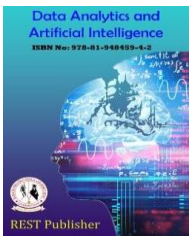
Data Analytics and Artificial Intelligence

Vol: 3(3), 2023

REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/



Cloud Intrusion Detection Based on Machine Learning Algorithm

S. Sujatha, M. Darshan, K. Hariharan, *D. S. Karthi, U. Niranjan

Adhiyamaan College of Engineering,Hosur Tamilnadu,India.

*Corresponding Author: dskarthi2706@gmail.com

ABSTRACT-*The diversification of wireless network traffic attack characteristics has led to the problems what traditional intrusiondetection technology with high false positive rate, low detection efficiency, and poor generalization ability. In order to enhance thesecurity and improve the detection ability of malicious intrusion behavior in a wireless network, It proposes a wireless network intrusion detection method based on convolutional neural network (CNN). First, the network traffic data is characterized and pre-processed, then modelled the network intrusion traffic data by CNN. The low-level intrusion traffic data is abstractly represented asadvanced features by CNN, which extracted autonomously the sample features, and optimizing network parameters to converge themodel. Finally, we conducted a sample test to detect the intrusion behavior of the network. During the training process, and the keyfeature information loss and parameter tuning difficulty are easily caused during the training process. It considers using the end-to end semi-supervised network training classifier of convolutional neural network (CNN), and the multi-layer feature of CNN to detect network, learn the sample features and discover the rules in the data training process to simplify the implementation process.*

keywords: *Attacks, machine learning, CNN, Decision Tree, prediction*

1. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process) is a field at the intersection of computer scienceand statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract informationfrom a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves databaseand data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post processing of discovered structures, visualization, and online updating. One of the main challenges in the security management of large-scale high-speed networks is the detection of suspicious anomalies in network traffic patterns due toDistributed Denial of Service (DDoS) attacks or worm propagation. A secure network must provide the following:

- Data confidentiality: Data that are being transferred through the network should be accessible only to those that have been properlyauthorized
- Data integrity: Data should maintain their integrity from the moment they are transmitted to the moment they are actually received.No corruption or data loss is accepted either from random events or malicious activity.
- Data availability: The network should be resilient to Denial-of-Service attacks.

What is IDSIntrusion detection is often used as another wall to protect computer systems. Intrusion detection (ID) is defined as theproblem of identifying individuals Who are using a computer system without authorization (i.e., `crackers') and those who have legitimate access to the system but are abusing their privileges. The goal of IDS is to detect malicious traffic. In order out going traffic. There are several approaches on the implementation of IDS.This technique is based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. Various different implementations of this technique have been proposed, based on the metrics used for measuring traffic profile deviation. This technique looks for patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that antivirus software operates. Data mining can help improve intrusion detection by addressing each and every one of the above-

mentioned problems.

Intrusion Detection in RBAC: A considerable effort has been recently devoted to the development of Database Management Systems (DBMS) which guarantee high assurance security and privacy. An important component of any strong security solution is represented by intrusion detection (ID) systems, able to detect anomalous behavior by applications and users. To date, however, there have been very few ID mechanisms specifically tailored to database systems. In this paper, we propose such a mechanism. The approach we propose to IDis based on mining database traces stored in log files. The result of the mining process is used to form user profiles that can model normal behavior and identify intruders. An additional feature of our approach is that we couple our mechanism with Role Based Access Control (RBAC). Under a RBAC system permissions are associated with roles, usually grouping several users, rather than with single users. We considered three models, of different granularity, to represent the log records appearing in the database log files. In that way, we managed to extract useful information from the log records regarding the access patterns of the users. Since role information was available in the log records, we used it for training a classifier that was then used as the basic component for our intrusion detection mechanism.

