Network Attack Detection in IoT using Artificial Intelligence

Muhammad Jahanzaib Gul
Dept. of computer science
Muhammad Ali Jinnah University
Karachi, Pakistan
fa20mscs0031@maju.edu.pk

Muhammad Khaliq-ur-Rahman Raazi Syed
Dept. of computer science
Muhammad Ali Jinnah University
Karachi, Pakistan
raazi.m.syed@ieee.org

2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT) | 979-8-3503-4770-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMCERT57083.2023.10075102

***Abstract*— We like to have simple and automated solutions, but these simple and automated solutions in technology could also contains risks if not deal properly. Due to no international standard of compatibility for IoT, security and privacy concerns are there which needs to be focus. There can be multiple types of attack on IoT networks which can damage the device or steal the sensitive information. Therefore, artificial intelligence (AI) techniques has an ability to detect and classify an unknown network behaviour by learning the network attacks patterns based on large volumes of historical data. We considered Aposemat IoT-23 which is a labelled dataset and created in the Avast laboratory. Basically, the goal of this large dataset is to provide labelled and real IoT attacks. In this paper, we used this dataset, considered the relevant workings, investigate the background and implement the machine learning algorithms such as Decision Tree, Random Forest and Naive Bayes. We also compared the accuracy among these machine learning algorithms on the IoT-23 dataset and showed the most efficient machine learning algorithm is Random Forest as per results by using Aposemat IoT-23 dataset, as well as showed feature engineering techniques to preprocess the mentioned dataset for detection and classification of IoT network attacks.**

Keywords*— *Network Attack Detection, Malware Detection, Malware in IoT, Machine Learning Classifiers, Aposemat IoT-23 dataset

I. Introduction

Today, we are living in the society where many things are going to be automated and digitalize. Technology is now involving in our daily life and there are many simple examples for that such as mobile phones, personal computers etc. Converting things to smart devices and making these processes automated, IoT is one of the technology which plays an important role for that purpose. So we can say that it is one of the most important technologies for businesses as well as for our daily life. But, it is important to remember that as the technology increases there are also a number of issues increases related with that technology. Similarly, as the number of devices connected it means the more information is sharing between these devices and if there is any type of bug in the sharing system, there is a chance that each connected device could corrupt and confidential information could steal by the hacker.

There should be an international standard for compatibility of IoT here which is not yet, therefore it is very difficult for devices which are manufactured from different companies to communicate with each other. Also there are many IoT devices which requires and ask to input user personal information such as name, location and contact as well as data which are important to hackers such as social media information. Therefore, the information sharing between IoT devices needs to be secured. Also IoT privacy and security are cited as major concerns. There are number of attacks on IoT including malware. Malware can be defined as a malicious software or bug which is designed to gain access and damage your device, device could be computer or IoT device. IoT devices are vulnerable to network attacks therefore, malware and network attack detection in IoT is the focus of research in recent years.

There are many workings are there to address the issue and detect network attacks. In comparison, ML and DL which can be defined as machine learning and deep learning in artificial intelligence has the power to detect unknown network behavior by automatically learning the networks attacks and malware patterns based of large datasets. In this paper we will focus on the security aspect of networks of IoT by understanding the use of machine learning based algorithms in artificial intelligence for the detection of network attacks and malwares. For this purpose, we will consider Aposemat IoT-23 which is a labeled dataset and created in the Avast laboratory. This dataset also provides benign IoT traffic which is helpful to develop or implement machine learning based algorithms in artificial intelligence.

This paper investigates and implement the machine learning algorithms, and show the most efficient algorithm based on time and accuracy with feature engineering techniques for classification which could be helpful to enhance the efficiency to detect the types of attacks in IoT networks.

The questions of which we will try to get the answers in this paper are:

Q1. Why do we need AI based algorithm to detect malware and traditional programming is not enough?

Q2. Which machine learning algorithm is the most efficient based on the results?

