Penerbit UTHM

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm
e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

A Review on Cybersecurity based on Machine Learning and Deep Learning Algorithms

Alan Fuad Jahwar[1*], Siddeeq Y. Ameen[2]

[1]Akre Technical College, Duhok Polytechnic University,
Duhok, Kurdistan Region, IRAQ

[2]Quality Assurance Directorate Duhok Polytechnic University,
Duhok, Kurdistan Region, IRAQ

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2021.02.02.002
Received 22 May 2021; Accepted 01 October 2021; Available online 15 October 2021

Abstract: Machin learning (ML) and Deep Learning (DL) technique have been widely applied to areas like image processing and speech recognition so far. Likewise, ML and DL play a critical role in detecting and preventing in the field of cybersecurity. In this review, we focus on recent ML and DL algorithms that have been proposed in cybersecurity, network intrusion detection, malware detection. We also discuss key elements of cybersecurity, the main principle of information security, and the most common methods used to threaten cybersecurity. Finally, concluding remarks are discussed, including the possible research topics that can be taken into consideration to enhance various cyber security applications using DL and ML algorithms.

Keywords: Cybersecurity, Intrusion Detection (ID), Machine Learning (ML), Deep Learning (DL).

1. Introduction

Cyber security since the start of the computer network era, cyber security systems have been of great importance. Cyber security is a major field of investigation because all government-based, military, business, financial, and civil operations collect, process, and store enormous volumes of data on computers and other systems. Businesses must organize their activities through their entire information infrastructure in order to be on the defensive side of cyber security. Cyber security encompasses a variety of components, including network security, application security, mobile security, data security, and endpoint protection [1],[2].

Recently, cybersecurity has undergone massive transformation technological shifts and operations in a computer context. Machine Learning (ML) and Deep Learning (DL) are a subset of Artificial Intelligence (AI), and ML algorithms play an important part in several cyber security applications for early detection and the prediction of various attacks such as spam classification. spam classification [3], [4], spam filtering [5], fraud detection [6], [7], malware detection [8], [9], [10], [11], phishing [12], dark web or deep web sites [13], and intrusion detection [14]. Also, ML needs to use similar languages to solve cybersecurity problems. Python has become famous for its succinct, readable code and multitude of ML algorithm libraries, and a common ML language [15]. DL is an AI method that simulates the functioning of the human brain in data processing and in decision making. DL is a port of ML in AI, which has networks that can learn from unstructured or unstructured data without supervision. Often known as deep neural research or profound neural network [16]. Additionally, DL methods and their ensembles and hybrid approaches can be used to intelligently resolve a variety of cybersecurity problems, including intrusion detection, malware or botnet recognition, phishing, and predicting cyberattacks, such as DoS, fraud detection, or cyber-anomalies. Due to its increased accuracy, DL is advantageous for developing security models, especially when learning from large amounts of security datasets [17].

*Corresponding author: alan.jahwar@dpu.edu.krd

An intrusion Detection System (IDS) is a software-hardware computer hybrid that captures, analyzes, and detects any unwanted, suspicious, or malicious network traffic. IDS is classified into the following categories: Network Intrusion Detection System (NIDS); Host Intrusion Detection System (HIDS); Perimeter Intrusion Detection System (IDS); and VM-based Intrusion Detection System (IDS). However, there are two primary types of intrusion detection systems: NIDS and HIDS. NIDS tracks, captures and analyzes network traffic in order to detect malicious data contained within packets. Network IDS tracks both external and internal network traffic and also observes data traversing between devices within the network. In HID, it is installed on individual computers rather than network hosts. It will only track data packets sent from the system and will notify the administrator if any suspicious behavior is detected on the network. IDS can be used to assess the quality and nature of attacks and to detect intrusions [16]. IDSs may be used in combination with other security measures such as access control, authentication protocols, and encryption techniques to assist in defending networks against cyberattacks [14]. Data mining, the term used to characterize information discovery, may assist in implementing and deploying IDSs with increased precision and robust behavior compared to more conventional IDSs that might be less effective against sophisticated cyber-attacks [18].

This study presents ML and DL approaches for cybersecurity purposes. Overall, several ML and DL methods for network intrusion detection are described. This paper does not discuss all of the available techniques for detecting network anomalies; rather, it focuses exclusively on ML and DL techniques. The remainder of this paper is divided into the following sections: Section 2 focuses on cybersecurity. Section 3 describes ML methods in cybersecurity. Section 4 describes DL methods in cybersecurity. Section 5 discusses the research status. Section 6 presents conclusions.

2. Cybersecurity

Cybersecurity applies to steps to protect the privacy and safety of electronic information from harm or theft. It is common practice not to misuse these devices and details. Cybersecurity refers to software, hardware, and Internet information and can be used to keep everything from personal information to complicated government systems. For the key cyber security concerns, organizations must maintain efficient vulnerability management systems, including remedial, detection, monitoring, and evaluation, emphasizing information security vulnerabilities [19][2].

Cyber security is the primary component of technology activities and processes aimed at defending networks, devices, and programs from assault, harm, and unauthorized access. In the sense of computing, the term "defense" refers to cyber security. Any industry, including financial institutions, enterprises, governments, and companies, collects, processes, and stores confidential information on computers and transmits it over networks or other computers. Thus, the number of cyber-attacks is increasing. It is also approached in Fig 1. The concepts are described as follows. The primary components of cybersecurity are as follows: Application security, information security, and network security [20].

