. In addition to having a reduces false alarm, also the capability of detecting small changes in the environment at a rapid pace along with the mitigate approach.

The authors [7] developed a solution for resolving authentication and security challenges connected to smart vessels in sea transport. An identity-based approach is used to authenticate the access for smart vessel and devices. But the method is limited to maritime transportation.

An auction with many attributes was proposed in [8] to mitigate distributed DDoS attacks on a net-work. A reputation-based detection approach was proposed, in which the minimal utility defines the user's reputation. A payment plan for normal users and another for fraudulent users, as well as an identifying mechanism are proposed in addition with the identification method. In this method, a greedy re-source for allocation strategy is used to ensure that resources are distributed appropriately among legitimate users. Differential payment systems are designed to penalize malevolent users that manipulate their offers in order to obtain the maximum possible share of limited resources.

The authors [9] describe an approach for detecting distributed denial of service attacks that makes use of Bayesian game theory. It is assumed that the service provider as well as legitimate users monitor the network in order to collect probabilistic information to ensure that another user is acting maliciously on their behalf or not. As a result of having this probabilistic knowledge, both the service provider and authorized users have the ability to alter their actions and replies in reaction to harmful activity on the network. The authors propose a Bayesian pricing [10] and auction approach for obtaining Bayesian Nash Equilibrium points in a variety of settings in which genuine consumers and service providers benefit from probabilistic knowledge. This is accomplished by taking into consideration the aforementioned assumptions and facts. In addition to this, a reputation evaluation and updating system is offered to determine a user's dependability based on factors such as the user's payment history and the amount of time spent participating in the platform (Table 1).

Table 1. Comparative table of existing work.

References	Techniques	Merits	Limitation
[10]	Used SDN (Software Defined Network	Detection Rate is high	Only work on Volume based DDoS
[11]	Worked on IBE (Identity Based Encryption)	Detection Rate is high	Overhead is high
[12]	Worked on IBS (Identity Based Signature) along with IBE	Detection rate is moderate	Overhead is high
[13]	Based on Boosting Algorithms	Detection rate is moderate	False alarms

3 Proposed Work

3.1 Approach

In our paper, we are primarily concerned with data preprocessing, selection of significant features [13], machine modelling through a classifier, and then finally prediction on testing dataset. After performance evaluation on results, the research work is concluded. The approach includes the following activities [14–19]:

Preparation of information: This phase is concerned with preparation of the data which is comprised of tasks helps on processing the raw data into a clean dataset.

If the raw data is not in a usable state at the time of completion, the type and order of activities may change, and Some of the unrelatable features may be removed. A few examples of the responsibilities involved are data cleansing, feature selection, and data transformation. In our work, Fig. 2 depicts the best 15 features by using extra tree classifiers. The selected features are: ACK Flag Count, Inbound, URG Flag Count, Destination IP, Source IP, Init_Win_bytes_forward, Timestamp, Flow ID, Pro-tocol, Min Packet Length, min_seg_size_forward, Destination Port, Max Packet Length, Average Packet Size, and Packet Length Std.

Modelling: In the modelling phase, modelling techniques are applied to the data. This is done in order to achieve the best possible performance by adjusting the parameters of the models in question.

As previously indicated, this step is closely tied to data preparation because modelling might disclose previously unknown data errors. Depending on the situation, the data preparation method can result in the employment of several models.

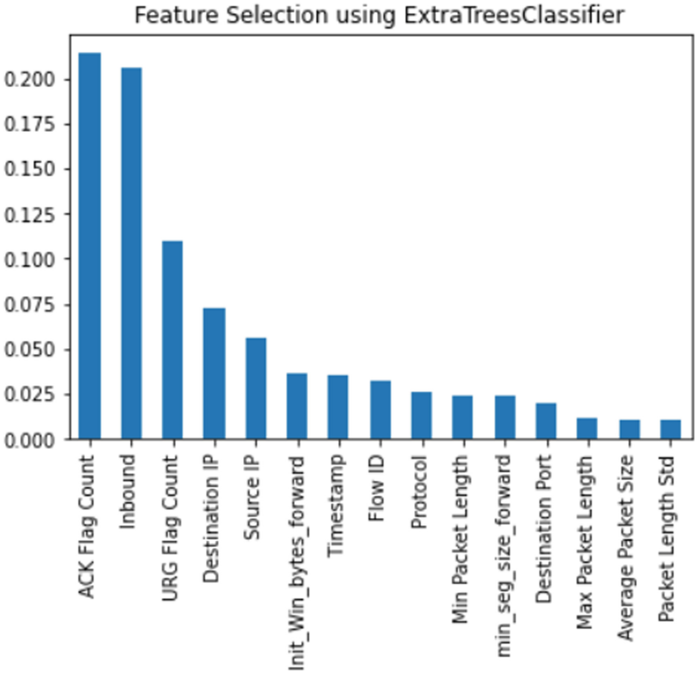

Fig. 2. 15 Best Features are selected by using extra trees classifiers

3.2 Modelling

In order to compare and contrast the datasets, three different supervised learning classifiers are selected based on a number of parameters [15–17], including the parametric models and nonparametric models, applications and use of algorithms have been discussed and used in previous work.

