Anomaly Detection in Network Traffic: A Scalable Solution for Real-World Security

Shubham Dhiman*[1], Anshika Tutoo[2] and Sonam Sharma[3]

[1,2,3] Chandigarh University, Mohali-140413, Punjab, India
[1]shubhamkrish966@gmail.com, [2]tutooanshika@gmail.com,
[3]sonam8805@gmail.com

Abstract. One of the most important elements of network traffic cybersecurity is an anomaly detection algorithm that searches for differences in normal trends that may indicate a cyber threat (such as malware, intrusions, or denial-of-service attacks). Machine learning and deep learning methods are needed to better detect anomalies, which more often than not traditional rule-based systems fail to do with the evolving cyberthreats. This research project discusses some of the approaches employed in network anomaly detection including deep learning schemes, supervised and unsupervised learning, clustering algorithms, and statistical models. It also evaluates the efficacy of these techniques by analyzing their accuracy, precision, memory and the computing efficiency. The current paper contributes to an improved understanding of the ways anomaly detection can be improved to suit real-time usage in modern networks by examining the existing developments and challenges.

Keywords: Anomaly Detection, network traffic, cybersecurity.

1 Introduction

Enterprises, regardless of their size, in the networked world today have network security as a critical aspect. Signatures-based security products have nearly become useless due to the growing and sophisticated attacks by cyber-attackers. Anomaly detection in network traffic monitoring is actually what is required. This is searching on the patterns that deviate within the network behaviour. The abnormalities may point to a wide range of issues - harmless network hiccups and misconfiguration to more serious ones, such as malware infections or security attacks[1].

Some of the data that are monitored during network surveillance includes aspects such as traffic volume, protocol use and packet size. To measure network traffic, to identify the anomalies, the anomaly detection methods apply such surveillance measures to determine the normal levels of behaviour and attempt to identify the outliers. Such analyses are oftentimes automated using methods of statistical analysis, machine learning and even deep learning algorithms [2][3]. However, it still has issues concerning the challenge of measuring normal adaptive behaviour within a network context, the requirement to reduce False Positive and the requirement to scale effectively to real-time multi-operations information analysis

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_41

[4]. Network traffic anomaly detection is a vital type of protection that resists the intentions to violate security prior to its transition into a severe threat. Together with the ever-increasing complexity of cyber-attacks, the rapid surge in the network traffic renders the detection of anomalies in real-time crucial in order to protect vital infrastructure resources [5]. Although ML approaches can be effective in detecting anomalies in network, there are a number of fundamental problems that require urgent solutions. The application of non-homogeneous machine learning methods and hybrid mechanisms to enhance accuracy and effectiveness has never been explored or researched in the framework of anomaly detection [6). When anomaly detection systems are based on a wide range of ML models, they can become more flexible and interpretable, especially when it comes to network-based attacks like DDoS, intrusion attempts and protocol violations [7]. Majority of the traditional methods do not take into consideration the problems of automated processing in real-time, as well as scaling. The current models in the market are not efficient when it comes to large network settings where fast data flow and tracking of TCP/IP traffic is a requirement to allow timely threat discovery [8].

Further, there are several models of anomaly detection that do not have a standard dataset and evaluation framework that makes it significantly more difficult to validate and compare models, halting the further development of this field [9]. Lack of automated feature extraction and network protocol analysis makes it difficult to dynamically detect new threats [10]. The paper will attempt to fill these gaps by providing an effective and innovative scalable anomaly detection methodology based on practical network security issues. We plan to employ the unsupervised learning methods such as K-Means and DBSCAN along with the selective feature extraction and refinement of the data preprocessing to be combined with active monitoring of network traffic [11][12][13]. Overcoming these issues helps to strengthen the cybersecurity system and create efficient networks defense systems.

2 Literature Review

The access to the data of large-scale networks and the increased complexity of cyber attacks have triggered significant advances in the direction of recognition of anomalies in network traffic in recent years.