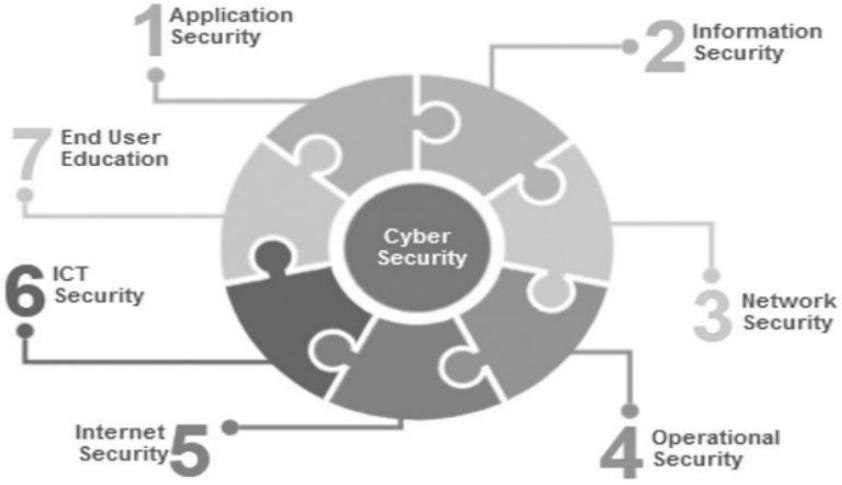

Fig. 1 Cybersecurity and other domains of security [19]

- **Application security:** it represents taking different steps to secure the application. This is often accomplished by monitoring the application for security vulnerabilities and identifying, resolving, and preventing them.
- **Information Security:** This is a collection of procedures or practices in which business data and knowledge are kept confidential, integrity, and accessible in various forms.
- **Network security:** This is the process of defending a computer network from intruders, whether they are deliberate attackers or opportunistic malware.

- **Operations security:** This is the method of defining and safeguarding unclassified sensitive information that is often desirable to competitors or adversaries seeking to obtain real information.
- **Internet security:** Involves a variety of security procedures that are used to ensure the safety of online transactions. It involves guarding browsers, networks, operating systems, and other applications against attacks by establishing precise rules and regulations.
- **ICT security:** It is generally known as the ISO's Technical Root of Computer Security or the CIA Theory of operational security.
- **End-User Knowledge:** This is critical since people are the chain's weakest connection of cybersecurity. User ignorance of cybersecurity threats accounts for 50% of cyberattacks, and nearly 90% of cyberattacks are the result of human actions.

The cyberattacks' goal is to disable or obtain access to the target device. The objective can be accomplished by using a variety of attacks against the target structure. Numerous cyber-attacks occur and continue to develop on a daily basis. The following section discusses some of the most recent cyberattacks [19], [21]:

- **Malware:** Malware is a type of malicious software that is intended to inflict harm on a single device or network. This category includes both traditional malicious software such as worms, malware, and trojans and more recent malicious software such as spyware and ransomware. When a user clicks on a dangerous link, opens an email file, or installs risky applications, the malware infects the device or network. The critical point to remember is that malware reproduces and spreads by interaction with other systems or devices. Several of the causes include blocking network access, installing additional malicious software, and gathering information.
- **Phishing:** Phishing is the act of sending malicious messages that appear to originate from a legitimate source, usually via email. The objective is to steal personal data such as credit card and login information or to infect the victim's computer with malware. Phishing is an increasingly prevalent form of cybercrime.
- **Spam:** It is an unwelcome email message. 12 Spam emails can be a time sink for users, but they can also contain Java applets that run automatically when the message is read.

3. Machine Learning (ML) in Cybersecurity

Machine learning (ML) is a data analysis technique that automates the process of developing analytical models. As a subset of the field of AI, ML is used in every area of computational work where algorithms are designed and performance is increased [22]. The main types of machine learning are supervised and unsupervised learning [23]. The aim of machine learning is to ensure that a machine can learn and automate tasks without human intervention [24]. The most common supervised methods of learning, called classification and regression, are popular in classifying or predicting a security problem in the future. For instance, in the cybersecurity domain, classification techniques can be used to predict a denial of service attack (Yes, No) or classify various network attack groups such as scanning and spoofing [25]. ML could play a major role in cybersecurity. Additionally, ML applications in cybersecurity in a variety of domains, including power system protection, industrial control system security, intrusion detection in supervisory control and data acquisition (SCADA) systems, intrusion detection for vehicular ad-hoc network (VANET) systems, and malware analysis [26]. In the following, we discuss the most common and popular methods that can be used to solve ML tasks and how they are related to cybersecurity tasks.

3.1 Naïve Bayes (NB)

NB is a simple but surprisingly efficient algorithm for predictive modeling. It is a classifier for binary (two-class) and multi-class classification problems. The technique is most easily understood when input values are binary or categorical. It is referred to as naive Bayes or idiot Bayes because the probability calculations for each hypothesis are simplified to make them tractable [27].

The study by, Bhosale [28] proposes the Modified Naïve Bayes Intrusion Detection System (MNBIDS) To enhance DDoS attack detection with the KDDCup99 dataset. The authors have used Data Pre-processing, Hybrid Feature Selection, and Modified Naïve Bayes Algorithms to improve the system accuracy. Then, compare the SVM, CNN, KNN, ANN, and proposed Naive Bayes Classifiers to determine which method the best results. The experiment results show MNBIDS system is better as compare to others. Talita. A [29] has used NB with feature selection based on Particle Swarm Optimization (PSO) method KDD CUP'99 specifically on one of the benchmark datasets for IDS problem. The best classification result was obtained when 38 features were used, an accuracy of 99.12 %.

3.2 K-Nearest Neighbor (K-NN)

K-NN is one of the most basic machine learning algorithms known. The K-NN algorithm can be used for regression and classification, but for classification problems, it is most used and is also called an algorithm for lazy

learners. It is a subset of supervised learning and is widely used in pattern recognition, data processing, and intrusion detection [30].

Chen. F [31] The Author has used Tree-Seed Algorithm (TSA) to extract the most useful characteristic of the input data and K-NN for classification. The proposed model (KNN-TSA) to improve the classification efficiency of intrusion detection with KDD CUP 99 datasets. The experimental results demonstrate that the proposed model can remove redundant features and reduce the classifier's input dimensions. Rao. B [32] Proposed Indexed Partial Distance Search K-NN (IKPDS), Partial Distance Search K-NN (KPDS) techniques for Network Intrusion Detection and comparison to standard K-NN classification. The author evaluates K-NN classification using the NSL-KDD dataset. The results show 99.6% accuracy of the proposed method. Fauzi. In [33], the authors have used the IDS process to select the best features from 41 to 18 using Genetic Algorithms (GA) and KNN with the KDD99 dataset. The results of the K-NN algorithm for training data, which has an accuracy of 99.98%, and for testing data accuracy of 97.52%

3.3 Support Vector Machine (SVM)

SVM is an extensively used supervised learning algorithm for classification and regression issues. However, it is mainly utilized to solve classification issues in machine learning. SVM generates a hyperplane or several hyperplanes in a high-dimensional vacuum. The best hyperplane is one that optimally partitions the given data into different classes using the main partition. A non-linear classifier uses a variety of kernel functions to determine the margins between hyperplanes. The main goal of these kernel functions, such as linear, polynomial, radial basis, and sigmoid, is to maximize the margins between hyperplanes [34]. Fig 2 illustrates the fundamentals of SVM classification [35].