3.2.1 Random Forest

A machine learning algorithm that is used to classify a dataset into a specific category. It is a combination of many decision trees. The results of the decision trees are combined in a way that helps reduce the error rate of the classification. This is similar to how a forest is grown [20–22].

Random forest is a machine learning technique that groups examples together by their similarity, rather than grouping them by their distance to the target classification. This is often referred to as "many small decisions," as opposed to "one big decision," which is how other machine learning techniques work. This means that random forest will, on average, get more things right than other machine learning techniques. However, it is also more likely to get things wrong.

A machine learning technique that finds patterns in large numbers of variables. For example, in a medical diagnosis problem, instead of just looking at a patient's symptoms and lab results, a machine learning technique might look at millions of different combinations of symptoms and lab results to find patterns that help make a better diagnosis. In a financial prediction problem, instead of just looking at a stock's past performance, a machine learning technique might look at millions of past stock trans-actions to find patterns that help predict whether a stock will rise or fall. The same is true for any other problem: finding the right combination of input variables is critical for making accurate predictions [23].

The random forest technique is often more effective than traditional decision trees, because it is more likely to capture non-linear relationships in data.

3.2.2 Decision Tree

Decision trees are a machine learning technique that finds patterns in large amounts of data. In a traditional decision tree, the data is split into two groups: examples that should be classified as "yes" or "no" in the question being asked, and examples that should be classified as "yes" or "no" on their own. This is a two-class classification problem. For example, if the question being asked is "does this dog have fleas?" [24, 25].

A machine learning technique that uses decision trees to make a classification. Each decision tree is built on a subset of the original data. The random forest technique is often more effective than traditional decision trees, because it is more likely to capture non-linear relationships in data. Data is split into a number of groups, and each group is given a separate decision tree.

A machine learning technique that finds patterns in large numbers of variables. For example, in a medical diagnosis problem, instead of just looking at a patient's symptoms and lab results, a machine learning technique might look at millions of different combinations of symptoms and lab results to find pat-terns that help make a better diagnosis.

In a financial prediction problem, instead of just looking at a stock's past performance, a machine learning technique might look at millions of past stock transactions to find patterns that help predict whether a stock will rise or fall. The same is true for any other problem: finding the right combination of input variables is critical for making accurate predictions. It is a tree-based method that uses the outcome of a single decision tree as the input for the next tree in the forest. This helps reduce the error rate of the classification. This is similar to how a forest is grown [26].

Decision trees are one of the most basic machine learning techniques. They are a machine learning technique that finds patterns in large numbers of input variables. To build a decision tree, a machine learning technique starts by choosing a subset of the original data as the root node and then from there, the machine learning technique divides the original data into subsets, or nodes, based on some criteria, such as variable type or variable range.

3.2.3 Naïve Bayes

A machine learning technique that uses Bayes theorem to make predictions. The Naive Bayes ma-chine learning technique assumes that each input variable is independent of the others. For example, if a machine learning technique is trying to predict whether a person has a certain disease, Naive Bayes would assume that the presence or absence of a symptom has no effect on the prediction. This assumption turns out to be surprisingly accurate in many cases [27, 28].

An example of a machine learning technique that uses decision trees is the naïve Bayes classification machine learning technique and often used in text classification problems. In a text classification problem, the naïve Bayes machine learning technique uses decision trees to classify text into different categories. The naïve Bayes machine learning technique uses decision trees instead of other machine learning techniques because decision trees are able to capture non-linear relationships in data better than other machine learning techniques.

Naive Bayes [18, 19] used to make predictions. It is often one of the first machine learning techniques that people learn because it is easy to understand. The Naive Bayes machine learning technique is also often used as a simple baseline to compare the accuracy of other machine learning techniques. For ex-ample, if a machine learning technique is twice as accurate as the Naive Bayes machine learning technique, then it is likely that the first machine learning technique is a good one.