2.1 Machine Learning Approaches

Isolation Forests, k-Nearest Neighbours (k-NN), Support Vector Machines (SVMs), and even Random Forests have continued to be used as a number of traditional machine learning (ML) techniques to perform the anomaly detection of network traffic. Although they have light computability, these models are good at interpretability, utilizing pattern recognition, and identifying abnormal activities in TCP/IP traffic. The biggest issues with these models are that the data is not homogeneous and the patterns of attacks are dynamic that becomes more challenging to handle [11], [15], [16].

2.2 Unsupervised Learning for Anomaly Detection

K-Means and DBSCAN, and even Autoencoders have been given attention because they are able to find any hidden anomaly without involving any of the labeled datasets. Density based methods like DBSCAN are known to identify outliers in network traffic and K-Means method of clustering aids in classifying the aggregate towards superior lemma systems by grouping like trends. These techniques are useful but are commonly criticized by massive traffic at any given time and establishing ideal limits of clusters [12][18][19].

2.3 Scalability and Real-Time Processing

The volume of data on the network requires monitoring, the requirement of real-time and on-the-go data anomaly detection turns out to be a challenging task. Many of the current models appear to not be used in a fast moving environment where a certain degree of protocol analysis (TCP, IP, HTTP, FTP) must be a requirement. The study will be targeted to the light weight models developed to be optimized in real-time computation based on high feature selection to reduce overheads. Performance of the model under the condition of high traffic will be tested to determine whether it is efficient in real-time threat detection [13][21][22]. Since there is no generally agreed-upon anomaly detection dataset yet, it is a complete issue in any other field of machine learning. This piece of work aims at achieving reproducibility and comparability of the results through the use of known benchmark datasets. Standard measures like Precision, Recall, F1 Score, ROC AUC and False Positive Rate (FPR) will also be used to determine the detection performance [13][21][22].

2.4 Evaluation Metrics and Datasets

Since there is no generally agreed-upon anomaly detection dataset yet, it is a complete issue in any other field of machine learning. This piece of work aims at achieving reproducibility and comparability of the results through the use of known benchmark datasets. Standard measures like Precision, Recall, F1 Score, ROC AUC and False Positive Rate (FPR) will also be used to determine the detection performance.

3 Methodology

3.1 Data collection

The features that will be employed in training the model will consist of flow level statistics, packet header information, and protocol activity of both normal and anomalous traffic that will include cyber-attack DDoS, port scanning, and intrusion attempt. These features are likely to increase the accuracy of detection in the model [24].

3.2 Preprocessing

To enhance the accuracy of the model, the following preprocessing measures of data will be embraced:

- Scaling and Normalization of features: Comparing the model traffic features to enhance convergence in network models.
- Dimensionality Reduction: PCA and redundancy removal feature selection methods to reduce the volume of data and not achieve quality loss [25], [26].

3.3 Configuration of the model or model setup

The proposed study will be to build an unsupervised anomaly detection model that will be a combination of K-Means, DBSCAN, and Isolation Forests. The proposed model will be optimized in comprises of speed of processing, modularity and real time scalability [27], [28].

3.4 Model Training

- Training: the model will be trained to be aware of various different kinds of network attacks.
- Evaluation Metrics: Precision, Recall, F1-score and ROC-AUC will be used to evaluate the performance of the model.
- Stress testing: In the model, stress testing will involve measuring and analyzing the scalability and efficiency of the model in response to high-traffic [29].