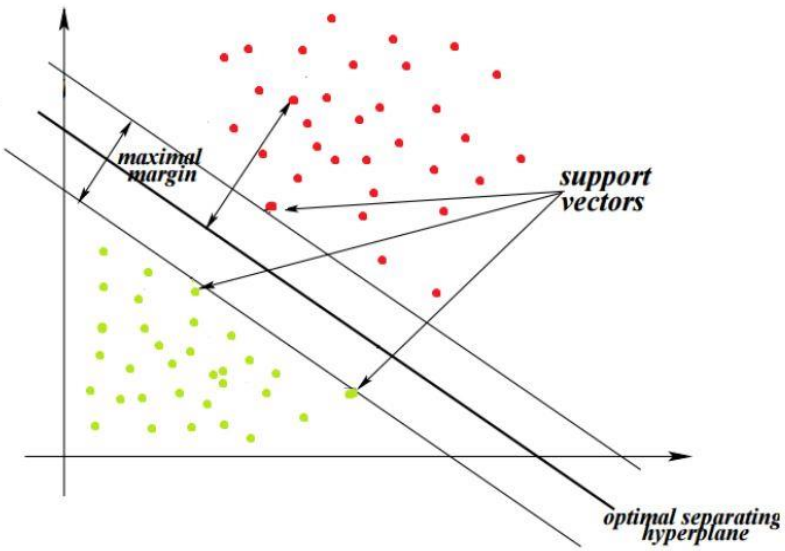

Fig 2 - The fundamental example of SVM classification [35]

The study by Khalvati. L [36] Proposed Intrusion Detection System (IDs) based on novel hybrid learning, that approach combines the K-Medoids clustering and selecting feature using the SVM method. The experimental results on the KDDCUP'99 dataset have shown that this method is capable of achieving 91.5 % for accuracy, 90.1% for detection rate and 6.36 for False alarm rate. The work by, Gu and Lu [37] Proposed a system for efficient intrusion detection using SVM and Naive Bayes with multiple datasets. The experimental results on the UNSW-NB15 dataset, the accuracy was 93.75 %; on the CICIDS2017 dataset, it was 98.92 %; on the NSL-KDD dataset, it was 99.35 %; and on the Kyoto 2006+ dataset, it was 98.58 %. However, Wang. W [38] proposed a Cloud Intrusion Detection based on unsupervised Deep Learning Stacked Denoising Autoencoders (SDAE) method and supervised learning SVM methods with NSL-KDD dataset. The author has used SDAE for dimensionality reduction, and SVM is used to build a classifier that can identify cyberattacks. The test results demonstrate that the approach suggested performs relatively well Classification accuracy, precision rate, recall rate, and f-measure are all factors to consider when compared to the other approaches. The study that is made by Prabakar [39] aims to improve the simulation of annealing and SVM in Wireless Sensor Networks Intrusion Detection Systems. Compared to GWO-SVM and PSO-SVM, the feature selection methods give greater performance. Increased accuracy by 8.71 %, false alarm rate reduction by 81.74 %, identification rate increase by 3.92 %, and execution time reduction by 43.64 %.

3.4 Decision Tree (DT)

DT is a nonparametric supervised learning approach that can be used for both classification and regression. There are two nodes in a DT: the Decision Node and the Leaf Node [40]. ID3 [41], J48 [42], CART algorithm [43]. Classification of instances is accomplished by inspecting the attribute identified by that node, beginning at the root node and continuing down the tree branch corresponding to the attribute value. The most frequently used criteria for splitting are "Entropy" in (eq.1) for the Gini impurity and "Gini" in (eq.2) for the knowledge gain represented mathematically as [41].

$$Entropy = -\sum_{i=1}^{n} p(xi)\, Log_2\, p(xi) \quad (1)$$

$$Gini = 1 - \sum_{i=1}^{n} pi^2 \quad (2)$$

Taghavinejad et al. [44] proposed a smart method for detecting intrusions by combining three DTs. The proposed method's output was compared to that of the SVM, K-NN, and DT methods. The experimental results indicate that the proposed approach outperforms other intrusion detection methods in IoT-based SG. However, Nejati. E [45] Described how to build an intrusion detection system using tree-based classifiers such as DT, RF, and Gradient Boosted Trees (GBT). The model is evaluated using a variety of feature selection methods and the NSL-KDD dataset.

In addition, to the most popular ML methods mentioned above, several other ML techniques exist in the field for a variety of purposes. For instance, a random forest (RF) is a multiple DT are usually more accurate in a learning model than a single DT [25]. Regression analysis consists of a variety of ML techniques that allow the prediction of a continuous (y) outcome variable based on the values of one or more (x) predictor variables. Some well-known forms of regression methods include linear, polynomial, lasso, and ridge regression, among others [46]. Cluster analysis, alternatively referred to as clustering, is an unsupervised ML method used to identify and group similar data points in large datasets without concern for the particular outcome [47]. Clustering can be used in various fields of use, such as cybersecurity, electronic commerce, mobile data processing, health analysis, and user modeling [48]. The most common clustering algorithms are K-means [49][50] and K-Medoids [51].

4. Deep Learning in (DL) Cybersecurity

Deep Learning (DL) typically is known to be a subset of Machine Learning (ML) in the field of Artificial Intelligence (AI), which is a computer simulation based on the biological neural networks found in the human brain [52], Artificial Neural Networks (ANN) are often used in DL, and the most widely used neural network algorithm is backpropagation [52]. It is a neural learning network composed of several layers. There are three layers in total: an input layer, one or two hidden layers, and an output layer. The main distinction between the two is DL, and traditional DL is the quality that DL produces as the amount of security data increases. By and large, DL algorithms perform better than ML algorithms when data volumes are large, while ML algorithms outperform DL algorithms when data volumes are small [17]. In the following section, we will discuss the most common neural network and DL algorithms in the context of cybersecurity, including the multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), and long-short term memory (LSTM) network. In the area of cybersecurity, DL may be used for a variety of purposes, including detecting network intrusions, detecting and classifying malware, performing security threat analysis, forecasting cyberattacks or anomalies, and conducting backdoor attacks [53] [14].