Naive Bayes is a machine learning technique that is used to make a prediction. It is a classification technique. For example, if the question being asked is "Will it rain today? a Naive Bayes technique might be used to predict whether it will rain today. It is a machine learning technique that finds patterns in large numbers of variables. It works on bayes theorem by combining the probability that a certain in-put will lead to a certain output with the probability that a different input will lead to the same output.

4 Result Analysis

We have the following metrics for evaluating the performance of classification machine learning [29–31]. Performance metrics for machine learning such as precision, recall, and f1-score are used to evaluate the quality of the model. These metrics can be used to compare the performance of different models and to evaluate the impact of different training regimes on model performance. Precision is a measure of the number of correctly classified examples. It is calculated as the ratio of the number of examples correctly classified by the model to the total number of examples in the training set.

The precision is the percentage of the time that an output was actually produced. The recall is the per-centage of the time that a specific output was actually identified. Table 2, Fig. 3 and 4 are used to show the results as precision, recall , f1 score and accuracy.

Table 2. Classification report of applied algorithms.

	Decision Tree			Random Forest			Naïve Bayes		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Benign	94	98	96	94	100	97	84	100	94
DDoS	100	100	100	100	100	100	100	99	100

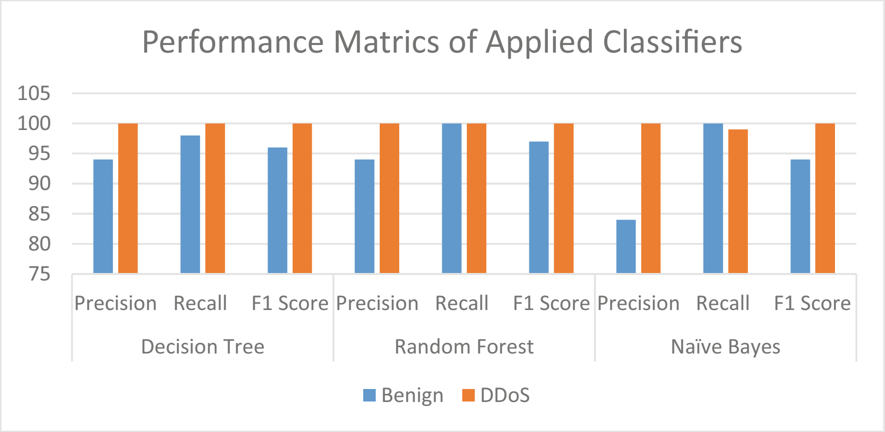

Fig. 3. Performance metrics of applied classifiers

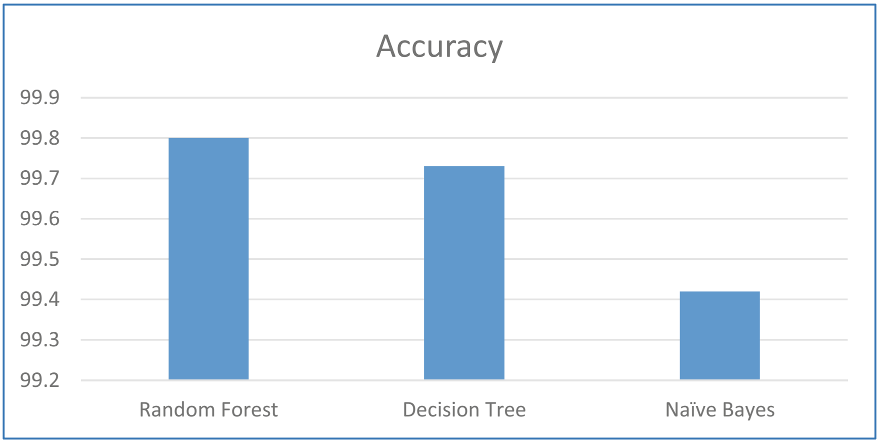

Fig. 4. Accuracy of applied classifiers

As shown in Figs. 5, 6 and 7, the results are discussed and analyzed. All three of them performed admirably when it came to classifying DDoS traffic. The results of the performance metric evaluations can be confirmed through the examination of the RoC Curve.

