JOURNAL OF INNOVATION AND KNOWLEDGE

COMPARATIVE ANALYSIS OF MACHINE LEARNING METHODS FOR THREAT ANALYSIS AND PREDICTION IN IOT DEVICES

KONTAGORA Muhammad Mamman [1]
PhD Candidate, Department of Cyber Security, Nasarawa State University, Keffi, Nassarawa State, Nigeria.
***Email**: mohakontagora14@gmail.com*

ADESHINA A. Steve[2]
Professor, Department of Computer Engineering, Nile University of Nigeria, Abuja, Nigeria.
***Email**: steve.adeshina@nileuniversity.edu.ng*

MUSA Habiba[3]
Associate Professor, Department of Public and International Law, Nasarawa State University, Keffi, Nassarawa State, Nigeria.
***Email**: habibamusa@nsuk.edu.ng*

ABSTRACT

The growing number of linked devices in IoT contexts has resulted in a proportional rise in security vulnerabilities. In order to ensure the security and privacy of IoT networks and devices, it is essential to have threat analysis and prediction system in place. Machine learning algorithms have emerged as a potential tool for threat analysis and prediction in the Internet of Things. Nevertheless, there exist several machine learning algorithms to choose from, each possessing its own advantages and disadvantages. This research paper presents a comparison of commonly used machine learning approaches for threat analysis and prediction in IoT scenarios. The study analyzes the merits and drawbacks of each algorithm, including their capacity to identify both familiar and unfamiliar assaults, their rates of false positives, their computing efficacy, and their prerequisites for training data. This study provides a comprehensive analysis of several machine learning techniques, such as Support Vector Machines, Artificial Neural Networks (ANN), Logistic Regression (LR), Decision Trees (DT), K-Nearest Neighbour (kNN), Random Forest (RF), Naive Bayes, and Deep Learning. The research encompasses an examination of the algorithms' efficacy in various scenarios, along with their inherent constraints. The research concluded by providing advice for choosing the optimal machine learning technique for threat analysis and prediction in IoT contexts. The suggestions take into account the particular use case, the data that is accessible, and other pertinent variables. The report offers significant insights for enterprises seeking to enhance their IoT security stance and safeguard their devices and networks from possible attacks.

Keywords: Threat Analysis, Machine Learning Methods, Cyber Security.

Introduction

The digital transformation is linked to the notion of the Internet of Things (IoT), which refers to decentralized and diverse networks of networked objects. This area integrates wireless sensor networks, real-time computing, embedded systems, and actuation technologies. The Internet of Things (IoT) is closing the divide between Operational Technology (OT) and Information Technology (IT) by combining physical and business processes, together with control and information systems (Chilamkurti et al, 2018).

Nevertheless, the merging of hitherto separate systems and technologies encounters significant security obstacles. Internet of Things (IoT) devices sometimes include software and communication protocol vulnerabilities, as well as

inadequate physical security and limitations in available resources (Papa (2012). Malware attacks provide a significant risk to IoT systems. A self-replicating virus, like Mirai, has the ability to infiltrate several vulnerable devices and create a network of bots to carry out multiple cyber-attacks. Cyber-attacks against IoT devices may be classified into two types: passive and aggressive. Passive attacks, such as eavesdropping and traffic analysis, do not directly affect the functioning of the system. Conversely, active assaults include a wide variety of techniques, including probing, man-in-the-middle, brute-force, and Denial-of-Service (DoS) (Zade, 2020).

Given the vulnerability of IoT to hostile actions, a dependable threat analysis and prediction system is very necessary. A threat analysis and prediction is also an Intrusion Detection System (IDS) actively monitors an environment in order to detect and identify potentially malicious activities, allowing for the timely mitigation of such risks. Applying machine learning methods to threat analysis and prediction is a potential approach to address the rising quantity and escalating intricacy of cyber-attacks (Rampone, 2015).

One significant obstacle in the IoT sector is the susceptibility of IoT devices. This is mostly due to manufacturers' lack of understanding of the importance of IoT security concerns. Furthermore, manufacturers may not prioritize the implementation of security measures on devices, even when they are aware of security problems, owing to financial limitations (Usmonov et al, 2017).