4.1 Muti-Layer Perceptron (MLP)

MLPs are a subclass of feedforward ANN, which is a form of a supervised learning algorithm. Furthermore, it is referred to as the fundamental architecture for DL or Deep Neural Networks (DNN) [54]. A typical MLP is a fully connected network with an input layer that receives data, an output layer that makes between these two layers, a judgment or inference about the input signal, and one or more hidden layers [55]. Fig. 3 displays ANN modeling for cyber anomalies or attacks, considering the input, hidden, and output layer [54].

The study by, Tavoli [56] A novel approach for intrusion detection is proposed that utilizes a multilayer perceptron (MLP) neural network. The proposed approach is divided into two phases: training and testing with the KDD99 dataset. Backpropagation error algorithms are used to train MLP neural networks. MLP is good at lowering the rate of false positives. The proposed method produces significantly better results than other approaches.

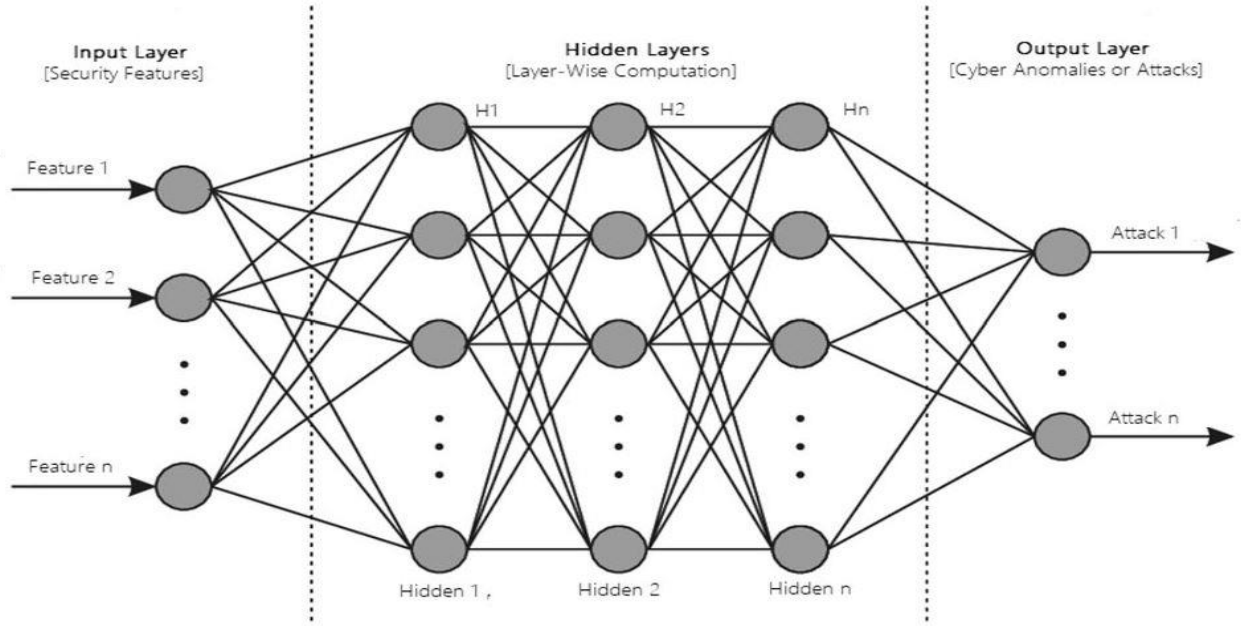

Fig 3 - A structure of ANN modeling for cyber anomalies considering the input, hidden, and output layer [54]

4.2 Convolution Neural Network (CNN)

A type of ANN, CNN is capable of extracting local features in data [57], and the biggest advantage is that its multilayer structure has the ability of automatic learning [58]. Each layer in CNN takes optimized parameters into consideration to ensure a significant result while minimizing complexity [59]. This type of network has been successfully used to recognize images and video, classification, and text processing. In general, CNN often consists of multiple layers are f an input layer, convolutional layer, pooling layer, fully connected layers, and an output layer, as depicted in Fig 4 [60]. The CNN's are most frequently used for visual image analysis, they can also be used for cybersecurity, it has been used for tasks like intrusion detection, IoT Networks [61], classification and detection of malware variants [62], android malware detection, etc. [63].

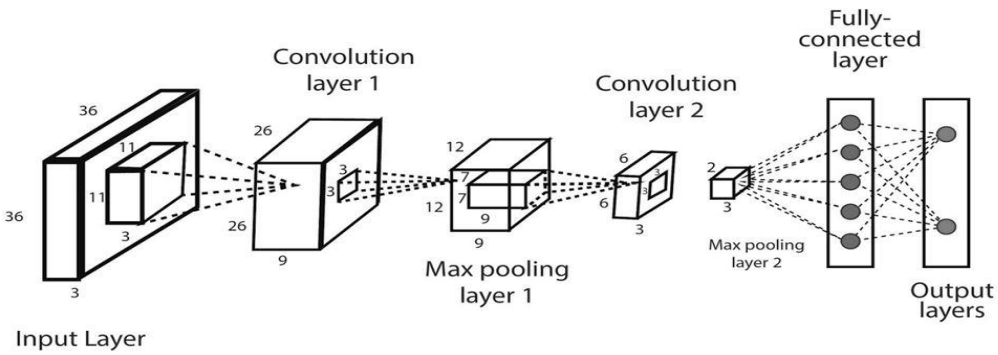

Fig 4 - An example of CNN [60]

Nguyen et al. [64] Proposed an IDS framework based on CNN to detect DoS attacks. The experiments demonstrate that our CNN-based DoS detection algorithm achieves a high level of precision, up to 99.87 %. Additionally, comparisons to other ML algorithms, such as KNN, SVM, and Naive Bayes, show that our proposed method outperforms established techniques. However, [65] has used network attack IDS focused on Convolutional Neural Networks (CNNs) for protecting the CAN bus system. The experiment results show that our classifier effectively detects CAN bus machine attacks, with a detection rate of 0.99 and a high accuracy of 99.99 %.