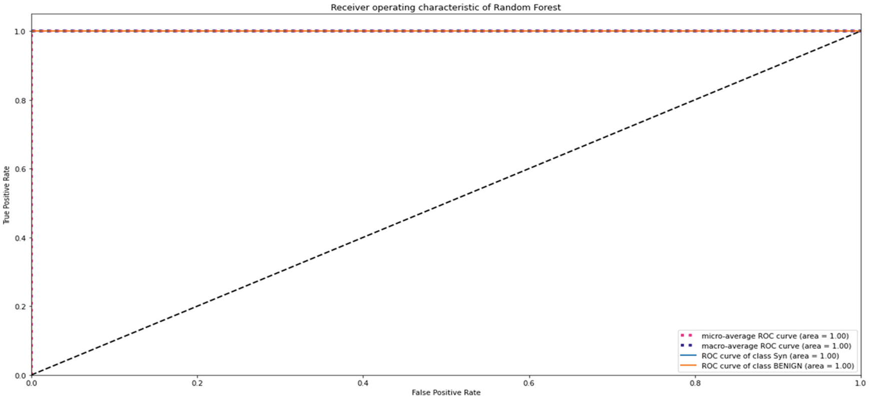

Fig. 5. RoC curve of Random Forest

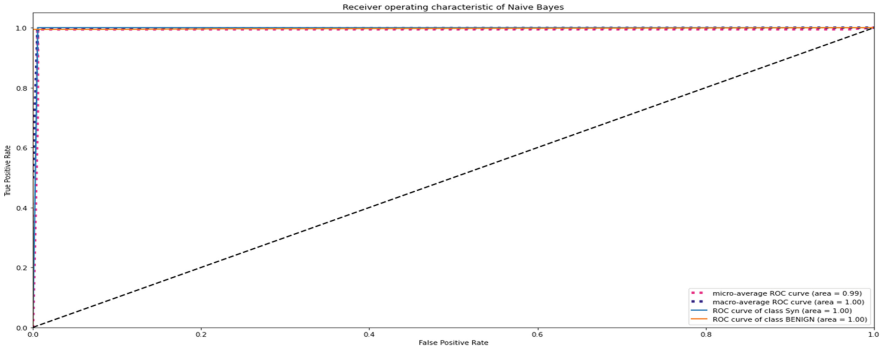

Fig. 6. RoC curve of Naïve Bayes

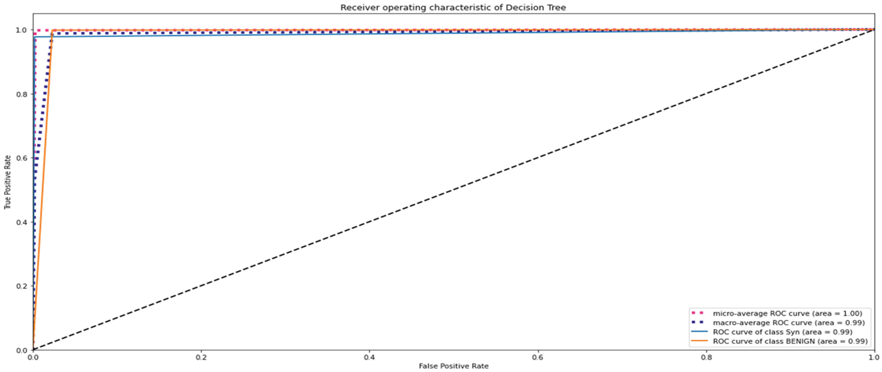

Fig. 7. RoC curve of Decision Tree

5 Conclusion

In this study, the datasets were classified into binary classification using machine learning classifiers, and each class was detected and validated properly. A comprehensive analysis of multiple machine learning algorithms was carried out for the purpose of identifying DDoS cyber threats, with the Random Forest with the highest accuracy score of 99.80 percent. The naive bayes method achieved 99.42 percent accuracy, while the decision tree achieved 99.75 percent accuracy in achieving the target. For future work, types of DDoS attacks can be targeted for classification and prediction in the future.

References

1. Badve, O.P., et al.: Taxonomy of DoS and DDoS attacks and desirable defense mechanism in a cloud computing environment. Neural Comput. Appl. **28**(12), 3655–3682 (2017)
2. Gupta, B.B., et al.: A comprehensive survey on DDoS attacks and recent defense mechanisms. In: Handbook of Research on Intrusion Detection Systems, pp. 186–218. IGI Global (2020)
3. https://radar.cloudflare.com/notebooks/ddos-2022-q1. Accessed 2 Apr 2022