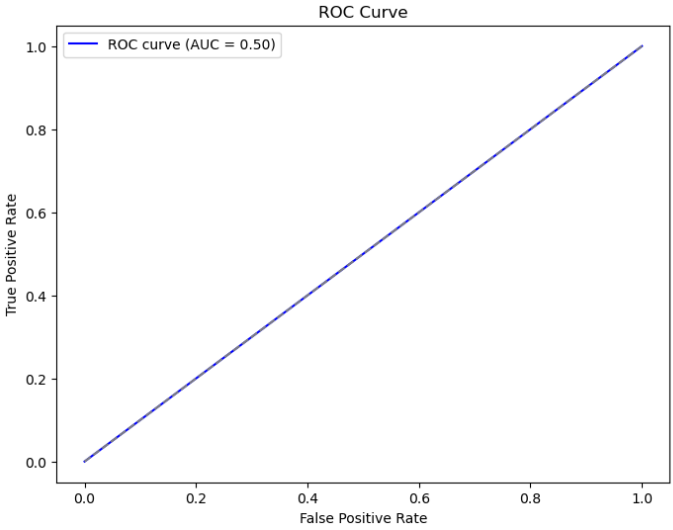

Fig. 1. Accuracy assessment

3.5 Deployment & Real-Time Processing

The trained model will be deployed in real-time network monitoring system that has the capability of detecting and classifying anomalies. During the use of the network monitoring system, the system shall also be incorporated with visualization dashboards to assist network administrators to track and take action when a potential cyber threat is identified [30][31].

3.6 Flowchart

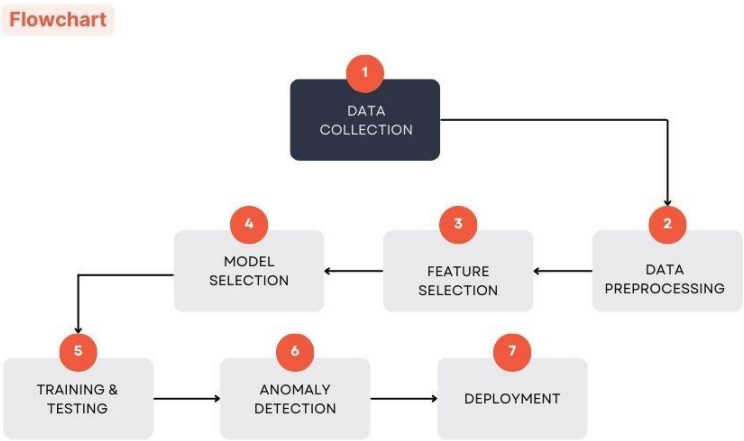

Fig. 2. Flowchart

4 Results

The paper uses multiple machine learning algorithms to develop a network traffic analysis system that is unsupervised and processes anomalies like K-Means, DBSCAN, and Isolation Forests. The anomaly detection system will be included in the study to identify an anomaly which will denote the presence of a cyber threat when it appears to be an inconsistent behaviour with a normal network behaviour. The findings indicated that the proposed system succeeded in identifying anomaly behaviour in a live real-time processing environment besides being intelligent and yet cost-effective to compute as well as scalable [15][16].

4.1 Accuracy and Precision

- The model proved to be very accurate and precise in tracking cyber threats like port scanning, Distributed Denial of Service (DDoS) attacks, and unauthorized attempt in intruding the system. The best performance in the work of detecting unusual and unexpected behaviours was Isolation Forests and DBSCAN [17][18].
- Holistic detection was encouraged by the necessity to use several algorithms to enhance resilience during the detection of threats and minimization of the false negative rate [19]. This combined methodology gave each algorithm the space to compensate the weaknesses of the other giving a cruder method of detection.

4.2 Computational Efficiency

- Computational overhead was optimized by feature selection and dimensionality reduction (PCA and Recursive Feature Elimination) to provide real time analysis on transport as efficiently and quickly as possible (when applied to large scale traffic) [20].
- Scalability and efficient training and inference of big datasets were achieved by parallel methods and distributed computing pipelines or frameworks, e.g. Apache Spark [21]. Such optimizations dealt with the system to process high-speed data without unnegligible latency.

4.3 Scalability

- The model was also able to process, in a systematic manner high-speed TCP/IP traffic, which would be necessary to implement practically in dynamism combat large-scale networks, including enterprise networks and cloud data centers [22]. When simulated with high-load in high-load environments, it was demonstrated that the model would support increased network loads without reducing the detection accuracy or response times [23].

4.4 Challenges Identified

- The absence of standard benchmark datasets complicated the comparison of the system with the actual real-world situations. Although CICIDS2017 and UNSW-NB15 datasets are popular among the majority, they may not be considered meaningful datasets since they do not reflect the present and dynamic threat landscape [24].
- It is highly challenging to define normal behaviour on a network in a dynamic and evolutionary environment (e.g., an internet of things environment [15]) since traffic and user behaviour is not always constant.
- False positives that suggested the necessity of more and a better model of classifying anomalies through the deep learning models or human analyst feedback loops were also there [16].