4.3 Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN)

Recurrent Neural Networks (RNNs) have made significant contributions to computer vision and natural language processing are two areas on which researchers are working (NLP), semantic comprehension, speech recognition,

language modeling, translation, image representation, and recognition of human behavior, among others [66]. The RNN model architecture is a feedback loop that connects layers and has the ability to store data from previous inputs, thus improving the model's reliability. For modeling sequence data, RNNs are extremely efficient [67]. In particular, a Short-Term Long Memory (LSTM) architecture can resolve the RNN's vanishing gradient problem, allowing it to achieve long-term dependence. LSTM is an advantageous method for dealing with time-sequential data [68]. Fig 5 illustrates an LSTM cell in which the 'Forget Gate', the 'Input Gate', and the 'Output Gate' all function cooperatively to monitor the flow of information in an LSTM unit [69].

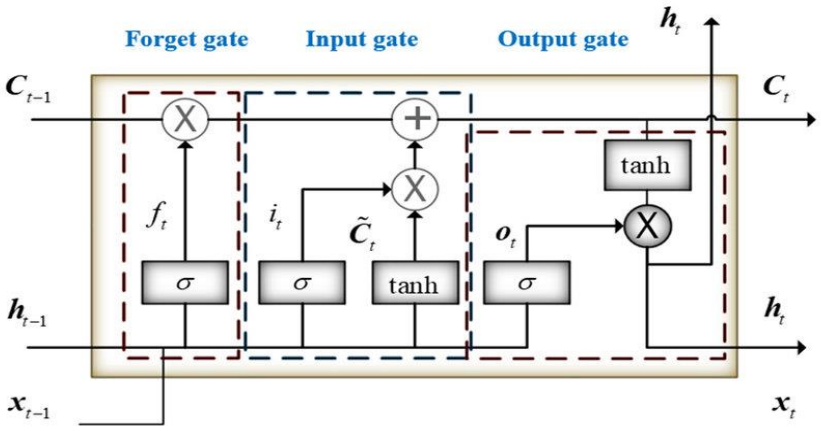

Fig. 5 - The basic structure of the LSTM unit [69]

As a novel strategy, there are several studies on these models in the area of cybersecurity, including the following: Detection of anomalies in 5G networks [70], Detection of distributed attacks in fog-to-things computing. The study by [71] proposed an intrusion detection system for SDNs that utilizes a Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) with the NSL-KDD dataset. The experimental results indicated that with just six raw features, the accuracy was 89 %. Tang et al. [72] propose a Software-Defined Network (SDN) framework. The authors have used DNN to construct an anomaly detection model. They trained their model using only six fundamental features drawn from the NSL-KDD Dataset's 41 features. Also, Kim [73] has used the LSTM structure to construct an IDS model that was trained on the KDD 99 dataset. The experimental results demonstrated that the LSTM structure is efficient for intrusion detection, and a feasible solution was suggested for reducing the false positive rate. This further demonstrated the suitability of the LSTM structure for intrusion detection. Kim et al. [74] used KDD 1999 data. The proposed intrusion detection model is comprised of two parameters: four hidden layers and one hundred hidden units. For deep neural network training, the ReLU function is used as the activation function, and the stochastic optimization approach is used. The proposed model is approximately 99 % accurate. Sokolov [75] proposed a method focused on DNN for analyzing cyber-threats in cloud-based applications. The proposed system employs four neural classifiers for network traffic, spam comments, spam email, and photos. The obtained findings are comparable to those obtained using contemporary approaches. The accuracy achieved for each aspect is comparable to that of contemporary works.

Along with the above-mentioned most popular DL methods, many other deep learning approaches exist in the field for a variety of purposes. For example, the Self-Organizing Map (SOM) is a form of ANN that uses unsupervised learning to convert high-dimensional data to a two-dimensional grid map, thereby achieving dimensionality reduction [76]. The Auto Encoder (AE) is another learning method that is commonly used in unsupervised learning tasks for dimensionality reduction and feature extraction [77]. The Restricted Boltzmann Machines (RBM) method is essential for dimension reduction, classification, regression, collaborative filtering, feature learning, and topic modeling [78]. A Deep Belief Network (DBN) is a type of DNN unsupervised network that is similar to RBM or autoencoders. It is also a type of Backpropagation Neural Network (BPNN) [79]. A Generative Adversarial Network (GAN) is a form of DL network that can generate new data with properties similar to the original data input. Additionally, GANs are commonly used in natural image synthesis, medical image processing, bioinformatics, data augmentation, video generation, and voice generation, among other applications. Additionally, it is beneficial in the field of cybersecurity [80].

5. Assessment and Recommendation

Our work examines a large number of academic intrusion detection studies based on Machine Learning (ML) and Deep Learning (DL), as shown in Table 1.

Table 1 - A summary of cybersecurity-related machine learning and deep learning tasks

Ref	Years	Techniques	Datasets	Purpose	Accuracy
[81]	2018	PSA and KNN	NSL-KDD	Intrusion Detection System	94 %
[82]	2019	KNN and DNN	CICIDS-2017	Anomaly Analysis for the Classification Purpose of Intrusion Detection System	KNN= 0.9293% DNN= 0.8824%
[83]	2019	DT	KDDCUP99	Enhance hybrid Intrusion Detection System	99.8 %
[84]	2019	DBN	NSL-KDD	Anomaly Detection System	97.5%
[85]	2020	RNN	NSL-KDD	Anomaly intrusion detection system and attack classification	98.27%
[86]	2020	PCA and RF	KDDCUP99	Intrusion Detection System	96.78%
[87]	2020	J48 DT and SVM	KDDCUP99	hybrid Intrusion Detection System	99.6%
[88]	2020	Signature Apriori algorithm	CICID2017	Novel Network Intrusion Detection System	99.56%
[89]	2020	MLP and PID	CICID2017	Intrusion Detection System	98.96%
[90]	2020	MLP	KDDCUP99	Intrusion Detection and Prevention System	91.4%