The complexity of the IoT automated network system is escalating as demand and expansion persistently climb. The increase has been propelled by the cost-effectiveness of sensors, the emergence of wireless connection, and the advancement of cloud computing. The rise of data-driven infrastructure has led to a growing emphasis on the use of machine learning (ML) in combination with the Internet of Things (IoT) (Anthi, 2018). The fields in which IoT and ML approaches are used include smart homes, industrial automation, healthcare, agriculture, smart cities, retail, and transportation. For instance, in the context of smart homes, Internet of Things (IoT) devices have the capability to automate several elements of residential properties, including lighting, temperature control, and security (Papa, 2012).

Similarly, in the field of healthcare, these devices may be used to remotely monitor the health status of patients. Although IoT and ML applications provide several advantages, their growing complexity makes them susceptible to unexpected flaws, resulting in security breaches and other irregularities. Moreover, ML techniques are necessary for performing intricate tasks such as interpreting ECG, identifying disorders using X-Ray analysis, and analyzing genetic data. The aircraft sector may get advantages from machine learning methodologies (Sheikhan, 2017). IoT devices are susceptible to attacks due to their wireless nature. While attacks on local networks are often confined to nearby nodes or a small local domain, attacks on IoT systems have the capacity to propagate over a larger geographical region and have substantial impact on IoT locations (Brun, 2018).

In order to protect against cybercrime, it will be essential to have a secure Internet of Things (IoT) infrastructure in the future. Nevertheless, the susceptibility of IoT devices to attacks undermines the effectiveness of the implemented security measures. Data serves as the currency for some stakeholders and business owners, with certain information being classified and sensitive for government and commercial institutions. The vulnerability of an IoT node might serve as an entry point for attackers to illicitly get sensitive data from any crucial organization (Lopez-Martin, 2017).

As previously said, there are several simple and direct methods to tackle the difficulties. The signature-based technique involves storing assaults and abnormalities in a database and periodically comparing them against the database. Nevertheless, this method may be computationally demanding and is also vulnerable to unanticipated risks. The Internet of Things (IoT) devices produce a substantial volume of data, a significant portion of which contains sensitive information pertaining to people, enterprises, and smart cities (Papa, 2012).

Conducting a comparative examination of machine learning algorithms used for intrusion detection in IoT contexts might give significant insights into their strengths and drawbacks. Such analysis may assist companies in selecting the best suitable algorithm for their unique use case and identifying areas that need more study to improve the efficiency of intrusion detection systems in IoT settings. Implementing this method may enhance the overall security stance of IoT, which is essential for protecting devices and networks from possible attacks.

The literature review for this paper was executed through a structured, multi-stage process with a specific focus on Machine Learning (ML) applications in IoT threat analysis and prediction. Initially, a broad collection or screening of literature was conducted, drawing from an extensive pool of 22,688 potential sources related to IoT threat analysis and prediction across scientific databases like IEEE, Springer, Science Direct, Scopus, and Web of Science. This initial collection included 14,014 conference papers, 4983 journal articles, 213 magazine articles, 96 books, 114 early access articles, 7 standards, and 4 courses.

During the systematic selection phase, we focused on sources that integrated ML in IoT Security, reducing the selection to about 10 % of the initial pool (approximately 710 sources). The systematic selection phase then followed, during which around 29 sources were selected based on specific criteria: relevance to the ML for IoT Security topic,

author, and journal reputation (preferring those with an impact factor above 3), originality of the content, publication date (prioritizing those published within the last five years), and impact (considering papers with at least 50 citations). The selected references were then classified into two main categories: technical papers (about 90 % of the selected works) and survey papers (10 %). In the final analysis phase, critical information was extracted from approximately 80 % of the technical papers and 90 % of the survey papers. This information was thoroughly analyzed and synthesized into the comprehensive survey presented in this paper, offering a detailed insight into the comparison of ML methods and IoT Security.

1. Review of Related Literature

Prior studies in the domain of Internet of Things (IoT) have shown encouraging outcomes. Pahl et al. (2017) created a method for detecting anomalies and protecting against unauthorized access in IoT microservices deployed at IoT locations. This study used clustering techniques, namely K-Means and BIRCH, to categorize distinct microservices. The clustering model was updated using an online learning approach, and clusters were merged if their centroid was within a distance of three standard deviations. The system attained an overall accuracy of 96.3% using the implemented techniques.

