International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 07 | July 2023 www.irjet.net p-ISSN: 2395-0072

Deep Learning and Big Data technologies for IoT Security

Pranesh Kathavate[1], Suvidhi Solanki[2]

[1]Student, Computer Engineering Dept, Vishwakarma University, Pune, India
[2]Student, Computer Engineering Dept, Vishwakarma University, Pune, India

---

Abstract - *An As there are millions to billions of connected Internet of Things (IoT) devices and systems sending enormous raw and processed data through the IoT network, we need to be able to use big data analytical techniques and solutions with an effective classification of possible attacks using deep learning in order to protect the security and privacy of IoT data and services from a wide range of attackers. The vast volume of organized, semi-structured, and unstructured data that is produced in the digital era has made the term "big data" popular in recent years. But this vast amount of data is trackable and can be used for financial gain, which violates people's privacy. The attack surfaces of the IoT system are investigated, along with any potential risks related to each surface. The use of deep learning to find security flaws has proved successful in earlier research. The data generated by IoT devices is abundant, diverse, and reliable. We then discuss the advantages, disadvantages, and strengths of each Deep Learning technique for IoT security. As a result, big data technology makes it possible to organize and perform tasks better. In this paper, we have focused on combining the potentiality of Big Data processing with Deep Learning*

Key Words: IoT Security, Big Data, Deep Learning, Botnet, DDoS, Access Control, Malware detection, RNN

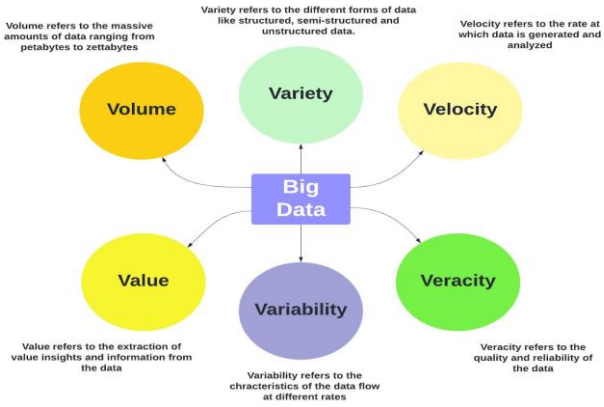
1.INTRODUCTION

The number of common devices equipped with sensors and capable of internet communication has significantly increased during the last few years. According to an article in IoT Business News, the number of linked devices in the IoT globe is predicted to reach about 75 billion by 2030.[1] These devices are capable of recognizing their current situation, sharing, and managing data that can be used for a variety of purposes. The Internet of Things (IoT) was quickly developed and adopted in a variety of industries, including business, agriculture, and the military. Big data is getting bigger and more complicated, especially from new sources. These enormous datasets can't be processed using conventional methods. Despite the overwhelming amount of data, it might be helpful in resolving business issues that you were previously unable to handle. These gadgets are equipped with a range of sensors that enable them to collect data in real time from distant physical devices. For example, the Internet of Things (IoT) has substantially increased traditional detection of surrounding situations. IoT technologies have the potential to gather, quantify, and grasp data about their surroundings, enabling modernizations that improve people's quality of life[3].The Internet of Things will encompass all aspects of our lives, with applications ranging from home automation to smart transportation to smart agriculture to wearable devices and e-Health. IoT devices are easy targets for botnet developers due to their vulnerability. The proliferation of IoT malware presents new security challenges for network managers. In this part, we will first describe the intrinsic characteristics of IoT ecosystems that make them difficult to safeguard. Then, we discuss the most frequent types of vulnerabilities found in IoT ecosystems, as well as how IoT botnets operate. We then explain a number of proposed solutions for IoT security, as well as their drawbacks. Deep learning techniques are widely used for high-dimensional object analysis. This is owing to their high ability to generalize data. The method for adjusting such structures is focused at approximating the formation of such a set that includes both training elements and elements not encountered during the training process. In [4], the author suggested a deep learning-based approach for detecting internet intrusions in the IoT. The authors underline the framework's effectiveness within the context of the ''smart city'' notion. The investigated strategy has the advantage of improving the classification accuracy and creating a trade-off balance between the correctness of attack detection and the speed of this process. In [5}, the research focuses on developing a distributed attack detection system using a deep learning model. The created system's application is the detection of assaults in the social IoT.