II. Literature Review

In this regard, there are several works here in which the authors purposed different methods and techniques to detect network anomalies using machine learning and deep learning. Some of authors compared the results on the bases of applied ML algorithms on different datasets. Some of related notable works are discussed here. Dutta et al. [1] they used ensemble method and idea to use with deep learning models and other classifiers. They also used IoT-23 and other datasets. The compared and test network anomaly detection techniques. Ngo et al. [2] they discussed that how to deal with multi architecture using static type methods and provide survey regarding that. They compared and analyzed existing malware detection methods in last some years. Wu et al. [3] discussed and summarized the IoT security issues and its feasibilities of technical type. They compared different algorithms and solutions. They also discussed that there are many abilities of devices to address the security issues. Zaman et al. [4] they discussed about the security threats in devices, and also discussed that how advanced encryption methods cannot implemented efficient on devices because of limits. Benjamin Vignau et al. [5] they discussed about the some history and evolution of malwares in internet of things, they also discussed and compared the features of different programs regarding internet of things malware. Stoian [6] he discussed the use of machine learning based algorithms in the anomalies detecting, applied different ML algorithms on dataset and compared the results regarding accuracy. Al-Zewairi et al. [7] they discussed detection studies of unknown attacks since last 10 years, they proposed categorization of attacks. They also conducted experiments for intrusion detection based on deep artificial neural network. Imtiaz Ullah and Qusay H. Mahmoud [8] They provide technique to detect if the device connected to network is IoT using machine learning, they result 100% in precision, recall and F score. Booij et al. [9] They used ToN_IoT dataset and compare other datasets, the discussed that these could be high impact on performance due to dataset differences. Cai et al. [10] They focused port scanning and P4 switch, provide E-Replacement method. They showed the improvement of detection compared to traditional methods. Tian et al. [11] They proposed DC Adam approach, with the other related experiments and comparisons, they showed the asynchronous federated learning by accuracy of 12.8%. Kalinin et al. [12] They provide traffic analysis method which is based on Needleman-Wunsch algorithm, basically it is a prototype of IDS. They also showed the experiment results of their proposed approach. Daniel et al. [13] As we are using IoT-23 dataset which includes zeek log files, they compare eight ML models performance on dataset. They showed that the result is almost has 90.3% accuracy. Haas et al. [14] Integrated zeek osquery proposed by them which is open source platform, it combines osquery and zeek host monitor. The reason behind this is to extend network IDS scope. Where IDS is Intrusion Detection System. Gustavsson [15] He used NIDS, implement machine learning algorithms using Scikit-learn and showed that efficient algorithms are KNN, Random Forest and Decision Tree on CICIDS2017 dataset. Sarker et al. [16] they discussed ML and DL approaches such as classification, regression, clustering and rule based techniques and give comprehensive overview to address the issue. They also shared the scope of study as well as they showed illustration of machine learning and deep learning methods. Abdullahi et al. [17] a systematic literature review study, they categorized and mapped literature on artificial intelligence methods which are exist and used for detecting attacks in IoT. They also shared the investigation regarding artificial intelligence methods. They also provide the studies as well as pros and cons. There are approximately eighty studies which were selected.

III. Methodology

A. *Review the Gist of Simple and Machine Learning Algorithms*

Basically we need an input and rules in simple or non AI based programming to work, it works on the set of rules to do a specific task. On the other hand, ML based algorithms need input with sample or historical data to work, a model is trained on the basis of given sample data and then algorithm predicts the output for given input. Therefore, it is generally not necessary to make set of rules for each prediction in ML based algorithms.

B. *Dataset*

It is the most important step to choose right and relevant dataset when you are working with machine learning algorithms. For this purpose, we selected Aposemat IoT-23 which is a labeled dataset and created in the Avast laboratory. The dataset has different internet of things devices captures of malware which are around 20 and also has benign anomalies capture which are around 3. The mentioned dataset has original network captured .pcap files and also has conn.log.labeled files, created using network analyzer zeek.

Zeek network analyzer: Zeek is a passive, open-source network traffic analyzer. Many operators use Zeek as a network security monitor (NSM) to support investigations of suspicious or malicious activity. Zeek also supports a wide range of traffic analysis tasks beyond the security domain, including performance measurement and troubleshooting.

conn.log files from zeek: The connection log, or conn.log, is one of the most important logs Zeek creates. It may seem like the idea of a “connection” is most closely associated with stateful protocols like Transmission Control Protocol (TCP), unlike stateless protocols like User Datagram Protocol (UDP). Zeek’s conn.log, however, tracks both sorts of protocols.

Inspecting the Dataset and conn.log.labeled files:

Following are the quantity of total captures provided in the Aposemat IoT-23 labeled dataset, which shows that approximately 90.5% captures are malicious.