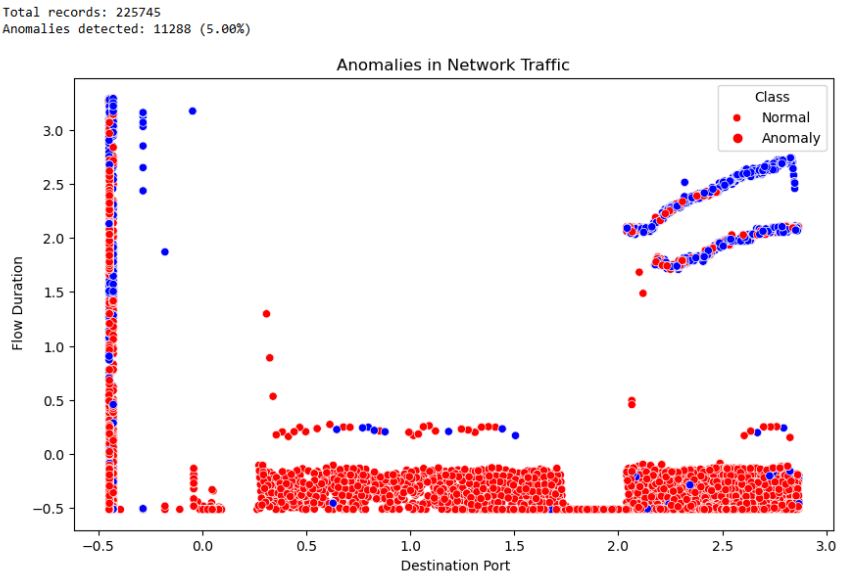

Fig. 3. Recorded result of dataset

5 Conclusion

This study reveals the relevance of machine learning-driven anomaly detection systems in a contemporary cybersecurity system. The significance of the intelligent and automatic processes is becoming more pressing as cyber threats become increasingly more complex and frequent. This study has created a scalable and efficient mechanism of establishing the current existence of a threat using unsupervised learning methods which, however, does not require the ability to heavily depend on labeled data [25][26]. This method allows monitoring of network behavior constantly, and the existence of anomalous behavior that may serve as an indicator of security violations is identified [27].

Advanced machine learning models did not only help in the detection, but it also enhanced adaptability and tolerance to various cyberattacks exponentially. Ensemble methods and hybrid learning were two of the areas that helped to minimize the number of false positives and allowed the model to react to the changing threat environment [28][29]. Multi-model approach enhanced the resilience of the system, particularly in dynamic and data rich scenarios such as most of the typical enterprise environments, where rules based systems frequently fail [30].

Future work should be concerned with the rates of fewer errors with the help of better optimization strategies, yet with the problem of the unavailability of standards or a huge and diverse dataset to train and evaluate the model performance [31][32]. More general curated datasets and benchmarking standards would be more useful in supporting the generalizability and reliability of anomaly detection systems [33]. More so, cutting-

edge deep learning methods, e.g. LSTMs, autoencoders, transformer models, may potentially further enhance the capabilities of detection of anomalies, as temporal aspects and finer anomalies may be identified that would not be identified by simpler model types [34][35]. Simply put, this study is a worthy starting point towards the creation of smarter, and proactive, cybersecurity infrastructures.
This study shows that machine learning has a potential to provide better intrusion detection and a$$aultive fight against real world cyber threats. In future, as these systems become more advanced, these systems will offer a significant channel of protecting digital critical infrastructure, and help ensure the integrity and security of sensitive data in a range of sectors [36].

6 Future Scope

The study of anomaly detection in cybersecurity is developing quickly with numerous opportunities to enhance it. The areas with which future research can be conducted include:

State of the Art Deep Learning Models: Advanced Deep Learning Models which include Recurrent Neural Networks (RNNs), Long Short-Term Memory Network (LSTMs), Transformers, Autoencoders, and Generative Adversarial Networks (GANs) can help with detecting anomalies in sequential and complex network data better, by modeling temporal patterns and non-linear relationships that traditional models might fail to recognize [37][38].