1.1 Big Data

Big data is being particularly collected in massive amounts, which raises complexity. Modern technology is unable to handle these large data quantities. Innovative data processing techniques are necessary because of the massive volume of data collected and its quick change. This terminology is required to improve judgment, gain a deeper understanding of processes, and make them function more effectively. We may refer to data as "Big Data" if traditional or modern technology has trouble gathering, processing, storing, filtering, and visualizing it. In a word, "Big Data" technology involves gathering, storing, and extracting knowledge from enormous data volumes. Huge data is a term for a significant volume of

information that is characterized by deep relationships between different data sets and a more complex form of information. Apache applications like Hadoop, Spark, Storm are a viable use-case.



1.2 Deep Learning

Neural networks are used in deep learning to carry out complex calculations on vast volumes of data. It is made up of numerous artificial neural network layers. Some of the neurons in each of the layers have activation functions that can be used to generate non-linear outputs. There are three main types of neural networks namely ANN, CNN, RNN. The neuronal structure of the human brain is allegedly the source of inspiration for this technology.

Three methods namely supervised, semi-supervised, and unsupervised learning are used in deep learning. In supervised learning, the input is correctly labeled and utilized to train the architecture; in unsupervised learning, the input is incorrectly labeled and the architecture tries to build a structure by extracting key information.

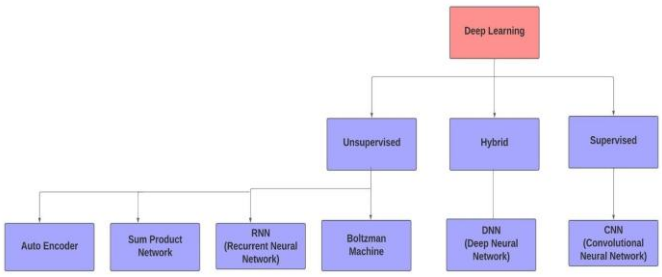


Fig 2. Branches of Deep Learning

1.2 IoT Security

Due to the fact that IoT is widely employed in mobile devices, transportation facilities, public facilities, and home appliances, this equipment can be used for IoT data collection. Furthermore, equipment utilization in various applications that are connected to the IoT network can be operated remotely. IoT security is defined as the protection of the entire IoT deployment architecture from assaults. Various considerations must be considered when building IoT security solutions. The following are the security requirements that must be addressed when building IoT security solutions. Deep learning and big data technologies have enormous capabilities that can be used to detect a pool of security breaches related to security criteria. The connected cyber-physical infrastructure causes IoT vulnerabilities. There is a strong demand for solutions with real-time, scalable, and distributed monitoring infrastructure technologies to reduce IoT security threat that can be used to identify a pool of security breaches related to security requirements. Confidentiality allows information to be securely transmitted during all communications. When information is delivered without authentication or encryption, enemies have the opportunity to breach the owner's privacy. Big data technologies typically include safe data transport via encryption approaches, preventing data from being compromised by adversaries. An adversary may compromise the integrity of an IoT system. As a result, integrity ensures that the data received was not tampered with during transmission. Furthermore, Apache Spark, a big data solution, supports data quality checks in the Spark DataFrame. Users can use this to perform data integrity checks on the IoT system. In IoT systems, availability refers to ensuring that legitimate users may access the system and that illegitimate access is denied. One of the key purposes of big data technology is to assure the user's omnipresence. Furthermore, they can be run on several nodes, ensuring the application's high availability. Authentication is the process of verifying the identity of the peer with whom IoT devices communicate. Furthermore, it is concerned with valid users having suitable network access for network functions such as IoT devices and network control. Furthermore, big data solutions like Apache Spark have authentication techniques for Remote Procedure Call (RPC) channels. Access control in an IoT system should act as a means of ensuring that the authenticated nodes are limited to access what they are privileged to and nothing more. Furthermore, it is known that big data technologies provide access control support for its applications. A filter is necessary for this to be achieved and each application can be equipped with its own access control list.

2. PROBLEM STATEMENT

There have been various sorts of studies in the domain of big data security in IoT contexts. However, it has been discovered that massive data security is a difficult operation. Deep learning algorithms for classification must be employed for attack categorization. Furthermore, an optimization mechanism is necessary to obtain the optimal answer, so that deep learning-based training and testing processes deliver improved accuracy as well as better performance. Filtering a dataset on the basis of

optimal value would result in excellent accuracy and performance.