TABLE I
Captures quantity in Aposemat IoT-23 dataset

Total captures around	Malicious captures around
325,307,990	294,449,255

Following are the types of detected attacks in dataset, by using the selected dataset with ML algorithm we can probably capture and predict the following types of attack on real IoT network.

TABLE II
Detected attacks in Aposemat IoT-23 dataset

Name	Description
Attack	A general label of anomalies
Command and Control (C&C)	This type can take device control and perform future attacks
Command and Control File Download	Receiving a file in device from the server
Command and Control Mirai	Mirai type of attack
Command and Control Torii	Torii type of attack which is more complex type of Mirai attack

DDoS	Distributed denial of service doing by device
Command and Control Heart Beat	This attack can be defined as the monitoring time to time from source
Command and Control Heart Beat Attack	Similar as attack no 8, but the attack method is not prominent
Command and Control Heart Beat File Download	Monitoring is done using small file rather than packet
Command and Control Part of a Horizontal Port Scan	Receiving packages to collect information for perform an attack in future
Okiru	Okiru type of attack which is also more complex type of Mirai attack
Okiru Attack	Type of attack has detected but method is not identified
Part of a Horizontal Port Scan	Collection of information is done to perform attack in future
Part of a Horizontal Port Scan Attack	Purpose of attack detected but methods are not identified

Since, it is important to mention here that the dataset is very large, therefore we decide to take some records from each sub-dataset of main dataset and then combine these records to generate a new dataset. Using this technique, we will be able to handle the workload on my own computer for the combined dataset. Also the new combined dataset contains most of the types of attack in aposemat IoT-23 dataset.

Data Preprocessing:

We need to load all 23 sub-datasets of the selected dataset, for the purpose of preprocessing we used Pandas library of Python programming language. First we loaded all sub-datasets separately into data frames (DF) and implement a condition to skip the starting ten (10) rows due to these rows can contains information regarding dataset, also implemented a condition to read one lac (100000) rows further. After generating all these 23 separate data frames we combined these data frames to a new single data frame.

After combining it is needed to sort out the labels in simple form, so we implemented a condition to sort out the following labels of attacks.

TABLE III
Sorted labels of attacks

Label in dataset	Sorted Label
- Malicious PartOfAHorizontalPortScan	PartOfAHorizontalPortScan
(empty) Malicious PartOfAHorizontalPortScan	PartOfAHorizontalPortScan
- Malicious Okiru	Okiru
(empty) Malicious Okiru	Okiru
- Benign -	Benign
(empty) Benign -	Benign
- Malicious DDoS	DDoS
- Malicious C&C	C&C
(empty) Malicious C&C	C&C
- Malicious Attack	Attack
(empty) Malicious Attack	Attack
- Malicious C&C-HeartBeat	C&C-HeartBeat
(empty) Malicious C&C-HeartBeat	C&C-HeartBeat
- Malicious C&C-FileDownload	C&C-FileDownload
- Malicious C&C-Torii	C&C-Torii
- Malicious C&C-HeartBeat-FileDownload	C&C-HeartBeat-FileDownload
- Malicious FileDownload	FileDownload
- Malicious C&C-Mirai	C&C-Mirai
- Malicious Okiru-Attack	Okiru-Attack

After that, there are some columns which can be removed and they not have effect on the results. So we decided to drop that columns from combined dataset file, these columns are: ts, uid, id_orig.h, id_orig.p, id_resp.h, id_resp.p, service, local_orig, local_resp, history.

After these steps, there are some missing values in the following columns which are needed to be replace with 0.

TABLE IV
Replaced values of columns

Column	Exist Values	Replaced Values
duration	-	0
orig_bytes	-	0
resp_bytes	-	0

As we know that machine learning algorithms required numerical values of features to work, in the above combined dataset we have two important columns or feature which are 'proto' and 'conn_state'. Is it important to use these features in the machine learning algorithms but the problem is that these columns have string values and need to be replaced with numerical data. For that purpose, we used one hot encoding technique to replace these columns values.

One-Hot Encoding:

Basically it is a process to convert the categorical data into the form that can be provide to machine learning algorithms for better accuracy. In this process the values in column transpose and for each string value a separate column is create in which binary values 0 or 1 insert according to the existence of that string value. For each string type column, the process repeats if you want to use and convert that column with one-hot encoding.

There are some other encoding methods as well but for the selected dataset it is better to use one-hot encoding method.