2. RELATED WORK

In the literature review section, we briefly explained all the related model and the closest rival to our proposed study. We studied thelatest research papers of the past two years for this research work and also Gozde Karatas et al. [2] proposed a machine learning approach for attacks classification. They used different machine learning algorithms and found that the KNN model is best for classification as compared to other research work. Nuno Martins et al. [1] proposed intrusion detection using machine learning approaches. They used the KDD dataset which is available on the UCI repository. They performed different supervised models to balance classification algorithm for better performance. In this work, a comparative study was proposed by the use of different classification algorithms and found good results in their work. Laurens D'hooge et al. [6] proposed a systematic review for malwaredetection using machine learning models. They compared different malware datasets from online resources as well as approaches for the dataset. They found that machine learning supervised models are very effective for malware detection to make a better decision in less time. Xianwei Gao et al. [7] proposed a comparative work for network traffic classification. They used machine learning classifiers for intrusion detection. The dataset is taken is CICIDS and KDD from the UCI repository. They found support vector machine SVM one of the best algorithms as compare to others. Tongtong Su et al. [3] proposed adaptive learning for intrusiondetection. They used the KDD dataset from an online repository. These models are Dtree, R-forest, and KNN classifiers. In this study, the authors found that Dtree and ensemble models are good for classification results. The overall accuracy of the proposed work is 85%. Kaiyuan Jiang et al. [4] proposed deep learning models for intrusion detection. The dataset is KDD, BAT-MC, BAT,and Recurrent neural network. The overall model's performance was very good.The accuracy is improved from 82% to 85% **.**

Arun Nagaraja et al. [5] proposed a hybrid model deep learning model for intrusion detection. They combined two deep learning models for the classification of CNN+ LSTM from the RNN model. The dataset was used in this work is KDD. They found an 85.14% average accuracy for the proposed. Yanqing Yang et al. [8] proposed a similarity-based approach for anomaly detection using machine learning. They used k mean cluster model for feature similarity detection and naïve Bayes model used for classification. Hui Jiang et al. [4] used an auto-encoder for labels and performed deep learning classification models on the KDD dataset. They found an 85% average accuracy for the proposed model [9]. SANA ULLAH JAN et al. [10] proposed a PSO-Xgboostmodel because it is higher than the overall classification accuracy alternative models, e.g. Xgboost, Random-Forest, Bagging, and Adaboost. First, establish a classification model based on Xgboost, and then use the adaptive search PSO optimal structure Xgboost.NSL-KDD, reference dataset used for the proposed model evaluation. Our results show that, PSO-Xgboost model of precision, recall, and macro-average average accuracy, especially in determining the U2R and R2L attacks. This work also provides an experimental basis for the application group NIDS in intelligence. Maede Zolanvari et al. [11] proposed a recurrent neural networkmodel for classification intrusion detection. They compared other deep learning models with RNN. Finally, they found RNN is thebest model for intrusion detection by using the KDD dataset. Yijing Chen et al. [12] proposed a domain that generates an algorithmfor botnet classification. It was a multiple classification problem. They used advanced deep learning LSTM for multiple classification problems. They found good results with 89% average accuracy for the proposed work. Larriva-Novo et al. [13] proposed two benchmark datasets, especially UGR16 and UNSW-NB15, and the most used dataset KDD99 were used for evaluation. The pre-processing strategy is evaluated based on scalar and standardization capabilities. These pre-processing modelsare applied through various attribute arrangements. These attributes depend on the classification of the four sets of highlights: basic associated highlights, content quality, fact

attributes, and finally the creation of highlights based on traffic and traffic quality basedon associated titles Collection. The goal of this inspection is to evaluate this arrangement by using different information pre-processing methods to obtain the most accurate model. Our proposition shows that by applying the order of organizing traffic and some preprocessing strategies, the accuracy can be improved by up to 45%. The pre-processing of a specific quality set takes into account more prominent accuracy, allowing AI calculations to effectively group these boundaries identified as potential attacks. Zeeshan Ahmad et al. [14] proposed a scientific classification approach, which depends on the well-known ML and DL processes included in the planning network-based IDS (NIDS) framework. By examining the quality and certain limitations of the proposed arrangements, an extensive review of the new clauses based on NIDS was conducted. By then, regarding the proposed technology,evaluation measurement, and dataset selection, the ongoing patterns and progress of NIDS based on ML and DL are given. Takingadvantage of the deficiencies of the proposed technology, in this paper, we put forward different exploration challenges and give suggestions. Muhammad Aamir et al. [15] proposed AI calculations were prepared and tried on the latest distributed benchmark dataset (CICIDS2017) to distinguish the best performance calculations on information, which contains the latest vectors of port checks and DDoS attacks. The permutation results show that every variation of isolation check and support vector machine (SVM)can provide high test accuracy, for example, more than 90%. According to the abstract scoring criteria cited in this article, 9 calculations from a bunch of AI tests received the most noteworthy score (highest) because they gave more than 85% representation(test) accuracy in 22 absolute calculations. In addition, this related investigation was also conducted to note that through the k-foldcross approval, the area under the curve (AUC) check of the receiver operating characteristic (ROC) curve, and the use of principal component analysis (PCA) for size reduction in preparation for AI execution model. When considering such late attacks, it was found that many checks on different AI calculations of the CICIDS2017 datasets were not sufficient for port checks and DDoS attackResults show that the proposed intrusion detection tool can effectively detect SQL-based attacks with no false positives and no overhead to the server.

Existing System: Intrusion detection technique can be dividing in two approach signature based approach and anomaly based approach.In signature based approach it stores existing attack pattern, and on each new transaction first it is matched with the existing attackpattern then if it succeeds then those transactions are declared as Malicious. In anomaly detection-based approach, it stores normal behaviour and if new transaction request is far more diverse from specified threshold value, then it is called an attack.

3. PROPOSED METHOD

In this method the end-to-end semi-supervised network training classifier of convolutional neural network (CNN).And the multi- layer This paper considers using the end-to-end semi-supervised network training classifier of convolutional neural network (CNN).And the multi-layer feature of CNN to detect network, learn the sample features and discover the rules in the data training process to simplify the implementation process.

To this end, we proposed an improved convolutional neural network (ICNN). Select Percentile and Decision Tree algorithm is used for feature selection and ranking. In this research, we design a framework for the DDoS attack classification and prediction based on the existing dataset that used machine learning methods. This framework involves the following main steps.

- The first step involves the selection of dataset for utilization.
- The second step involves the selection of tools and language.
- The third step involves data pre-processing techniques to handle irrelevant data from the dataset. In the fourth step featureextraction and label.
- Encoding is performed to convert symbolical data into numerical data.
- In the fifth step, the data splitting is performed into a train and test set for the model.

4. MATERIAL AND METHODS

Machine Learning: We proposed different algorithms for classification because the current algorithms have a lot of flaws and drawbacks. First, they cannot work with irrelevant values and feature engineering because the confusion matrix results are notaccurate. Some labeled results are zero that means algorithms do not work well. So, this is important to train the model precisely. Another problem is that some results show (Null) that means missing values also included in data that was not computed. Similarly,we need to justify existing algorithms with an advanced algorithm to find out the fastest and sufficient model. They also showed that random forest is not better than the KNN model because the result is less for the KNN model.

In Select percentile and Decision tree both are two different algorithms that can be used for different purposes. For example, selectpercentile is used for feature extraction and Decision tree is used for regression in time series data utilization. They used the modelfor intrusion detection. However, this is a very long and time-consuming process.

Therefore, it is very important to perform advanced machine learning techniques to model optimization that train the best model for highly accurate work. Here, in this paper, intrusiondetection is a classification problem. Therefore, it is a very serious problem to handle these implemented algorithms. In the last one,no such methodology is used for data mining to improve the quality of data. Among the machine learning techniques, CNN are powerful supervised learning models. Both are applicable and used for classification problems. The random forest algorithm is approximately 100 times faster than other algorithms and best working for classification problems.The application plane includes various applications to implement different functions, such as the anomaly detection application.The anomaly detection applicationis used to achieve three main functions: (1) data preprocessing, where the network traffic is transformed and standardized, (2) classifier training, where the CNN model used for feature extraction and classification detection is trained from the preprocessed network traffic, and (3) attack recognition,where the trained classifier is used to detect intrusion on the testing dataset or online network traffic We selected the KDD dataset that contains features' data about the attacks. The total numbers of rows and columns in the dataset. The dataset consists of different features about the attacks including an ID number.which presents medium of the network, label of the attacks,and attacks' cat which presents the severity of the attacks. All algorithms in artificial intelligence and machine learning employ input data to generate output results. The characteristics and features in this input dataset are in the form of structural columns. To operate well with the algorithm, all algorithms require data characteristics with certain characters. The basic aims of feature engineering are to provide an input dataset that is compatible with the criteria of machine learning and artificialintelligence models. As a result, we begin by converting all classified attributes into equivalent numerical labels. The second goal ad objective is to improve the performance of machine learning and artificial intelligence models. Feature Element scaling is a method of standardizing the existence of autonomous elements in the information within an appropriate range. The scaling is performed in the process of information pre-processing to deal with the magnitude or value or unit of height changes. If the component scaling is not completed, then the AI calculation will weigh greater mass, greater magnitude, and treat the more generalquality as a lower value, and rarely consider units with important values. There are two most ideal ways to apply the highlight zoom. The first is normalization, and the second is standardization. In normalization, your perception is taken away through all perceptualmethods, and when the parts are separated by the standard deviation, at this point, the perception is scaled. The attached recipe is used for the normalization strategy in AI. This is a very effective strategy to readjust the quality to achieve nothing but the same difference with one.

Xnew=(Xi-Xmean)/(standarddeviation)

In standardization, divide your perception by the basis of all perceptions, and then, at this point, subtract the smallest perception from the most extreme perception, and then perform highlight scaling at that point. This process re-adjusts the components or perceptions, and spreads the value somewhere within the scope of nothingness.

Xnew=(Xi-min(X))/(max(x)-min(x))

In our proposed work, we use the standard scalar element scaling method for element scaling. This is due to the fact that it is thebest strategy to use, most of the time, in including zooming.

5. METHODOLOGY

The main contribution is to generate the best model for data utilization, as well as, model optimization; and which performs best formodel learning. After getting the results, we performed performance measures in terms of precision, recall, and f1 score. In this research work, we used two well-known supervised learning models which are: CNN and RNN. Python language is considered a suitable programming language both for simulations and real-world programming. It is considered the most powerful high level language for model learning. . We used a jupyter notebook as a tool. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data can have missing values for a number of reasons such as observations that were not recorded and data corruption. The module includes two operations of numerical processing and normalization. The purpose of the module is to provide a standardized input data set for network training. Feature extraction is the second class of methods for dimension reduction. This function is useful for reducing the dimensionality of high-dimensional data. (ie you get lesscolumns). Then, a threshold is set to select sensitive features to construct an optimal feature vector. Through this work, many uselessand small influence features are pruned. The selected optimal features are used to train the underlying neural network. Neural network is used to train the data, which is resulted as optimal features. Neural network model is composed of 3 layers: the input, thehidden and the output layers. Algorithms utilized by neural network generally incorporate two phases: the forward propagation and the backward propagation. The forward propagation starts to work when it receives a positive propagation signal. The backward propagation is invoked when the model detects the occurrence of evident deviation between the calculation result of the output layerand the actual value. CNN is used to classify the data of intrusion detection. ICNN's back propagation process is to calculate the error vector backwards from the last layer. This reverse movement is based on the output of the loss function in the network.In orderto grasp the law that the loss function

changes with the weight and offset of the previous layer, the stochastic gradient descent method and the repeated use of the chain rule can obtain the desired expression result in reverse, so that the optimal parameter combination can be obtained. In the ICNN training process, as the number of network layers increases and the number of convolutionkernels increases, the feature information extracted from the network is also increasing. The network itself converges the model through continuous feature selection and parameter optimization, which reflects the effectiveness of ICNN's deep learning processin modeling intrusion detection.