The research conducted in the domain of IoT has significantly contributed to the development of many technologies designed to identify and prevent security breaches. In Wang et al (2019), a smart home system was reported that used a deep learning technique using a Dense Random Neural Network (DRNN) to identify and detect Denial of Service (DoS) and Denial of Sleep (DoS) threats. The system used a collection of parameters derived from packet captures to forecast the likelihood of a network assault. The authors provide a comprehensive description of the system's structure and assessment outcomes, showcasing the efficacy of the method.

Liu et al. (2014) conducted a research where they created a detector to identify and counteract hostile network nodes that engage in On and Off assaults in an industrial Internet of Things (IoT) setting. These attacks occur when a malicious node targets an IoT network when it is in an active or "On" state. However, the network operates normally while the malicious node is in an inactive or "Off" state. The system utilizes a light probe routing approach to detect irregularities and calculates trust estimations for each neighboring node.

In their study, Diro et al. (2014) examined the fog-to-things architecture in order to assess its effectiveness in detecting threats. The article's authors conducted a comparison of deep and shallow neural networks using a publicly available dataset. The main objective of this research was to classify four distinct forms of assault and abnormalities. The accuracy attained for four unique classes is as follows: 96.75% for shallow neural networks (SNN) and 98.27% for deep neural networks (DNN).

Usmonov and colleagues have examined the security concerns that occur with the development of embedded technology for the Internet of Things (IoT). An important issue that was identified is the preservation of data integrity while moving data across the physical, logical, and virtualized components of an IoT system. The paper's authors suggested using digital watermarks as a solution to tackle these challenges.

Anthi et al. (2016) developed a technique for detecting unauthorized access in the Internet of Things (IoT). The research used several machine learning (ML) classifiers to accurately identify network scanning probing and basic forms of Denial of Service (DoS) attacks. The study's data set was created by recording network traffic over a span of four consecutive days using the Wireshark program. The ML classifiers were implemented using the Weka program.

Ukil et al. (2016) conducted study to uncover abnormalities in healthcare data using the Internet of Things (IoT). The research presented a smartphone-compatible model for detecting heart anomalies. The authors used several methodologies, such as IoT sensors, biomedical signal analysis, predictive analytics, medical image analysis, and big data mining, to detect anomalies in healthcare.

Pajouh et al. (2019) proposed an intrusion detection model that utilizes a two-layer dimension reduction and two-tier classification module to accurately detect malicious actions such as User to Root (U2R) and Remote Local (R2L) attacks. The experiment used the NSL-KDD dataset and included dimension reduction techniques such as component analysis and linear discriminate analysis. The focus of the experiment was on U2R (User to Root) and R2L (Remote to Local) assaults. The research used component analysis and linear discriminate analysis to reduce dimensions, using the NSL-KDD dataset for the experiment.

In this study by Angelo et al. (2018), the authors analyzed the binary NSLKDD dataset and Real Traffic Data from Federico II University of Napoli using the Uncertainty-managing Batch Relevance-based Artificial Intelligence (U-BRAIN) approach. The U-Brain model operates dynamically on several computers and has the capability to manage incomplete data. The authors used the J-48-based classification technique to choose six features out of the total 41 characteristics in the NSL-KDD dataset. The research found accuracy rates of 94.1% and 97.4% (10-fold training mean) for NSL-KDD and Real Traffic Data, respectively.

Kozik et al. (2021) proposed a threat detection service that utilizes a cloud architecture and HPC cluster resources to train classifiers, which are time-consuming and expensive. The research focused on the Extreme Learning Machines (ELM) classifier, which allows for efficient calculations and analysis of collected data in edge computing contexts. An analysis was conducted on the structure and qualities of the subject. The major emphasis of the study was on three scenarios of IoT systems: scanning, infected host, and command and control. The study findings for each of these examples indicated accuracy ratings of 0.99, 0.76, and 0.95, respectively.

2. Threat Analysis and Prediction in IOT Devices

Figure 2 illustrates the categorization of the threat analysis and prediction system in the context of the Internet of Things (IoT). The classification of intrusion detection systems (IDS) may be divided into three categories: topology-based IDS, attack-based IDS, and IDS based on the intrusion detection technology used. The intrusion detection approach is categorized into four distinct groups: hybrid IDS, anomaly IDS, specification IDS, and signature IDS. The network structure-based Intrusion Detection System (IDS) may be categorized into three types: Centralized IDS (CIDS), Distributed IDS (DIDS), and Host-based IDS (HIDS). Furthermore, Intrusion Detection Systems (IDS) may be used to identify and detect several sorts of attacks, including denial of service, wormhole, Sybil, fake data injection, reply, and jammer assaults.