Following is the example of implementation of one-hot encoding on the dataset columns: proto and conn_state

TABLE V
One-Hot encoded columns

Column	Binary Values
proto_icmp	0 or 1
proto_tcp	0 or 1
proto_udp	0 or 1
conn_state_OTH	0 or 1
conn_state_REJ	0 or 1
conn_state_RSTO	0 or 1
conn_state_RSTOS0	0 or 1
conn_state_RSTR	0 or 1
conn_state_RSTRH	0 or 1
conn_state_S0	0 or 1
conn_state_S1	0 or 1
conn_state_S2	0 or 1
conn_state_S3	0 or 1
conn_state_SF	0 or 1
conn_state_SH	0 or 1
conn_state_SHR	0 or 1

Following are the list of attacks in created combined dataset:

TABLE VI
Count of attacks in combined dataset

Name of attack	Count of attack
PartOfAHorizontalPortScan	825939
Okiru	262690
Benign	197809
DDoS	138777
C&C	15100
Attack	3915
C&C-HeartBeat	349
C&C-FileDownload	43
C&C-Torii	30
FileDownload	13
C&C-HeartBeat-FileDownload	8
C&C-Mirai	1

A summary of combined dataset file is:

TABLE VII
Summary of combined dataset

Total sub-datasets of captures in IoT-23 dataset	23 (20 Malicious + 3 Benign)
Total sub-datasets of malicious captures combined	20
Total rows in combined dataset (as per taken rows)	1444674
Total types of attacks in combined dataset	18
Total count of attacks in combined dataset	1444674
Total grouped types of attacks in combined dataset	12
Total grouped count of attacks in combined dataset	1444674

C. Implementation of ML algorithms to combined dataset

For the implementation of machine learning algorithms, preprocessed combined dataset splited for the training and testing, 0.8 size for the training and 0.2 for the testing.

The implemented ML algorithms in this paper are: Naive Bayes, Decision Tree and Random Forest.

Environment for the implementation

I used a laptop PC to implement the above explained algorithms. The configurations of PC are Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz 2.50 GHz, RAM is 8 GB, 64-bit OS, operating system is windows 10, IDE, python 3.8 and other libraries.

Matrices from algorithms

Following are the matrices from algorithms which are used to evaluate outputs:

TABLE VIII
Evaluation metrics from algorithms

Name	Explanation and Formula
Time	Total time of processing for an algorithm to give output
True Positive	Predicts correctly the positive class by algorithm
False Positive	Predicts not correctly the positive class by algorithm
Precision	Calculation of positives which are correctly identified by algorithm $precision = \frac{true\ positives}{true\ positives + false\ positives}$
Recall	Calculation of actual positives which are correctly identified by algorithm $recall = \frac{true\ positives}{true\ positives + false\ negatives}$
F1 score	Calculation of harmonic mean for recall & precision $F1 = 2 * \frac{precision * recall}{precision + recall}$
Support score	Predicts not correctly the positive class by algorithm

IV. Results and Discussion

After implement the above explained machine learning algorithms, got the following Results:

TABLE IX
Results of Naive Bayes

	Precision	**Recall**	**F1 Score**	**Support**
Attack	0.47	0.09	0.15	768
Benign	1	0.29	0.45	39562
C&C	0.62	0.1	0.18	3045
C&C-FileDownload	0.01	1	0.02	8
C&C-HeartBeat	0.06	0.51	0.11	76
C&C-HeartBeat-FileDownload	0.5	1	0.67	2
C&C-Torii	0	0	0	6
DDoS	1	0.82	0.9	27842
FileDownload	0.25	0.33	0.29	3
Okiru	0.21	1	0.35	52867
PartOfAHorizontal-PortScan	1	0	0	164756
Accuracy			0.30	288935
macro avg	0.46	0.47	0.28	288935
weighted avg	0.85	0.3	0.21	288935

TABLE X
Results of Decision Tree

	Precision	**Recall**	**F1 Score**	**Support**
Attack	1	0.99	1	747
Benign	0.96	0.55	0.7	39239
C&C	0.6	0.13	0.22	3041
C&C-FileDownload	0.82	0.82	0.82	11
C&C-HeartBeat	0.81	0.37	0.5	79
C&C-HeartBeat-FileDownload	0.5	1	0.67	1
C&C-Torii	0	0	0	3
DDoS	1	0.82	0.9	27661
FileDownload	0.67	0.5	0.57	4
Okiru	0.67	0	0	52342
PartOfAHorizontal-PortScan	0.68	1	0.81	165807
Accuracy			0.73	288935
macro avg	0.7	0.56	0.56	288935
weighted avg	0.75	0.73	0.65	288935