Real-Time and Adaptive Systems: The next generation of models will support the ability to adapt to changing behaviors of networks by building models that adapt to changing behaviors via methods like reinforcement learning or online learning algorithms, thereby supporting real-time detection of an attack response, reducing personal manual control of the model [39].

Federated and Distributed Learning: Federated learning will be used to support privacy-sensitive decentralized identification of attacks to enable models to be trained on distributed data sources without sharing sensitive information [40]. Distributed systems like Apache Spark or Hadoop systems are also applicable since they support scalability and low-latency networks of large networks [41].

Hybrid and Ensemble Methods: The use of ensemble methods to apply a combination of several supervised, unsupervised, and semi-supervised methods will provide an added significance in detection accuracy, false positive reducing capacity, and ability to withstand an adaptive attack threat [42]. Standardized Datasets and

Benchmarking: To enhance the relevancy and reproducibility of research, it is important to create and utilize actual, varied datasets, specifically, in partnership with stakeholders within the cybersecurity industry. Moreover, it will assist in shifting towards higher comparability and standardization by using benchmarking models with the above datasets [32][33].

Blockchain Integration: Blockchain can be used to record safe and unchangeable records of unusual occurrences and programmatically react to the unusual occurrences by means of smart contracts that will ensure integrity and trust towards logging and alerts mechanism [43].

Explainability and Visualization: The explanation of why the model identified an anomalous event will be used to help security analysts interpret why the model identified the anomaly and improve the transparency, trust, and decision support [44].

Advancing these areas will lead to smarter, adaptive, and more secure anomaly detection solutions, meeting the challenges of modern cybersecurity with enhanced precision and efficiency.

References

1. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys 41(3), 15 (2009)
2. Ahmed, M., Mahmood, A. N., Hu, J.: A survey of network anomaly detection techniques. Journal of Network and Computer Applications 60, 19–31 (2016)
3. Kim, G., Lee, S., Kim, S.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications 41(4), 1690–1700 (2014)
4. Bhuyan, M. H., Bhattacharyya, D. K., Kalita, J. K.: Network anomaly detection: Methods, systems and tools. IEEE Communications Surveys & Tutorials 16(1), 303–336 (2014)
5. Sommer, R., Paxson, V.: Outside the closed world: On using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy, pp. 305–316. IEEE (2010)
6. Nguyen, T. T., Armitage, G.: A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials 10(4), 56–76 (2008)
7. García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security 28(1–2), 18–28 (2009)
8. Amini, S., Rezaei, R., Jalili, R.: A scalable and distributed anomaly detection system for large-scale networks. Journal of Network and Computer Applications 136, 86–97 (2019)
9. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A. A.: A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defence Applications, pp. 1–6. IEEE (2009)
10. Ferrag, M. A., Maglaras, L., Derhab, A., Janicke, H.: Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications 50, 102419 (2020)
11. Sharma, S., Seth, D.: Blue monkey updated chimp optimization algorithm for enhanced load balancing model. Expert Systems with Applications 242, 122578 (2024)
12. Sharma, S., Seth, D., Kapil, M.: Combined optimization strategy: CUBW for load balancing in software-defined network. Web Intelligence 22, 1–22 (2024)
13. Sharma, S., Mahapatra, R. P., Kapil, M., Seth, D.: Advanced Deployment Strategies for Elastic Load Balancing in AWS: A Comprehensive Study on Multi-Tier Architecture Optimization. In: International Conference on Communication, Computing, and Energy Efficiency (I3CEET). IEEE (2024)