3. Machine Learning Algorithms

Machine learning is a field of study focused on creating computer algorithms that imitate the way humans learn, enabling them to automatically gain information. It is a multidisciplinary domain that encompasses computer science, statistics, psychology, and neuroscience. Machine learning algorithms are classified into three types depending on learning approaches: supervised learning, unsupervised learning, and reinforcement learning. Figure 1 depicts the many categories of machine learning (ML) algorithms.

The machine learning framework comprises several autonomous processes, as seen in Figure 2. The first step involves the gathering and observation of data, whereby the information is meticulously gathered and examined to determine its nature. Next, the dataset undergoes data pre-processing, which involves tasks such as visualization, data cleaning, feature engineering, and vectorization. These processes are carried out to transform the information into feature vectors. The feature vectors are then divided into a training set and a testing set, with a ratio of 80:20. The training set is used in the learning method to construct a definitive model via the implementation of an optimization approach. This study used several optimization algorithms for distinct classifiers.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a discriminative model that has similarities with logistic regression. The model is a commonly used supervised learning technique for regression, classification, and outlier identification. Support Vector Machines (SVM) are very advantageous for the analysis of non-linear data.

Decision Tree (DT)

A Decision Tree is an algorithm that allows nodes to evaluate various options by considering their costs, rewards, and probabilities. Essentially, it offers a systematic representation of potential results that arise from a sequence of interconnected decisions. Typically, it begins with a single node and expands into several outcomes, each of which connects to other nodes and generates further branches. Consequently, it has a resemblance to either a tree-like structure or a flowchart.

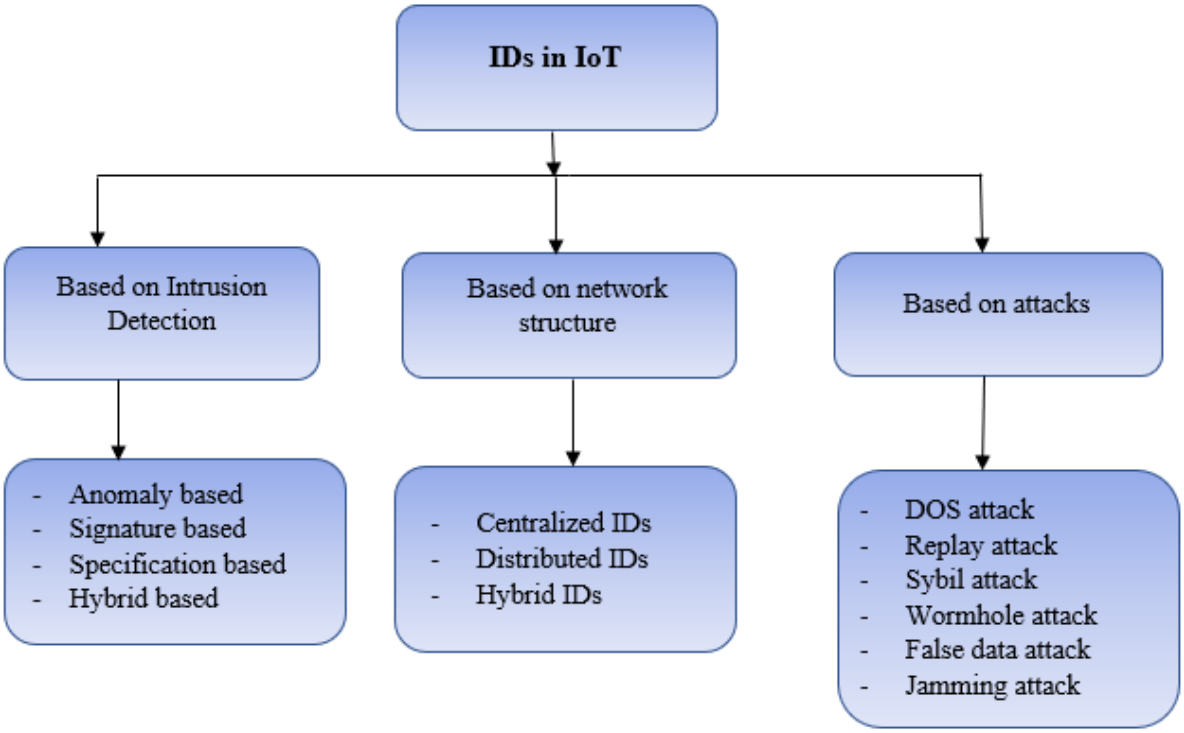

Fig. 1: Taxonomy of IoT

Logistic Regression (LR)

Logistic Regression (LR) is a discriminative model that relies on the quality of the dataset. Logistic regression is a statistical method that use past evaluations of a dataset to forecast a binary outcome, such as a positive or negative response. A logistic regression model predicts the value of a dependent variable by analyzing the relationship between one or more existing independent variables.