TABLE XI
Results of Random Forest

	Precision	**Recall**	**F1 Score**	**Support**
Attack	0.99	1	1	747
Benign	0.96	0.55	0.69	39239
C&C	0.52	0.14	0.22	3041
C&C-FileDownload	1	0.73	0.84	11
C&C-HeartBeat	0.75	0.38	0.5	79
C&C-HeartBeat-FileDownload	0.5	1	0.67	1
C&C-Torii	0	0	0	3
DDoS	1	0.82	0.9	27661
FileDownload	0.5	0.5	0.5	4
Okiru	0.68	0	0	52342
PartOfAHorizontal-PortScan	0.68	1	0.81	165807
Accuracy			0.73	288935
macro avg	0.69	0.56	0.56	288935
weighted avg	0.75	0.73	0.65	288935

Following are the summary of results:

TABLE XII
Summary of results

Algorithm	**Accuracy of test**	**Time Cost**
Naive Bayes	0.30	6.74 seconds
Decision Tree	0.73	7.44 seconds
Random Forest	0.73	5.69 seconds

As we can see the summary shows that in our experiment Naive Bayes has the slowest processing for our selected dataset and the Random Forest shows the better time cost which is 5.69 seconds as well as better accuracy which is 0.73. Basically Decision Trees are common supervised learning ML algorithms can have problems such as bias and overfitting, random forest builds multiple decision trees and merges them, when multiple decision trees form an ensemble in the random forest algorithm, they predict accurate results.

V. Conclusion

In this paper, we experiment for the network attack detection in IoT using artificial intelligence with supervised learning based machine learning algorithms. After investigation and implementation, results in this paper shows that the Random Forest has better accuracy with better time cost as compare to other implemented algorithms on the selected dataset. In future, Iot-23 and other relevant datasets with different pc environment could be tested with these as well as other algorithms including deep learning. By doing this we can compare the deep learning algorithms and better clarify the overall efficiency.

REFERENCES

[1] Vibekananda Dutta , Michał Chora´s, Marek Pawlicki and Rafał Kozik, "A Deep Learning Ensemble for Network Anomaly and Cyber-Attack Detection", Sensors, August 2020.

[2] Quoc-Dung Ngo, Huy-Trung Nguyen, Van-Hoang Le, Doan-Hieu Nguyen, "A survey of IoT malware and detection methods based on static features", ICT Express, December 2020.

[3] Hui Wu, Haiting Han, Xiao Wang, and Shengli Sun, "Research on Artificial Intelligence Enhancing Internet of Things Security: A Survey", IEEE Access, August 2020.

[4] Shakila Zaman, Khaled Alhazmi, Mohammed A. Aseeri, Muhammad Raisuddin Ahmed, Risala Tasin Khan, M. Shamim Kaiser and Mufti Mahmud, "Security Threats and Artificial Intelligence Based Countermeasures for Internet of Things Networks: A Comprehensive Survey", IEEE Access, June 2021.

[5] Benjamin Vignau, Raphaël Khoury, Sylvain Hallé, Abdelwahab Hamou-Lhadj, "The evolution of IoT Malwares, from 2008 to 2019: Survey, taxonomy, process simulator and perspectives", J. Syst. Archit., June 2021.

[6] Nicolas-Alin Stoian, "Machine Learning for Anomaly Detection in IoT networks: Malware analysis on the IoT-23 Data set", University of Twente, 2020.

[7] Al-Zewairi, Malek, Sufyan Almajali, and Moussa Ayyash. "Unknown Security Attack Detection Using Shallow and Deep ANN Classifiers." Electronics 9.12, November 2020

[8] Ullah, Imtiaz, and Qusay H. Mahmoud. "Network Traffic Flow Based Machine Learning Technique for IoT Device Identification." IEEE International Systems Conference (SysCon), 2021

[9] Booij, Tim M., et al. "ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Datasets." IEEE Internet of Things Journal, May 2021.

[10] Cai, Yun - Zhan, et al. "E - Replacement: Efficient scanner data collection method in P4 - based software - defined networks." International Journal of Network Management, November 2021

[11] Tian, Pu, et al. "Towards Asynchronous Federated Learning Based Threat Detection: a DC-Adam Approach." Computers & