Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering pp. 299–303 ISSN 2300-5963 ACSIS, Vol. 33
DOI: 10.15439/2022R39

Application of Machine Learning in Malicious IoT Classification and Detection on Fog-IoT Architecture

Duong Thi Van
Institute of Information Technology
Vietnam Academy of Science and Technology)
Hanoi, Viet Nam
dtvan@ioit.ac.vn

Tran Ba Hung
Institute of Information Technology
Vietnam Academy of Science and Technology)
Hanoi, Viet Nam
tbhung@ioit.ac.vn

Tran Duc Thang
Institute of Information Technology
Vietnam Academy of Science and Technology)
Hanoi, Viet Nam
thang@ioit.ac.vn

Nguyen Khac Giao
Institute of Information Technology
Vietnam Academy of Science and Technology)
Hanoi, Viet Nam
giaonk@ioit.ac.vn

***Abstract*—Due to the limitations in self-protection and information processing capabilities at IoT (Internet of Things) nodes, these nodes are susceptible to attacks, turning them into malicious nodes that cause damage or danger to the system. Early detection of these threats is essential to make timely recommendations and limit severe consequences for individuals and organizations. The study proposes applying a machine learning model to detect malicious traffic and IoT devices, which can be deployed and applied on the Fog IoT platform. This solution helps detect and early warn threats from IoT data before they are sent to the cloud. The model is evaluated on the IoT-23 dataset and gives good results.**

***Keywords*—Fog Computing; Security IoT; Malicious IoT Devices; Fog Node.**

I. INTRODUCTION

The world is fighting to transition to the Internet of Things (IoT) society [1], where everything can be connected via the Internet. The help of smart devices, embedded devices, and sensors makes life and work more efficient and productive. More and more individuals and organizations use the services that IoT brings. Therefore, securing data and providing user privacy is of great interest.

IoT devices are not standing alone. They are often part of Cloud computing architecture (Cloud computing - CC) to provide on-demand services such as resources, storage, and services [2] ... Users only need to request services without regard to the construction, management, or installation behind it. As technology develops and the number of IoT devices increases rapidly. According to iot-analytics.com, about 50 billion IoT devices are connected to the Internet [3].

Unlike standard personal computers, IoT nodes are limited in processing capacity and storage capacity, so they need support from the Cloud server for analysis and decision-making. Data collected from IoT devices is continuously sent to the cloud. In some cases, the amount of data obtained is huge, fighting, leading to overload, and the response time from the cloud is not enough to meet many real-time applications. To overcome the above limitation, CISCO first proposed the concept of Fog Computing - FC [4]. Fog Computing is a new platform extending new cloud services closer to IoT devices. Fog computing has gained additional advantages over cloud computing, making it a remarkable platform: Utilizing device resources, low latency, distributed instead of centralized Centralized and supporting real-time applications. Therefore, fog computing helps to facilitate the deployment of many applications.

Besides inheriting the advantages of Cloud Computing, this technology has its limitations, so the number of cyber attacks still increases in frequency and intensity [5]. Traditional security solutions, such as firewalls and intrusion detection systems, are not enough to overcome this drawback. Applying machine learning to detect threats and attacks based on network traffic is still a topic of interest to researchers.

In this study, we use a new machine learning model compared to traditional models, XGBoost [6] to detect anomalies and provide early warning of malicious IoT risks. Ongoing damage/type of attack for timely warnings. We also assess malicious IoT devices based on the above results to find out which devices have been infected with malicious code or are the source of the attack. We deploy simulation on Fog-IoT architecture, collect data from the IoT layer, and aggregate it into analytical characteristics, thus predicting whether IoT node activity is benign or malicious.

The next structure of the paper is as follows: In Part II, we give an overview of the Fog-IoT architecture. Some related studies are presented in section III. The proposed method and experimental results are presented in sections IV and V, respectively. Finally, conclusions and development directions.

II. FOG-IOT ARCHITECTURE

A. *Fog-IoT Architecture*

Ants Fog - IoT architecture consists of three layers [7]:

- Device layer: Consists of IoT devices that are physically distributed, with limited computing power and storage resources. These devices often collect and send raw data to a cloud server for storage and processing.

- Fog layer: Includes devices such as switches, routers, and access points. This layer is located between the device layer and the cloud server layer, capable of processing and calculating data before uploading to the cloud. The fog layer provides various services and real-time analytics for data from IoT devices.

- Cloud layer: The cloud layer includes many high-performance servers and long-term data storage capabilities, providing a variety of applications and services. These services are designed to be accessible anywhere and anytime.

B. Security issues in Fog –IoT

Fog Computing has many advantages in improving service quality, but new problems related to security and privacy arise. Regarding computing power, it is difficult for FC to implement a complete set of security solutions to detect and prevent attacks. Additionally, due to its proximity to IoT devices, Fog is first vulnerable to malicious IoT attacks. Besides, since FC gets data from IoT devices and the cloud, it becomes a potential target.

There are some attack techniques on layers of IoT. On the perception layer, attackers can exploit eavesdropping, spoofing, or radio frequency jamming techniques. The Network layer can face attacks such as denial of service, malicious code injection, or man in the middle. On the support and application layers, hackers can execute denial-of-service attacks, malicious code injection, or eavesdrop on packets.

Since many IoT devices collect and exchange sensitive and private data, data security is a central issue. Several solutions to ensure the safety of IoT have been proposed. However, security threats are increasingly complex due to the heterogeneous diversity and increasing number of IoT. Existing solutions need to be continuously improved to deal with new security threats in IoT systems.

III. RELATED STUDIES

The rapid increase in data collected from IoT devices and different communication protocols has increased security risks, demonstrating the need for an effective IDS system. Researchers have focused on anomaly detection methods using new techniques such as machine learning and deep learning. In the past, some researchers have used the KDD99 or NSL-KDD dataset to identify malicious behaviors. The survey's main findings underscore the need for a legitimate and modern dataset to get accurate outputs.

In an IoT node attack early detection study, Y. Meidan et al. [10] performed a snapshot of the traffic behavior for every IoT device to extract the attributes. These attributes are used as input to deep learning techniques to detect anomalies. They trained an autoencoder (one encoder per device) to learn the normal operation of the IoT device. Using an autoencoder is learning complex patterns and minimizing false alarms.

A recent study by Layla Albdour et al . [8] uses a crawler that acts as a security checker to monitor IoT nodes and collect data streams to analyze the behavior of the nodes. Based on that, to put the fake IoT alert action. However, placing the behavior analyzer at all Fog nodes in the system (distributed) is not efficient in terms of time. In addition, the algorithm becomes complicated when it has to process simultaneously—several jobs in different places. In addition, if an attack occurs at a branch Fog system, it will be challenging to take action to respond on time.

Hasan, Islam, Zerif, et al. [12] have implemented machine learning algorithms to detect if the system is behaving abnormally. If so, they use algorithms to detect the type of attack that the system is not behaving device in progress. DS2OS dataset [11] was used for evaluation. The Random Forest algorithm is the best choice, with an accuracy of 99.4%.

In their research, Idrissi and his colleagues [14] learned about security vulnerabilities in IoT. They identify several vulnerabilities and threats on IoT devices, thus offering recommended solutions. They use a neural network-based intrusion detection technique to solve the problem and achieve satisfactory results.

In this paper, we propose to apply the XGBoost machine learning model to detect malicious network traffic from IoT devices. XGBoost is a model that previous studies have not fully considered. At the same time, we also tested a new job to detect malicious IoT devices based on the results of malicious traffic classification. Our contribution to detecting and removing malicious IoT devices in the Fog IoT network.

IV. PROPOSED METHODS

The number of IoT devices in the Fog-IoT network can be huge, including malicious and benign nodes. The research objective is to quickly detect network traffic generated by malicious nodes at the Fog layer. At the same time, scan and warn those malicious nodes to the administrator. Our proposed algorithm aims to solve the above two problems.

A. Anomaly detection algorithm

In this problem, we propose to apply a recently published algorithm model called XGBoost (Extreme Gradient Boosting). It is an algorithm developed based on gradient boosting, with many improvements in terms of algorithm optimization. XGBoost is suitable for many problems, such as regression, classification, and ranking.

The algorithm is as follows:

- Data: Where n is the number of training samples, m is the number of attributes, $D = \{(x_i, y_i)\}$ is the data set. With x_i is the th data i and y_i is the label corresponding to $1 \leq i \leq n$. We have:

$$|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R} \quad (1)$$

- The tree structure includes: q is the structure of a tree, f_k is the structure of k independent trees, with w_i is the weight of the i-th leaf node.