Naive Bayes (NB)

Naive Bayes is a widely used machine-learning method used for classification jobs. Bayes' theorem is used to assess the probability of an event by using previous knowledge of relevant factors. Naive Bayes assumes that the presence or absence of each feature is independent and unrelated to the presence or absence of other features. This assumption is referred to as the "naive" assumption. Although it may not always be valid, it streamlines the calculation process and may enhance the efficiency of the algorithm.

Random Forest (RF)

The random forest algorithm constructs an ensemble of decision trees to perform supervised classification.
The term is derived from the fact that every tree is generated in a random manner, exhibiting minor differences in the feature set and data used. The program then calculates the mean of the forecasts generated by each tree in order to get a final prediction. The random forest technique, with its ensemble approach, often exhibits superior prediction accuracy compared to a single decision tree. Moreover, it is renowned for its exceptional processing speed, which makes it a compelling choice for handling large datasets. As the population of a random forest increases, its performance generally improves.

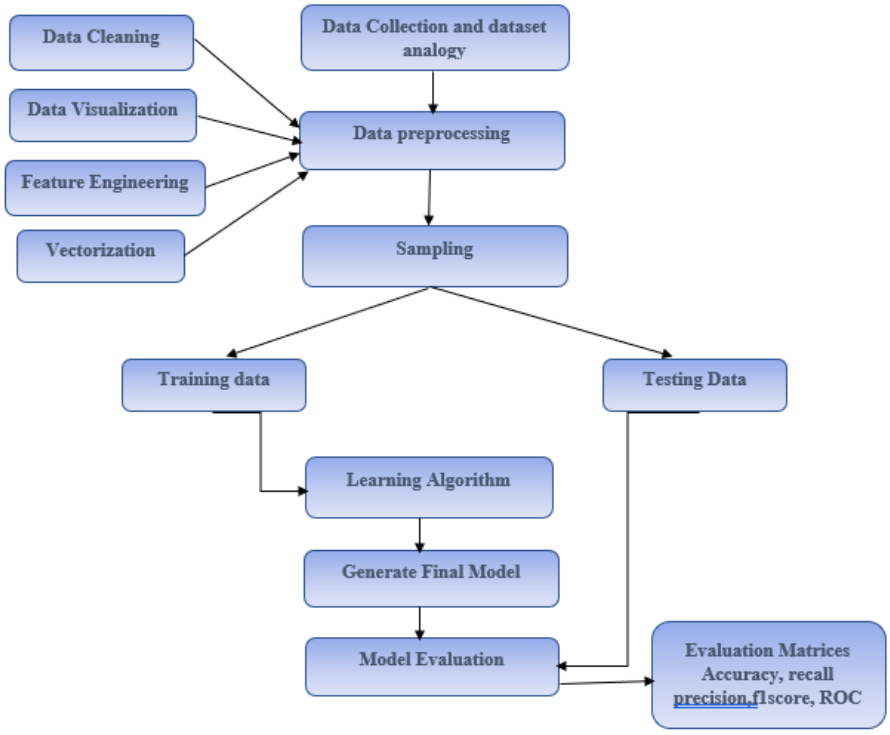

Fig. 2: Overall framework for attack detection in IoT

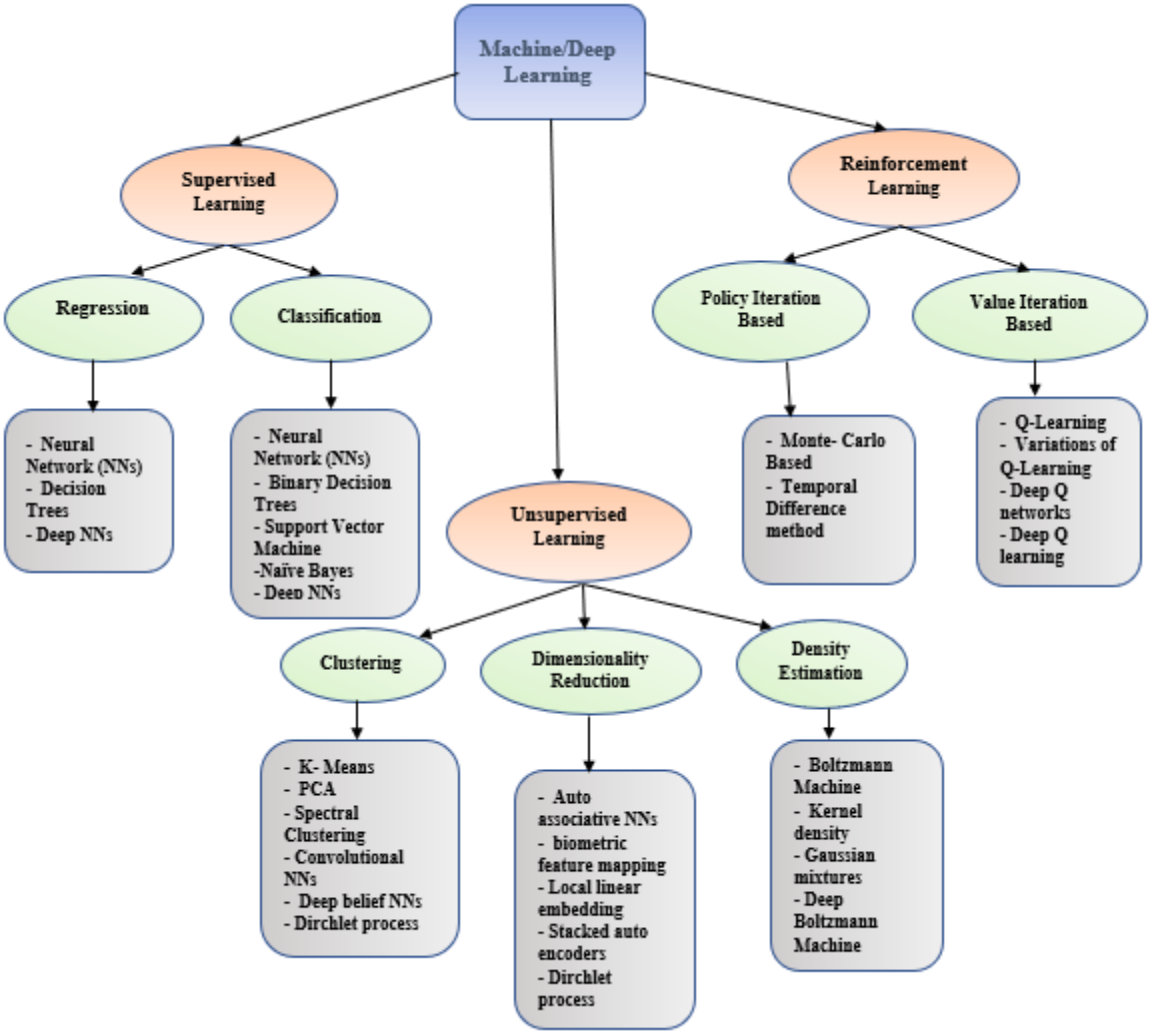

Fig. 3: Family of Machine Learning

Deep Recurrent Neural Network (DRNN)

The Deep Recurrent Neural Network (DRNN) is a neural network design that combines the principles of deep learning with recurrent neural networks (RNNs). DRNNs, or Deep Recurrent Neural Networks, possess a feedback loop similar to that of RNNs, enabling them to effectively handle sequential input. The feedback loop in the network allows it to retain a state or memory of previous inputs, which is crucial for tasks like natural language processing or voice recognition. Deep Recurrent Neural Networks (DRNNs) include numerous layers of neurons, enabling them to acquire hierarchical representations of the input data. This attribute is also seen in deep learning neural networks, which have the ability to acquire intricate characteristics by progressively integrating more basic ones.