In this paper, we reviewed most of the ML and DL tasks in the domain of cybersecurity. We realize that most of them have good accuracy. However, some research has higher accuracy through the literature in Table 1, ML and DL technique with variance datasets such as the authors [85],[87],[89], and [91]. The authors of [85] have superior accuracy by utilized the KDDCUP99 dataset to enhance the hybrid intrusion detection system. The authors are proposing a new method based on a decision tree of data mining techniques that are based on the C4.5 algorithm to provide better results with high accuracy of the detection rate and reduce the value of false-positive rate. In [89] Proposed hybrid IDS is a combination of two ML algorithms J48 DT and SVM. To select relevant features from the KDD CUP dataset. Also, the author [90] proposed an algorithm to use the known signature to find the signature of the related attack quickly. In addition, the author [91] In the learning process, two methods have been suggested PID (Perimeter Intrusion Detection) with MLP (Multi-Layer Perceptron), and these proposed methods used quantum classifiers to address issues in ID, and then PID with MLP and quantum classifier algorithm to achieve all parameters such as performance, accuracy, and higher consistency. Furthermore, [87] proposed a model that makes use of a multi-layered RNN. It is intended to be applied for fog computing security similar to end-users and IoT applications. It has been demonstrated that the proposed model used a balanced variant of the difficult dataset: NSL-KDD.

On the other hand, some techniques reviewed in this paper have lower accuracy than other algorithms that have high accuracy ,as mentioned above, such as [92] proposed Intrusion Detection System used a DL model called Multi-Layer Perceptron (MLP) trained on the kddcup99 dataset. The Intrusion detection and prevention systems are integrated as a single framework to accomplish the goal of intrusion detection and prevention activities in a more reliable and timely manner. Generally, as we reviewed most f o ML and DL algorithms, we realized that most of them have higher accuracy.

6. Conclusion

We presented a study of machine learning and deep learning approaches for network security in the last three years, highlighting the most recent implementations of machine learning and deep learning in the area of Cybersecurity intrusion detection. Network intrusion detection datasets are important for training and testing systems. Machine learning and deep learning algorithms will not work without representative data, and collecting such a dataset is challenging and time-consuming. A series of similar studies were explored in terms of intrusion detection, malware detection, and the use of deep learning and machine learning approaches, as well as their accomplishments and limitations.

Acknowledgment

The authors would like to acknowledge Duhok Polytechnic University for providing all financial support and support for this study.

References

[1] Yavanoglu, O., & Aydos, M. (2017, December). A review on cyber security datasets for machine learning algorithms. In 2017 IEEE international conference on big data (big data) (pp. 2186-2193). IEEE

[2] Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. IEEE Access, 9, 22351-22370

[3] Jain, A. K., Goel, D., Agarwal, S., Singh, Y., & Bajaj, G. (2020). Predicting spam messages using back propagation neural network. Wireless Personal Communications, 110(1), 403-422

[4] Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802
[5] Shajideen, N. M., & Bindu, V. (2018, March). Spam filtering: A comparison between different machine learning classifiers. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1919-1922). IEEE
[6] Lokanan, M., Tran, V., & Vuong, N. H. (2019). Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. Asian Journal of Accounting Research
[7] Shukur, H. A., & Kurnaz, S. (2019). Credit card fraud detection using machine learning methodology. International Journal of Computer Science and Mobile Computing, 8(3), 257-260
[8] Ma, Z., Ge, H., Liu, Y., Zhao, M., & Ma, J. (2019). A combination method for android malware detection based on control flow graphs and machine learning algorithms. IEEE access, 7, 21235-21245
[9] Jain, P. (2019). Machine learning versus deep learning for malware detection
[10] Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W., & Ye, H. (2018). Significant permission identification for machine-learning-based android malware detection. IEEE Transactions on Industrial Informatics, 14(7), 3216-3225
[11] Yerima, S. Y., & Sezer, S. (2018). Droidfusion: A novel multilevel classifier fusion approach for android malware detection. IEEE transactions on cybernetics, 49(2), 453-466
[12] Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and Applications, 31(8), 3851-3873
[13] Almukaynizi, M., Grimm, A., Nunes, E., Shakarian, J., & Shakarian, P. (2017, October). Predicting cyber threats through hacker social networks in darkweb and deepweb forums. In Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas (pp. 1-7)
[14] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419
[15] Shahriar, H., Qian, K., & Zhang, H. (2020, July). Learning Environment Containerization of Machine Leaning for Cybersecurity. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 1131-1132). IEEE
[16] Dhir, S., & Kumar, Y. (2020, August). Study of Machine and Deep Learning Classifications in Cyber Physical System. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 333-338). IEEE
[17] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. Ieee access, 6, 35365-35381
[18] Khalaf, B. A., Mostafa, S. A., Mustapha, A., Mohammed, M. A., & Abduallah, W. M. (2019). Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods. IEEE Access, 7, 51691-51713
[19] Torres, J. M., Comesaña, C. I., & Garcia-Nieto, P. J. (2019). Machine learning techniques applied to cybersecurity. International Journal of Machine Learning and Cybernetics, 10(10), 2823-2836
[20] Sarker, I. H. (2021). Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. SN Computer Science, 2(3), 1-16
[21] Saravanan, A., & Bama, S. S. (2019). A Review on Cyber Security and the Fifth Generation Cyberattacks. Oriental Journal of Computer Science and Technology, 12(2), 50-56
[22] Chaudhary, D., & Vasuja, E. R. (2019). A Review on Various Algorithms used in Machine Learning. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, 5(2), 915-920
[23] Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine learning supervised algorithms of gene selection: A review. Machine Learning, 62(03)
[24] Vinayakumar, R., Soman, K. P., Poornachandran, P., & Akarsh, S. (2019). Application of deep learning architectures for cyber security. In Cybersecurity and Secure Information Systems (pp. 125-160). Springer, Cham.
[25] Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1), 1-28
[26] Handa, A., Sharma, A., & Shukla, S. K. (2019). Machine learning in cybersecurity: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1306
[27] Sarker, I. H. (2019). A machine learning based robust prediction model for real-life mobile phone data. Internet of Things, 5, 180-193
[28] Bhosale, K. S., Nenova, M., & Iliev, G. (2018, December). Modified naive bayes intrusion detection system (mnbids). In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 291-296). IEEE
[29] Talita, A. S., Nataza, O. S., & Rustam, Z. (2021, February). Naïve Bayes Classifier and Particle Swarm Optimization Feature Selection Method for Classifying Intrusion Detection System Dataset. In Journal of Physics: Conference Series (Vol. 1752, No. 1, p. 012021). IOP Publishing