We have the following learning model:

$$\widehat{y_i} = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (2)$$

With K is the space of independent tree structures. The learning model determines the label of the value x_i based on the computation on each objective function in turn f_k. The final aggregate results help to find the label for the data.

- F is the objective function, expressed as follows:

$$F = \{f(x) = \omega_{q(x)}\} \quad (3)$$

Where $\omega_{q(x)}$ is the weight of the node $q_{(x)}$.

- L is a learning function, calculated by the following formula:

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (4)$$

The learning function L calculates the difference between the correct and predicted labels, determines the model weights, and evaluates the convergence of the model.

The XGBoost algorithm scales down the leaf nodes, improving the model's generality. Previous studies have shown the effectiveness of XGBoost in classification and regression problems.

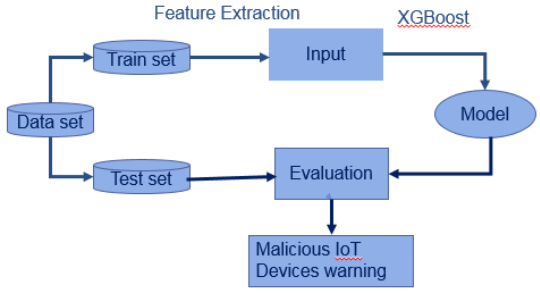

Fig. 1. Application model of the XGBoost algorithm

The flowchart of the proposed solution is described in Fig. 1. In which the Train Set data is attribute extracted and trained by XGBoost, the results are evaluated on the Test Set. Based on the labeling results, malicious IoT devices will be detected and alerted to users.

B. *Deployment model on Fog-IoT*

In this study, we propose a security control model on the Fog-IoT architecture. Adding a Server node behind the Fog layer monitors information flows when sending to the cloud, as shown in Fig. 2.

The Fog layer performs the function of processing information collected from the IoT layer. Data before going to the cloud will be analyzed and monitored at the server. Here, the server will do two things: (1) Detect malicious network traffic and stop them; (2) Report malicious IoT nodes.

V. EXPERIENCE AND ASSESSMENT

A. *Experimental data set*

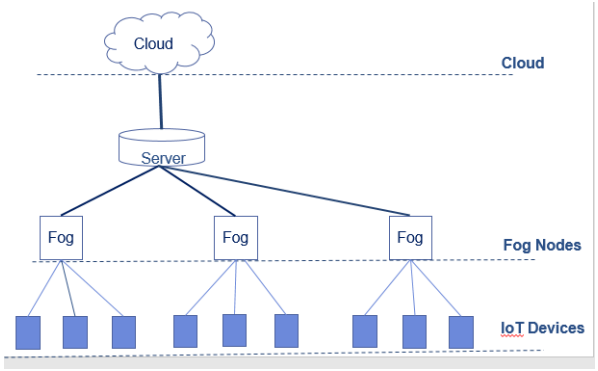

Fig. 2. Proposed security analysis model on Fog - IoT

This study uses the IoT-23 dataset [15], which Avast AIC lab generated. The dataset contains 20 types of malicious traffic logged from different IoT devices, which was collected from 2018 to 2019. The flow labels are the type of malicious traffic contained in the dataset, generated in the Stratosphere lab, and include the following: Attack, Benign, C&C, DDoS, FileDownload, HeartBeat, Mirai, Okiru, PartOfAHorizontalPortScan, Torii.

Collected network traffic for benign situations is obtained by collecting the network traffic of three IoT device types. These are actual deployed IoT devices, not simulators. This allows the demonstration of an IoT network model similar to reality. Evaluation results on this data set are also more reliable.

B. *Experimental results*

We perform our evaluations using the Python language, machine learning support libraries, and execution on the Google CoLab platform, Intel Xeon Processor 2.3GHz chip. The algorithm used is XGBoost.

The results of classification experiment on the IoT-23 dataset are given in Table I below:

TABLE I. CLASSIFICATION RESULTS OF NETWORK TRAFFIC

	Precision	**Recall**	**F 1 - score**	**support**
Attack	0.99	0.98	0.98	783
Benign	*0.95*	*0.56*	*0.71*	*39951*
C&C	1.00	0.11	0.20	3020
C&C-FileDownload	0.57	0.89	0.70	9
C&C-HeartBeat	0.88	0.21	0.34	70
C&C-HeartBeat-FileDownload	0.00	0.00	0.00	2
C&C-Torri	0.00	0.00	0.00	6
DDOS	1.00	0.82	0.90	27755
FileDownload	1.00	0.50	0.67	2
Okiru	0.48	0.00	0.00	52538
PartOfAHorizontalPortScan	0.68	1.00	0.81	165188
Accuracy			**0.73**	289324