Artificial Neural Network (ANN)

The fundamental basis of a deep learning method is an artificial neural network (ANN), which is a machine learning technique. The unprocessed data may be used to train the artificial neural network model. Unlike previous classifiers, this one has a greater number of tuning parameters, which contributes to its more advanced and intricate structure. Furthermore, the process of optimizing the mistake requires a longer duration compared to other procedures. CUDA programming is used to train neural network algorithms on the GPU. Each neuron node of the ANN is trained using a feature set X, which consists of unique properties X1, X2, X3, ..., Xn. The features are augmented with bias values, denoted as b = b1, b2, ..., bn, and multiplied by random weights, represented as W = W1, W2, W3, ..., Wn. Subsequently, the resulting values are provided as input to a non-linear activation function.

Conclusions

This research discovered that the analysis and prediction of threats in the context of the Internet of Things remains challenging. As the Internet of Things (IoT) evolves, the focus shifts from connection to data. To ensure data security, our endeavor focused on the latest advancements in intrusion detection and intelligent Internet of Things (IoT) methodologies. The study primarily focused on analyzing several works that addressed the topic and many efforts made by academics and the industry to develop efficient security methods that provide sufficient protection.

The research incorporates several ingenious methodologies that are used for threat analysis, prediction, and network security in computer networks. While these techniques strive to enhance the accuracy of intrusion detection, it is well acknowledged that the issue of false positives remains a persistent challenge that must be tackled in all research endeavors. Although several strategies may reduce the occurrence of false positives, they need further training and categorization. Nevertheless, several techniques may invert the procedure, ensuring a consistent false positive rate but requiring significant computing resources for both training and testing. This issue has significant importance in the field of intrusion detection, since the ability to detect intrusions in real-time is a crucial factor to consider.

References

Anthi, E. (2018). Pulse: An adaptive intrusion detection for the Internet of Things," in Living in the Internet of Things: Cybersecurity of the IoT - 2018, 2018, pp. 1-4.

Brun, O. (2018). Deep learning with dense random neural networks for detecting attacks against IoT-connected home environments, in Security in Computer and Information Sciences: First International ISCIS Security Workshop 2018, Euro-CYBERSEC 2018, London, UK, February 26-27, 2018, Revised Selected Papers 1, 2018, pp. 79-89.

Chilamkurti, A. A. et al (2018). Distributed attack detection scheme using deep learning approach for Internet of Things, Future Generation Computer Systems, vol. 82, pp. 761-768, 2018.

Hashem, M. H. et al (2019). Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches, Internet of Things, vol. 7, p. 100059, 2019.

Liu, X. et al (2017). Defending ON–OFF Attacks Using Light Probing Messages in Smart Sensors for Industrial Communication Systems," IEEE Transactions on Industrial Informatics, vol. 14, pp. 3801-3811, 218.

Lopez-Martin, M. (2017). Conditional Variational Autoencoder for Prediction and Feature Recovery Applied to Intrusion Detection in IoT, Sensors, vol. 17, 2017.

Pajouh, H. H. et al (2019). A Two-Layer Dimension Reduction and Two-Tier Classification Model for Anomaly-Based Intrusion Detection in IoT Backbone Networks, IEEE Transactions on Emerging Topics in Computing, vol. 7, pp. 314-323, 2019.

Palmieri, R. K. et al (2018). A scalable distributed machine learning approach for attack detection in edge computing environments," Journal of Parallel and Distributed Computing, vol. 119, pp. 18-26, 2018.

Papa, C. R. (2012). An Optimum-Path Forest framework for intrusion detection in computer networks, Engineering Applications of Artificial Intelligence, vol. 25, pp. 1226-1234, 2012.

Rampone, G. D. (2015). An uncertainty-managing batch relevance-based approach to network anomaly detection, Applied Soft Computing, vol. 38, pp. 408-418, 2015.

Sheikhan, H. B. (2017). Hybrid of anomaly-based and specification-based IDS for Internet of Things using unsupervised OPF based on MapReduce approach," Computer Communications, vol. 98, pp. 52-71, 2017.

Ukil, A. et al (2016). IoT Healthcare Analytics: The Importance of Anomaly Detection," in 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 2016, pp. 994-997.

Usmonov, B. et al (2017). The cybersecurity in development of IoT embedded technologies, in 2017 International Conference on Information Science and Communications Technologies (ICISCT), 2017, pp. 1-4.

Wang, C. C. et al (2014) A Framework for Clustering Evolving Data Streams, in Proceedings 2013 VLDB Conference, San Francissco, Morgan Kaufmann, 2003, pp. 81-92.

Zade, S. H. (2020). New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN," Computer Networks, vol. 173, p. 107168, 2020.