[30] Yu, H., Chan, P. P., Ng, W. W., & Yeung, D. S. (2010, July). Apply randomization in KNN to make the adversary harder to attack the classifier. In 2010 International Conference on Machine Learning and Cybernetics (Vol. 1, pp. 179-183). IEEE
[31] Chen, F., Ye, Z., Wang, C., Yan, L., & Wang, R. (2018, September). A feature selection approach for network intrusion detection based on tree-seed algorithm and K-nearest neighbor. In 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS) (pp. 68-72). IEEE
[32] Rao, B. B., & Swathi, K. (2017). Fast kNN classifiers for network intrusion detection system. Indian Journal of Science and Technology, 10(14), 1-10
[33] Fauzi, M. A., Hanuranto, A. T., & Setianingsih, C. (2020, October). Intrusion Detection System using Genetic Algorithm and K-NN Algorithm on Dos Attack. In 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-6). IEEE
[34] Bhatt, M., Dahiya, V., & Singh, A. (2019, February). Supervised Learning Algorithm: SVM with Advanced Kernel to classify Lower Back Pain. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 17-19). IEEE.
[35] Hu, T., Wu, L., Zhang, X., Yin, Y., & Yang, Y. (2019, October). Hardware trojan detection combine with machine learning: an svm-based detection approach. In 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID) (pp. 202-206). IEEE
[36] Khalvati, L., Keshtgary, M., & Rikhtegar, N. (2018). Intrusion detection based on a novel hybrid learning approach. Journal of AI and data mining, 6(1), 157-162
[37] Gu, J., & Lu, S. (2021). An effective intrusion detection approach using SVM with naïve Bayes feature embedding. Computers & Security, 103, 102158
[38] Wang, W., Du, X., Shan, D., & Wang, N. (2019, October). A hybrid cloud intrusion detection method based on SDAE and SVM. In 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 271-274). IEEE
[39] D. Prabakar, S. Sasikala, and T. R. Saravanan, "Enhanced Simulating Annealing and SVM for Intrusion Detection System in Wireless Sensor Networks," 202 Prabakar, D., Sasikala, S., & Saravanan, T. R. (2021). Enhanced Simulating Annealing and SVM for Intrusion Detection System in Wireless Sensor Networks
[40] Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. Mobile Networks and Applications, 25(3), 1151-1161
[41] Patil, S., & Kulkarni, U. (2019, April). Accuracy prediction for distributed decision tree using machine learning approach. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1365-1371). IEEE
[42] Posonia, A. M., Vigneshwari, S., & Rani, D. J. (2020, December). Machine Learning based Diabetes Prediction using Decision Tree J48. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 498-502). IEEE
[43] Shamrat, F. J. M., Ranjan, R., Md, K., Hasib, A. Y., & Siddique, A. H. Performance Evaluation among ID3, C4. 5, and CART Decision Tree Algorithms
[44] Taghavinejad, S. M., Taghavinejad, M., Shahmiri, L., Zavvar, M., & Zavvar, M. H. (2020, April). Intrusion detection in IoT-based smart grid using hybrid decision tree. In 2020 6th International Conference on Web Research (ICWR) (pp. 152-156). IEEE
[45] Nejati, E., Shakeri, H., & Sani, H. R. (2020, September). Ensembling tree-based classifiers for improving the accuracy of cyber attack detection. In 2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS) (pp. 70-76). IEEE
[46] Gambhir, E., Jain, R., Gupta, A., & Tomer, U. (2020, September). Regression analysis of COVID-19 using machine learning algorithms. In 2020 International conference on smart electronics and communication (ICOSEC) (pp. 65-71). IEEE
[47] Ahmed, M., & Barkat, A. (2019, May). Performance analysis of hard clustering techniques for big IoT data analytics. In 2019 Cybersecurity and Cyberforensics Conference (CCC) (pp. 62-66). IEEE
[48] Homsi, S., Quan, G., Wen, W., Chapparo-Baquero, G. A., & Njilla, L. (2019, May). Game Theoretic-Based Approaches for Cybersecurity-Aware Virtual Machine Placement in Public Cloud Clusters. In 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (pp. 272-281). IEEE
[49] Peng, C., Yongli, W., Boyi, Y., Yuanyuan, H., Jiazhong, L., & Qiao, P. (2020, December). Cyber Security Situational Awareness Jointly Utilizing Ball K-Means and RBF Neural Networks. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 261-265). IEEE
[50] Jahwar, A. F., & Abdulazeez, A. M. (2020). Meta-heuristic algorithms for k-means clustering: A review. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 12002-12020