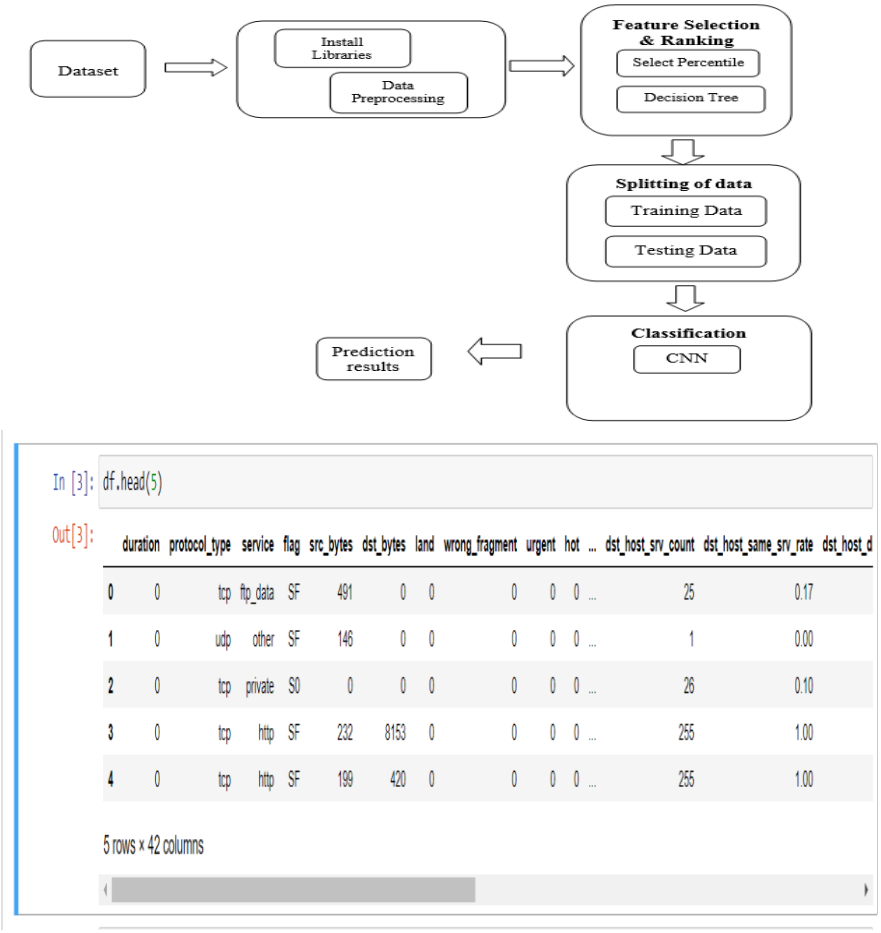


FIGURE 1. pre-processing

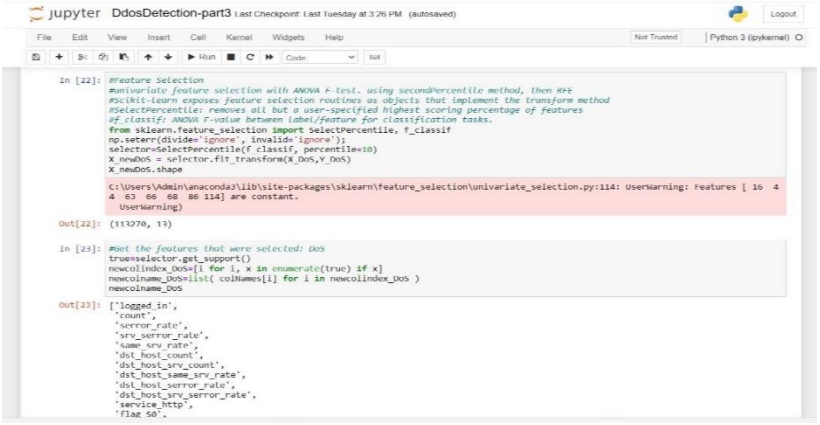


FIGURE 2. Feature Selection

6. RESULT AND DISCUSSION

In prediction, we used classification to generate the prediction results for future decisions. Then, we present the prediction results and outcomes,graphically. This section contains all the obtained results of our proposed models. This prediction showed Normal (7) = 11,147, Generic (6) = 5,526, Exploits (5) = 1,809, Fuzzers (4) =1,162, DoS

(3) = 971, Reconnaissance (2) =199, Analysis (1) = 163, Backdoor (0) = 112, attacks for future decision. As evident from the results, this prediction, as compared to the actual data, isapproximately accurate.

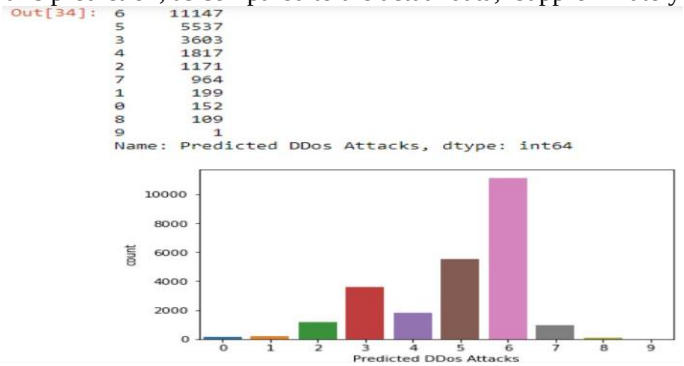


FIGURE 3

7. CONCLUSION

We proposed a complete systematic approach for detection of the Intrusion. First, we selected the KDD dataset that contains information about the attack. Then, Python and jupyter notebook were used to work on data wrangling. Secondly, we divided the dataset into two classes i.e. the dependent class and the independent class. Moreover, we normalized the dataset for the algorithm. After data normalization, we applied the proposed, supervised, machine learning approach. The model generated prediction and classification outcomes from the supervised algorithm. Then, we used CNN and Decision tree algorithms. Furthermore, we noted approximately 95% average Accuracy (AC) for the proposed model thatis enough good. Looking to the future, for functional applications, it is important to provide a more user-friendly, faster alternative to
deep learning calculations, and produce better results with a shorter burning time. It is important to work on unsupervised learning toward supervised learning for unlabeled and labeled datasets.

REFERENCES

[1]. N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, ‘‘Adversarial machine learning applied to intrusion and malware scenarios: A systematic review,’’ IEEE Access, vol. 8, pp. 35403–35419, 2020.

[2]. G. Karatas, O. Demir, and O. K. Sahingoz, ‘‘Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset,’’ IEEE Access, vol. 8, pp. 32150–32162, 2020.

[3]. T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, ‘‘BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset,’’ IEEE Access, vol. 8, pp. 29575–29585, 2020.

[4]. H. Jiang, Z. He, G. Ye, and H. Zhang, ‘‘Network intrusion detection based on PSO-xgboost model,’’ IEEE Access, vol. 8, pp. 58392–58401, 2020. 21452 VOLUME 10, 2022 Ismail et al.: Machine Learning-Based Classification and Prediction Technique for DDoS Attacks

[5]. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty, and V. S. Kiran, ‘‘Similarity based feature transformation for network anomaly detection,’’ IEEE Access, vol. 8, pp. 39184–39196, 2020.

[6]. L. D’hooge, T. Wauters, B. Volckaert, and F. De Turck, ‘‘Classification hardness for supervised learners on 20 years of intrusion detection data,’’ IEEE Access, vol. 7, pp. 167455–167469, 2019.

[7]. X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, ‘‘An adaptive ensemble machine learning model for intrusion detection,’’ IEEE Access, vol. 7, pp. 82512–82521, 2019.

[8]. Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, ‘‘Network intrusion detection based on supervised adversarial variational auto-encoder withregularization,’’ IEEE Access, vol. 8, pp. 42169–42184, 2020.

[9]. C. Liu, Y. Liu, Y. Yan, and J. Wang, ‘‘An intrusion detection model with hierarchical attention mechanism,’’ IEEE Access, vol. 8, pp. 67542– 67554, 2020.

[10]. S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, ‘‘Toward a lightweight intrusion detection system for the Internet of Things,’’ IEEE Access, vol. 7, pp. 42450–42471, 2019.

[11]. M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, ‘‘Machine learning-based network vulnerability analysis of industrial Internetof Things,’’ IEEE Internet Things J., vol. 6, no. 4, pp. 6822–6834, Aug. 2019.

[12]. Y. Chen, B. Pang, G. Shao, G. Wen, and X. Chen, ‘‘DGA-based botnet detection toward imbalanced multiclass learning,’’ Tsinghua Sci. Technol., vol. 26, no. 4, pp. 387–402, Aug. 2021.

[13]. X. Larriva-Novo, V. A. Villagrá, M. Vega-Barbas, D. Rivera, and M. S. Rodrigo, ‘‘An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets,’’ Sensors, vol. 21, no. 2, p. 656, Jan. 2021.

[14]. Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, ‘‘Net?work intrusion detection system: A systematic study of machine learning and deep learning approaches,’’ Trans. Emerg. Telecommun. Technol., vol. 32, no. 1, p. e4150, Jan. 2021.

[15]. M. Aamir, S. S. H. Rizvi, M. A. Hashmani, M. Zubair, and J. A. Usman, ‘‘Machine learning classification of port scanning and DDoS attacks: A comparative analysis,’’ Mehran Univ. Res. J. Eng. Technol., vol. 40, no. 1, pp. 215–229, Jan. 2021.