4. Mishra, A., et al.: Security threats and recent countermeasures in cloud computing. Modern Principles, Practices, and Algorithms for Cloud Security, pp. 145–161. IGI Global (2020)
5. Mishra, A., Gupta, N.: Analysis of Cloud Computing Vulnerability against DDoS. In: International Conference on Innovative Sustainable Computational Technologies (CISCT), pp. 1–6. IEEE (2019)
6. Mishra, A., Gupta, N., Gupta, B.B.: Defense mechanisms against DDoS attack based on entropy in SDN-cloud using POX controller. Telecommun. Syst. **77**(1), 47–62 (2021). https://doi.org/10.1007/s11235-020-00747-w
7. Gaurav, A., et al.: Identity-based authentication mechanism for secure information sharing in the mari-time transport system. IEEE Trans. Intell. Transp. Syst. (2021)
8. Nguyen, G.N., et al.: Secure blockchain enabled cyber–physical systems in healthcare using deep belief network with ResNet model. J. Parallel Distrib. Comput. **153**, 150–160 (2021)
9. Zhou, Z., et al.: A fine-grained access control and security approach for intelligent vehicular transport in 6g communication system. IEEE Trans. Intell. Transp. Syst. (2021)
10. Dahiya, A., Gupta, B.B.: Multi attribute auction based incentivized solution against DDoS attacks. Comput. Secur. **92**, 101763 (2020)
11. Cvitić, I., et al.: Boosting-based DDoS detection in internet of things systems. IEEE Internet of Things J. **9**, 2109–2123 (2021)
12. Dahiya, A., et al.: A reputation score policy and Bayesian game theory based incentivised mechanism for DDoS attacks mitigation and cyber defense. Future Generation Computer Systems (2020)
13. Han, J., et al.: Data Mining: Concepts and Techniques. Elsevier (2011)
14. DDoS 2019 | Datasets | Research | Canadian Institute for Cybersecurity | UNB. Accessed 28 Apr 2022
15. Alzahrani, R.J., et al.: Security analysis of DDoS attacks using machine learning algorithms in networks traffic. Electronics **10**(23), 2919 (2021)
16. He, Z., Zhang, T., Lee, R.B.: Machine learning based DDoS attack detection from source side in cloud. In: Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, USA, 26–28 June 2017, pp. 114–120 (2017)
17. Aamir, M., et al.: DDoS attack detection with feature engineering and machine learning: the framework and performance evaluation. Int. J. Inf. Secur. **18**, 761–785 (2019)
18. Liu, Z., et al.: The prediction of DDoS attack by machine learning. In: Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), vol. 12167, pp. 681–686. SPIE (2022)
19. Zewdie, T.G., Girma, A.: An evaluation framework for machine learning methods in detection of DoS and DDoS intrusion. In: 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 115–121 (2022)
20. Sahoo, S., et al.: Multiple features based approach for automatic fake news detection on social net-works using deep learning. Appl. Soft Comput. **100**, 106983 (2021)
21. Cvitić, I., Peraković, D., Periša, M., Gupta, B.: Ensemble machine learning approach for classification of IoT devices in smart home. Int. J. Mach. Learn. Cybern. **12**(11), 3179–3202 (2021). https://doi.org/10.1007/s13042-020-01241-0
22. Gupta, B.B., et al.: Machine learning and smart card based two-factor authentication scheme for pre-serving anonymity in telecare medical information system (TMIS). Neural Computing and Applications, 1–26 (2021). https://doi.org/10.1007/s00521-021-06152-x
23. Yamaguchi, S., Gupta, B.: Malware threat in Internet of Things and its mitigation analysis. In: Research Anthology on Combating Denial-of-Service Attacks, pp. 371–387. IGI Global (2021)

24. Peraković, D., et al.: A Big Data and Deep Learning based Approach for DDoS Detection in Cloud Computing Environment. In: 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), pp. 287–290. IEEE (2021)
25. Dahiya, A., et al.: A PBNM and economic incentive-based defensive mechanism against DDoS at-tacks. Enterp. Inf. Syst. **16**(3), 406–426 (2022)
26. Dahiya, A., et al.: A reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense. Futur. Gener. Comput. Syst. **117**, 193–204 (2021)
27. Chartuni, A., et al.: Multi-classifier of DDoS attacks in computer networks built on neural networks. Appl. Sci. **11**(22), 10609 (2021)
28. Zhu, X., et al.: Prediction of rockhead using a hybrid N-XGBoost machine learning framework. J. Rock Mech. Geotech. Eng. **13**(6), 1231–1245 (2021)
29. Teles, G., Rodrigues, J.J., Rabêlo, R.A., Kozlov, S.A.: Comparative study of support vector machines and random forests machine learning algorithms on credit operation. Software: Practice and Experience **51**(12), 2492–2500 (2021)
30. Gaurav, A., et al.: A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. Enterprise Information Systems, 1–25 (2022)
31. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)