[51] Madhulatha, T. S. (2011, July). Comparison between k-means and k-medoids clustering algorithms. In International Conference on Advances in Computing and Information Technology (pp. 472-481). Springer, Berlin, Heidelberg
[52] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," Morgan Kaufmann Ser. Data Manag. Syst., vol. 5, no. 4, pp. 83–124, 2011 Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier
[53] Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019, May). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 707-723). IEEE
[54] Van Efferen, L., & Ali-Eldin, A. M. (2017, May). A multi-layer perceptron approach for flow-based anomaly detection. In 2017 international symposium on networks, computers and communications (ISNCC) (pp. 1-6). IEEE
[55] Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 1-18
[56] Tavoli, R. (2019). Providing a method to reduce the false alarm rate in network intrusion detection systems using the multilayer perceptron technique and backpropagation algorithm. In 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI) (pp. 001-006). IEEE
[57] Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018, October). Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE
[58] Liang, C., & Xin, S. (2020, July). Research Status and Prospects of Deep Learning in Medical Images. In 2020 International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 380-382). IEEE
[59] Liu, H., Lang, B., Liu, M., & Yan, H. (2019). CNN and RNN based payload classification methods for attack detection. Knowledge-Based Systems, 163, 332-341
[60] Ker, J., Wang, L., Rao, J., & Lim, T. (2017). Deep learning applications in medical image analysis. Ieee Access, 6, 9375-9389
[61] Susilo, B., & Sari, R. F. (2020). Intrusion detection in IoT networks using deep learning algorithm. Information, 11(5), 279
[62] Wang, W., Zhao, M., & Wang, J. (2019). Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 10(8), 3035-3043
[63] McLaughlin, N., Martinez del Rincon, J., Kang, B., Yerima, S., Miller, P., Sezer, S., ... & Joon Ahn, G. (2017, March). Deep android malware detection. In Proceedings of the seventh ACM on conference on data and application security and privacy (pp. 301-308)
[64] Nguyen, S. N., Nguyen, V. Q., Choi, J., & Kim, K. (2018, February). Design and implementation of intrusion detection system using convolutional neural network for DoS detection. In Proceedings of the 2nd international conference on machine learning and soft computing (pp. 34-38)
[65] Hossain, M. D., Inoue, H., Ochiai, H., Fall, D., & Kadobayashi, Y. (2020, December). An Effective In-Vehicle CAN Bus Intrusion Detection System Using CNN Deep Learning Approach. In GLOBECOM 2020-2020 IEEE Global Communications Conference (pp. 1-6). IEEE
[66] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. Ieee Access, 5, 21954-21961
[67] Macas, M., & Wu, C. (2020, November). Deep Learning Methods for Cybersecurity and Intrusion Detection Systems. In 2020 IEEE Latin-American Conference on Communications (LATINCOM) (pp. 1-6). IEEE
[68] Kim, G., Lee, C., Jo, J., & Lim, H. (2020). Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. International journal of machine learning and cybernetics, 11(10), 2341-2355
[69] Jiang, C., Chen, Y., Chen, S., Bo, Y., Li, W., Tian, W., & Guo, J. (2019). A mixed deep recurrent neural network for MEMS gyroscope noise suppressing. Electronics, 8(2), 181
[70] Maimó, L. F., Gómez, Á. L. P., Clemente, F. J. G., Pérez, M. G., & Pérez, G. M. (2018). A self-adaptive deep learning-based system for anomaly detection in 5G networks. IEEE Access, 6, 7700-7712
[71] Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2018, June). Deep recurrent neural network for intrusion detection in sdn-based networks. In 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft) (pp. 202-206). IEEE
[72] Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2016, October). Deep learning approach for network intrusion detection in software defined networking. In 2016 international conference on wireless networks and mobile communications (WINCOM) (pp. 258-263). IEEE
[73] Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016, February). Long short term memory recurrent neural network classifier for intrusion detection. In 2016 International Conference on Platform Technology and Service (PlatCon) (pp. 1-5). IEEE.

[74] Kim, J., Shin, N., Jo, S. Y., & Kim, S. H. (2017, February). Method of intrusion detection using deep neural network. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 313-316). IEEE
[75] Sokolov, S. A., Iliev, T. B., & Stoyanov, I. S. (2019, May). Analysis of cybersecurity threats in cloud applications using deep learning techniques. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 441-446). IEEE
[76] Qu, X., Yang, L., Guo, K., Ma, L., Sun, M., Ke, M., & Li, M. (2019). A survey on the development of self-organizing maps for unsupervised intrusion detection. Mobile networks and applications, 1-22
[77] Sarker, I. H., Abushark, Y. B., & Khan, A. I. (2020). Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry, 12(4), 499
[78] YImamverdiyev, Y., & Abdullayeva, F. (2018). Deep learning method for denial of service attack detection based on restricted boltzmann machine. Big data, 6(2), 159-169
[79] Wei, P., Li, Y., Zhang, Z., Hu, T., Li, Z., & Liu, D. (2019). An optimization method for intrusion detection classification model based on deep belief network. IEEE Access, 7, 87593-87605
[80] Yin, C., Zhu, Y., Liu, S., Fei, J., & Zhang, H. (2018, May). An enhancing framework for botnet detection using generative adversarial networks. In 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 228-234). IEEE
[81] Benaddi, H., Ibrahimi, K., & Benslimane, A. (2018, October). Improving the intrusion detection system for nsl-kdd dataset based on pca-fuzzy clustering-knn. In 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 1-6). IEEE
[82] Atefi, K., Hashim, H., & Kassim, M. (2019, December). Anomaly analysis for the classification purpose of intrusion detection system with K-nearest neighbors and deep neural network. In 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC) (pp. 269-274). IEEE
[83] Ahmed, M. R. A. G., & Ali, F. M. A. (2019, September). Enhancing hybrid intrusion detection and prevention system for flooding attacks using decision tree. In 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) (pp. 1-4). IEEE
[84] Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A survey of deep learning-based network anomaly detection. Cluster Computing, 22(1), 949-961
[85] Almiani, M., AbuGhazleh, A., Al-Rahayfeh, A., Atiewi, S., & Razaque, A. (2020). Deep recurrent neural network for IoT intrusion detection system. Simulation Modelling Practice and Theory, 101, 102031
[86] Waskle, S., Parashar, L., & Singh, U. (2020, July). Intrusion detection system using PCA with random forest approach. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 803-808). IEEE
[87] Kumari, A., & Mehta, A. K. (2020, October). A Hybrid Intrusion Detection System Based on Decision Tree and Support Vector Machine. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 396-400). IEEE
[88] Chen, L., Kuang, X., Xu, A., Suo, S., & Yang, Y. (2020, December). A Novel Network Intrusion Detection System Based on CNN. In 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD) (pp. 243-247). IEEE
[89] Thirumalairaj, A., & Jeyakarthic, M. (2020, January). Perimeter Intrusion Detection with Multi Layer Perception using Quantum Classifier. In 2020 Fourth International Conference on Inventive Systems and Control (ICISC) (pp. 348-352). IEEE
[90] Krishna, A., Lal, A., Mathewkutty, A. J., Jacob, D. S., & Hari, M. (2020, July). Intrusion Detection and Prevention System Using Deep Learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 273-278). IEEE