3. RELATED WORK

The designed system architecture allows for the updating of each cooperative node's learning parameters on a single master node and the distribution of the resulting updates to the worker nodes. According to the research made by the author [6], The deep neural network is considered in the context of mobile big data analytics in the following distributed learning problem. The proposed method was implemented in the Spark environment and consists of the following steps, learning partial models (worker nodes tune neural networks using different data partitions), parameter averaging (model parameters are transmitted to the master node for averaging), and parameter dissemination (updated parameters are passed to worker nodes, which tune neural networks once more). This process is done multiple times until the convergence condition is met. Since a two-layer neural network has a straightforward structure and a respectable learning rate, we have decided to focus on it for the time being. Investigation into the practicality of deep neural networks is planned for the future.

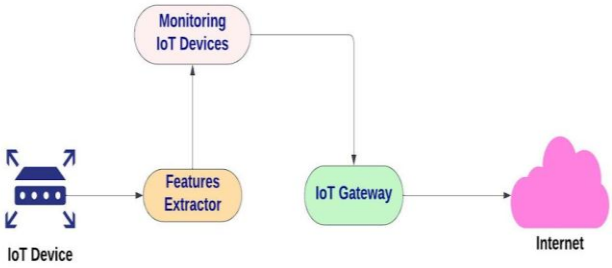
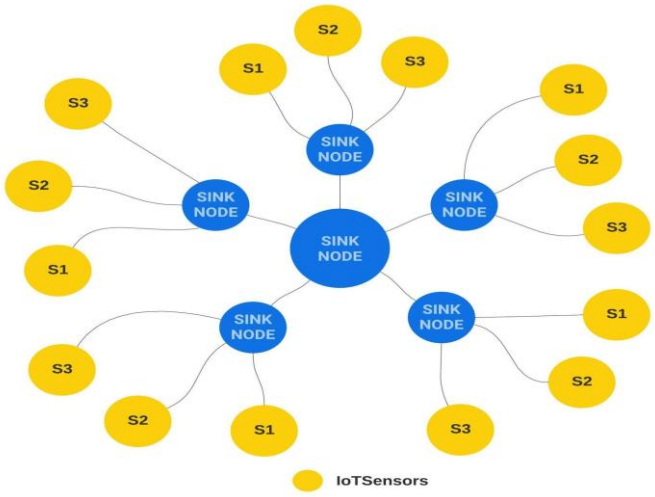


Fig 3. DL feature extraction for IoT Security

Big data analytics can be used in IoT security to identify patterns, anomalies, and potential threats. Deep Learning threat detection with a big data analytics model can be used to detect potential security threats by analyzing data from various sources such as network traffic, device logs, and user behavior. This can help identify abnormal patterns and behavior that could indicate a potential attack. By using real-time analytics, it is possible to monitor IoT devices and networks continuously, which can help detect and respond to threats as they occur. Big data analytics can be used to analyze historical data to identify trends and predict potential threats. This can help security teams to proactively implement security measures to prevent attacks before they happen. Big data analytics can be used to correlate data from multiple sources, including IoT devices, to identify potential security threats. By correlating data from various sources, it is possible to identify patterns and trends that may not be apparent from individual data sources. Apache Hadoop can perform processing of more than terabytes and petabytes of data. From the above figure, at a very large scale, we can see petabytes and exabytes of data being fetched from IoT sensors can be safely stored inside sink nodes and there would be a single junction sink node where data will be stored from all interconnected nodes.In the event of a security breach, big data analytics can be used to conduct forensic analysis of data to identify the root cause of the breach and help prevent similar attacks in the future. HDFS will play a vital role in storing this humongous sensor data.



Apache Hadoop is a batch processing solution that is both scalable and fault-tolerant.Hadoop can handle petabytes of data and allows programs to operate on numerous nodes. Furthermore, log data is broken down into blocks and distributed to Hadoop cluster nodes. Furthermore, Hadoop is popular because of its ability to quickly retrieve and search log data, as well as its scalability, faster data insertion, and fault tolerance. [6]

Apache Spark is a unifying architecture used for distributed data processing. Apache Spark adds a data sharing concept called Resilient Distributed Dataset (RDD) to the MapReduce architecture. Apache Spark can capture and process workloads such as SQL, streaming, machine learning, and graph processing using this extension. [6].