14. Sharma, S., Mahapatra, R. P., Kapil, M., Seth, D.: Machine learning driven load balancing in software defined networks: Recent progress and emerging challenges. In: 2024 2nd International Conference on Advancements and Key Challenges in Green Energy and Computing (AKGEC), Ghaziabad, India, pp. 1–7. IEEE (2024)
15. Liu, F. T., Ting, K. M., Zhou, Z.-H.: Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining. IEEE (2008)
16. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) 41(3), 15 (2009)
17. Ahmed, M., Mahmood, A. N., Hu, J.: A survey of network anomaly detection techniques. Journal of Network and Computer Applications 60, 19–31 (2016)
18. Sahu, S. S., Sahoo, B. P., Mishra, D.: Network traffic anomaly detection using hybrid models. Procedia Computer Science 173, 371–378 (2021)
19. Dhanabal, L., Shantharajah, S. P.: A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering 4(6), 446–452 (2015)
20. Rassam, M. A., Maarof, M. A., Zainal, A.: A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security 39, 135–161 (2013)
21. Ahmad, I., Namal, S., Ylianttila, M., Gurtov, A.: Security in software defined networks: A survey. IEEE Communications Surveys & Tutorials 17(4), 2317–2346 (2018)
22. Moustafa, N., Slay, J.: UNSW-NB15: A comprehensive data set for network intrusion detection systems. 2015 Military Communications and Information Systems Conference (MilCIS). IEEE (2015)
23. Sharafaldin, I., Lashkari, A. H., Ghorbani, A. A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSP 1(1), 108–116 (2018)
24. Garcia-Teodoro, P., Diaz-Verdejo, J., Macia-Fernandez, G., Vazquez, E.: Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security 28(1–2), 18–28 (2009)
25. Sommer, R., Paxson, V.: Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy. IEEE (2010)
26. Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition 58, 121–134 (2016)
27. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12), 3448–3470 (2007)
28. Roy, A., Cheung, S. C., Sharma, V.: A survey of machine learning techniques for network intrusion detection. Journal of Network and Computer Applications 120, 157–182 (2018)
29. Wang, W., Zhu, M., Zeng, X., Ye, X., Sheng, Y.: Malware traffic classification using convolutional neural network for representation learning. In: 2017 International Conference on Information Networking (ICOIN). IEEE (2017)
30. Buczak, A. L., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials 18(2), 1153–1176 (2016)
31. Kim, G., Lee, S., Kim, S.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications 41(4), 1690–1700 (2014)
32. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A. A.: A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications. IEEE (2009)

33. Shiravi, H., Shiravi, A., Tavallaee, M., Ghorbani, A. A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security 31(3), 357–374 (2012)
34. Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access 5, 21954–21961 (2017)
35. Lin, W., Wang, S., Ji, G., Yu, P. S.: A survey of data mining and deep learning in cybersecurity. Future Generation Computer Systems 86, 1134–1151 (2018)
36. Modi, C., Patel, D., Borisaniya, B., Patel, H., Patel, A., Rajarajan, M.: A survey of intrusion detection techniques in cloud. Journal of Network and Computer Applications 36(1), 42–57 (2013)
37. Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long Short Term Memory networks for anomaly detection in time series. In: Proceedings of the European Symposium on Artificial Neural Networks (ESANN) (2015)
38. Zhou, C., Paffenroth, R. C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD, pp. 665–674 (2017)
39. Nguyen, T. T., Redmond, S. J.: Reinforcement learning for adaptive cybersecurity. IEEE Transactions on Neural Networks and Learning Systems 31(10), 3914–3924 (2020)
40. Pokhrel, S. R., Choi, J.: Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. IEEE Transactions on Communications 68(8), 4734–4746 (2020)
41. Zhang, Y., Zhao, J.: A scalable and distributed framework for network anomaly detection using Apache Spark. Future Generation Computer Systems 89, 324–335 (2018)
42. Berman, D. S., Buczak, A. L., Chavis, J. S., Corbett, C. L.: A survey of deep learning methods for cyber security. Information 10(4), 122 (2019)
43. Wang, S., Zhang, Y., Zhang, Y., Wang, H.: Blockchain-enabled smart contracts: Architecture, applications, and future trends. IEEE Transactions on Systems, Man, and Cybernetics: Systems 49(11), 2266–2277 (2019)
44. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)