The evaluation results show that the XGBoost algorithm effectively detects various types of malicious traffic. Specifically, the Attack, C&C, DDOS, and FileDowload labels were detected with a very high rate, and Precision reached 0.99 or higher. However, the Recall rate is low in the case of C&C and FileDownload. Overall, the F1-score achieved is very good with the Attack and DDOS labels, with 0.98 and 0.90, respectively. It should also be noted that these labels have high Support. They make up most of the dataset.

Some other labels have low detection rates, such as Okiru (Precision is 0.48) and C&C's compound labels with malicious code on IoT. This could be explained by some of the labels compounded with C&C detected in the C&C label. At the same time, the number of samples for these labels is minimal, with the lowest being 02 and the highest being 70, resulting in the model lacking data to train. Overall, this low rate is acceptable because of its low specificity in the dataset.

The result correlation of the evaluation parameters of the network traffic labels is shown in Fig. 3 below.

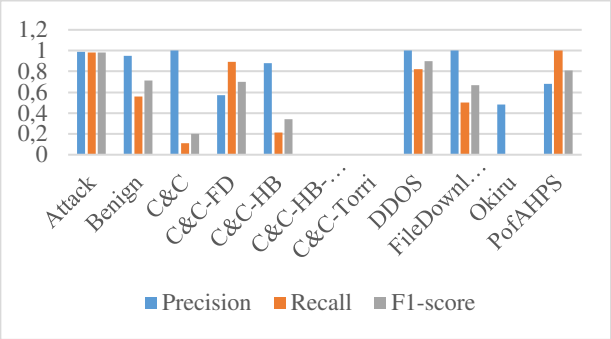

Fig. 3. Correlation of evaluation parameters of labels

The algorithm also performed well when correctly identifying benign data, labeled Benign, with Precision, Recall, and F1-scores reaching 0.95, 0.56, and 0.71, respectively. The support number is 39951.

We conduct a k-fold evaluation with k=5 to get a more comprehensive assessment of the overall accuracy of the XGBoost algorithm. The results per fold are given in Table II as follows:

TABLE II. RESULTS OF MODEL EVALUATION THROUGH EACH I-FOLD

i-fold	Precision	Recall	F1
0	0.7305	0.7305	0.7305
first	0.7304	0.7304	0.7304
2	0.7302	0.7302	0.7302
3	0.7304	0.7304	0.7304
4	0.7315	0.7315	0.7315
AVG	*0.7306*	*0.7306*	*0.7306*

Overall, the XGBoost model has an accuracy of 0.73. This is a pretty good detection rate in the case of the IoT-23 dataset.

We exclude benign network traffic with the Benign label to detect malicious IoT devices. The remaining traffic is identified as malicious, and look for the IPs of those malicious IoT devices. The ratio of malicious IPs to the total IPs is 72.53%. In which the detection results of malicious IoT devices are given in Table III as follows:

TABLE III. SEARCH RESULTS FOR MALICIOUS IP ADDRESSES

	0	**first**
0	<.>	57
first	4	<524>

Label 0 represents benign IoT devices, and label 1 represents malicious IoT and belongs to the group of IPs suggested by the algorithm. The evaluation parameters Precision, Recall, and F1-score have values of 0.90, 0.99, and 0.94, respectively. This result shows that the model effectively detects malicious IoT devices and allows suggestions to prevent malicious behavior from these devices.

C. Comparison with some other machine learning models

Comparing the evaluation results on the IoT23 dataset above with some other machine learning models, including NB, ANN, and SVM by Stoian et al. [16], the accuracy results are shown in Table IV:

TABLE IV. COMPARISON OF CLASSIFICATION RESULTS WITH SOME OTHER MACHINE LEARNING ALGORITHMS

	XGBoost	**NB**	**ANN**	**SVM**
Accuracy	0.73	0.25	0.52	0.59

It can be seen that XGBoost improves the accuracy a lot over models like NB, ANN, and SVM, with an accuracy of 0.73 compared to 0.25, 0.52, and 0.59, respectively. In general, XGBoost is an algorithm with many advantages over NB, ANN, and SVM algorithms.