Apache Storm is a real-time calculation system that is open source. Apache Storm makes it possible to process data streams in real time. It can also process millions of tuples per second per node. It is quick, scalable, fault-tolerant, and easy to use. It also has the ability to incorporate databases into the processing.[6].Overall, big data analytics can be a powerful tool in improving IoT security by enabling security teams to detect and respond to threats in real-time, predict potential threats, and conduct forensic analysis to prevent future attacks.

The vulnerabilities from open ended IoT sensors have increased over the period of time. These sensors can be attacked through some of the common malware attack practices such as BotNet, DDOS attacks, routing attacks, middleware attacks etc. A botnet is a network of multiple bots designed to undertake harmful activities on the target network and are controlled by a single unit known as the botmaster via command and control protocol. Bots are infected computers that are remotely controlled by the botmaster and are used to carry out malicious operations. Botnet sizes range from a few hundred bots in a small botnet to 50,000 hosts in a large botnet. Hackers distribute botnet malware and operate in secrecy, leaving no trace of their presence, and can remain effective and operational for years. Communication with bots is required to provide commands to the bots in order for them to carry out destructive operations [7].

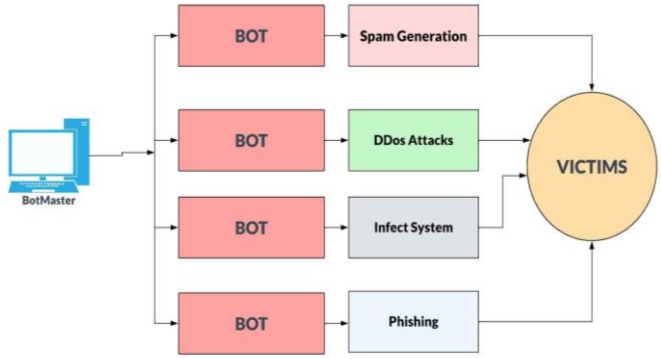


Fig 4. Botnet Architecture

Botmaster always stays hidden in the botnet network by using minimal bandwidth and providing concealed services. Botmaster and bots constantly communicate via command and control server. Bots' principal purpose is to remain inconspicuous until they are necessary to do given duties. Bots are difficult to identify since they do not disturb regular

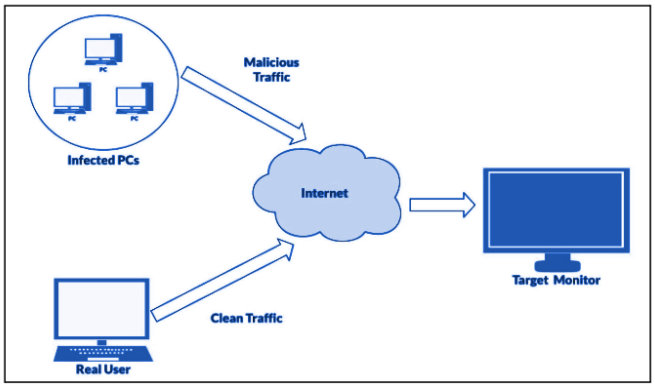


Fig 5. DDos Attacks

host operation and remain silent until they receive a command from the botmaster to undertake assigned operations. The botnet's life cycle is divided into numerous stages, including dissemination and infection, secondary injection, connection, malicious command and control, update, and maintenance. DDoS attacks are the most popular type of cyber-attack in which the attacker's computers simultaneously transmit a significant amount of malicious traffic to the target server in order to overwhelm the target network [7]. DDoS attacks aim to dramatically disrupt regular server operation by flooding the target device with large traffic, such as fraudulent requests, in order to oversaturate its capacity, causing disruption or denial of service to legitimate traffic. DDoS assaults have an impact on the server's system resources such as CPU and memory, and can also cause the network bandwidth to become oversaturated with a significant volume of data. As a result, genuine computers will be denied service since the server is preoccupied with dealing with the DDoS attack. DDoS attacks are launched by hackers using a botnet. Following below table demonstrates the most commonly detected attacks at various layers of execution.

Table 1. A list of some IoT security layer attacks

Physical Layer Attacks	**Network Layer Attacks**	**Application Layer Attacks**
Botnet	Man in the Middle	Malware
Sleep Deprivation	Denial of Service (DOS)	Phishing
Node Tampering and Jamming	Routing Attacks	Code Injection
EavesDropping	Middleware Attacks	BGP Hijacking

Datasets are one of the basic ways to practice and study detection of several attacks through processing some of the commonly found patterns in Big Data being fetched from the IoT Sensors. The following datasets are frequently used for experimental analysis on deep learning, big data technologies, and/or IoT security or network security.