VI. CONCLUSION AND DEVELOPMENT

Since IoT nodes collect sensitive information from users, securing Fog-IoT systems is a matter of life and death. Furthermore, the increasing demand for use due to IoT's conveniences leads to increased security vulnerabilities. Early identification of an IoT node as malicious so that appropriate action can be taken before an attack occurs is a method of constant concern.

The study applied a new machine learning model, XGBoost, to detect malicious network traffic, blocking them at the Fog layer before sending them to the cloud. The algorithm also allows for finding IoT devices with malicious behavior, which is a hint for system administrators to review and handle them. Overall, the XGBoost model gave positive results with 73% accuracy when detecting malicious network traffic while finding malicious IoT devices with 90% accuracy.

In the future, this study needs to be applied experimentally on large data sets and more diverse. From there, it is possible to develop an appropriate security policy for the Fog - IoT system, improve safety, and ensure customer service quality.

VII. ACKNOWLEDGMENT

This research was partly supported by the project CS21.12 managed by the Institute of Information Technology, Vietnam Academy of Science and Technology.

REFERENCES

[1] M. S. Roopa, S. Pattar, R. Buyya, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "Social Internet of Things (SIoT): Foundations, thrust areas, systematic review and future directions," Comput. Commun., vol. 139, pp. 32–57, 2019, doi: 10.1016/j.comcom.2019.03.009.

[2] H. Sabireen and V. Neelanarayanan, "A Review on Fog Computing: Architecture, Fog with IoT, Algorithms and Research Challenges," ICT Express, vol. 7, no. 2, pp. 162–176, 2021, doi: 10.1016/j.icte.2021.05.004.

[3] K. L. Lueth, "IoT market analysis : Sizing the opportunity," IoT Anal., no. March, pp. 0–12, 2015, [Online]. Available: http://iot-analytics.com/iot-market-forecasts-overview/.

[4] S. Khanagha, S. Ansari, S. Paroutis, and L. Oviedo, "Mutualism and the dynamics of new platform creation: A study of Cisco and fog computing," Strateg. Manag. J., vol. 43, no. 3, pp. 476–506, 2022, doi: 10.1002/smj.3147.

[5] A. Aljumah and T. A. Ahanger, "Fog computing and security issues: A review," 2018 7th Int. Conf. Comput. Commun. Control. ICCCC 2018 - Proc., pp. 237–239, 2018, doi: 10.1109/ICCCC.2018.8390464.

[6] T. Chen, T. He, and M. Benesty, "XGBoost : eXtreme Gradient Boosting," R Packag. version 0.71-2, pp. 1–4, 2018.

[7] S. P. Ahuja and N. Wheeler, "Architecture of Fog-Enabled and Cloud-Enhanced Internet of Things Applications," Int. J. Cloud Appl. Comput., vol. 10, no. 1, pp. 1–10, 2019, doi: 10.4018/ijcac.2020010101.

[8] L. Albdour, S. Manaseer, and A. Sharieh, "IoT crawler with behavior analyzer at fog layer for detecting malicious nodes," Int. J. Commun. Networks Inf. Secur., vol. 12, no. 1, pp. 83–94, 2020, doi: 10.17762/ijcnis.v12i1.4459.

[9] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, "Attack and anomaly detection in IoT sensors in IoT sites using machine

learning approaches," Internet of Things (Netherlands), vol. 7, 2019, doi: 10.1016/j.iot.2019.100059.

[10] M.-O. Pahl and F.-X. Aubet, "DS2OS traffic traces | Kaggle," [Online]. Available: https://www.kaggle.com/francoisxa/ds2ostraffictraces.

[11] I. Idrissi, M. Azizi, and O. Moussaoui, "IoT security with Deep Learning-based Intrusion Detection Systems: A systematic literature review," 4th Int. Conf. Intell. Comput. Data Sci. ICDS 2020, 2020, doi: 10.1109/ICDS50568.2020.9268713.

[12] Q. Tian, J. Li, and H. Liu, "A Method for Guaranteeing Wireless Communication Based on a Combination of Deep and Shallow Learning," IEEE Access, vol. 7, pp. 38688–38695, 2019, doi: 10.1109/ACCESS.2019.2905754.

[13] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic," Zenodo., 2020, [Online]. Available: http://doi.org/10.5281/zenodo.4743746.

[14] N. A. Stoian, "Machine Learning for Anomaly Detection in IoT networks: Malware analysis on the IoT-23 Data set," Univ. Twente, 2020.