IoTPOT: The IoTPOT dataset contains IoT network traffic, regular and malware-based network traffic, which is primarily utilized in DDoS assaults. The dataset is divided into five malware families: ZORRO, GAYFGT, KOS[7]

Kyoto: In 2006, the Kyoto dataset was created for Intrusion Detection System (IDS) research. This dataset is based on three years of real-world network traffic data. This dataset also includes 14 features obtained from the

KDDCUP99 as well as an additional 10 features. Furthermore, their honeypot data includes 50,033,015 normal sessions and 43,043,255 assault sessions. Furthermore, the three attack kinds, exploits, shellcodes, and malware, are covered.

CICIDS2017: The Canadian Institute for Cybersecurity (CIC) generated the CICIDS2017 dataset in 2017. It comprises data from real-world benign and malicious network traffic. This dataset has 225,746 records and 80 characteristics in total. Furthermore, this dataset includes Brute Force, DDoS[7].

CIDDS-001: CIDDS-001 is a labeled flow-based dataset designed for NIDS evaluation based on anomalies. The dataset includes normal and attack traffic statistics collected over a four-week period. This dataset also includes 14 features and four types of assaults, including DoS, PortScan, Brute Force, and Ping Scan. Sometimes, detecting advanced issues could not be possible by just merely going through the processing of the datasets mentioned above. We can achieve enhanced assault detection by making use of certain mathematical algorithms. Deep Learning based mathematical solutions provide a way to detect complex and difficult patterns used by hacking sources for IoT Security

The Support Vector Machine

The SVM is built by establishing a separating hyperplane with the property of equidistance from objects of different classes that are close to it. The algorithmic expression is depicted below:

$$F^{(1)}(\mathbf{z}) = \text{sign}\left(-b + \sum_{i=1}^{M_s} w_i \mathbf{x}_i^T \mathbf{z}\right)$$

where wi – weight coefficients, which are the product of nonzero Lagrange multipliers and the desired output values, xi – support vectors (i = 1, . . . , Ms), b – bias parameter.

KNN (K-nearest neighbors) method

The KNN can match the analyzed vector to a label of that class, whose instances have the highest number among all K training items that are closest to the provided vector z. Formally, the methodology is as follows:

$$F^{(2)}(\mathbf{z}) = \arg\max_{c \in \Omega} \sum_{i=1}^{K} [\mathbf{x}_i' \in c],$$

where x 0 1 , ... , x 0 k – training vectors for which the value PK i=1 z – x 0 i is minimal among all training vectors, omega classes.We may state that this strategy does not necessitate any prior training. It is sufficient to keep the whole training sample for it to function

The two layer Artificial Neural Network (ANN)

A two-layer artificial neural network is a layered structure that performs linear and nonlinear transformations on the input vector. The composition of the non-linear activation function and the weighted sum of the components of that vector that is output for the (k 1)-th layer is conducted after passing through each k-th layer.

$$F^{(4)}(\mathbf{z}) = \varphi\left(\theta_1^{(2)} + \sum_{i=1}^{N_1} w_{1i}^{(2)} \cdot \varphi\left(\theta_i^{(1)} + \sum_{j=1}^{n} w_{ij}^{(1)} \cdot z_j\right)\right)$$

customizable parameters are there in the learning process, ϕ - activation function, θ (1) and Hθ (2) are bias parameters.

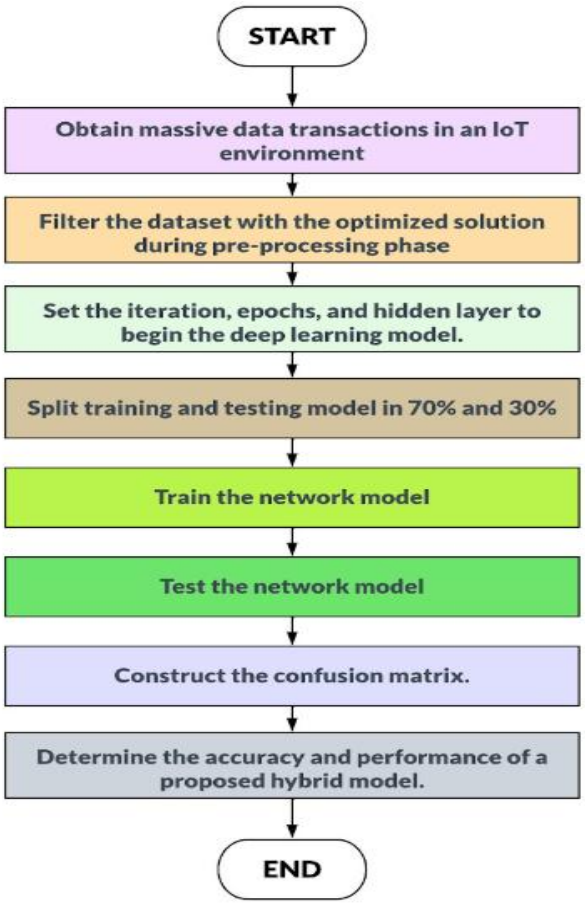


Fig 6. Algorithm detecting IoT security attacks through deep learning

The above algorithm is a high-level understanding of categorization of the raw amount of massive data fetched from IoT sensors in the sink nodes of the hadoop cluster. The algorithm works in an efficient way through fetching terabytes and petabytes of data transactions through various physical IoT devices which are connected in an integrated environment. Data gets transformed and collected into a CSV or Excel file format to perform further

processing. As the obtained data consists of a lot of unstructured or unfiltered content, we basically create and define an optimized solution which helps to appropriately filter the data values during the processing phase and gets ready for the conventional deep learning execution.

In the following table below, we have outlined the research that has included deep learning and big data technologies, including big data technologies, deep learning architectures and IoT security related possible threat detections.

Table 2. Big data technologies with Deep Learning architectures

Big Data Technologies	Deep Learning Model Architecture	IoT Security Domain
Apache Spark	RNN	Intruder detection
Apache Spark	ANN	DDoS attack detection
Apache Storm	CC4 neural network and MLP	Real time intrusion detection
Apache Hadoop	DNN	Malware detection

3. CONCLUSIONS

The growing number of IoT devices has increased the amount of awareness of the security dangers associated with them. IoT devices have been shown to be vulnerable as a result of recent increased attacks such as the Carna and Mirai botnets. Furthermore, IoT devices generate a significant volume, velocity, and variety of data. Existing solutions become less efficient as a result, necessitating the use of modern solutions. Deep learning has gained popularity among academics and organizations due to its high accuracy, capacity to learn deep features, and lack of human supervision.

Furthermore, big data technologies have sparked attention due to their ability to process enormous volumes of data as well as their ability to process data in a range of scenarios such as real-time, batch, and stream.Our findings show that many studies have combined deep learning with IoT security or deep learning with big data technologies; however, there is a lack of research in combining deep learning and big data technologies for IoT security; however, our investigations revealed that two studies have demonstrated the efficiency and feasibility of combining deep learning and big data technologies for IoT security over traditional models.

4. REFERENCES

[1] R.H. Weber, Internet of things–new security and privacy challenges, Comput. Law Secur. Rep. 26 (1) (2010) 23–30.

[2] Parker2005, "The IoT in 2030: 24 billion connected things generating $1.5 trillion"

[3] P. Mishra, V. Varadharajan, U. Tupakula, E.S. Pilli, ADetailed Investigation and Analysis of Using MachineLearning Techniques for Intrusion Detection, IEEECommunications Surveys & Tutorials, 2018.

[4] Y. Zhou, M. Han, L. Liu, J. S. He, and Y. Wang, "Deep learning approach for cyberattack detection," in Proc. IEEE INFOCOM IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), Apr. 2018, pp. 262–267.

[5] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," Future Gener. Comput. Syst., vol. 82, pp. 761–768, May 2018.

[6] Mohamed Ahzam Amanullah, Riyaz Ahamed Ariyaluran Habeeb, Fariza Hanum Nasaruddin, Abdullah Gani, Ejaz Ahmed, Abdul Salam Mohamed Nainar, Nazihah Md Akim, Muhammad Imran, "Deep learning and big data technologies for IoT security", Computer Communications, Volume 151, 2020, Pages 495-517, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2020.01.016.

[7] Pokhrel, Satish & Abbas, Robert & Aryal, Bhulok. (2021). "IoT Security: Botnet detection in IoT using Machine learning."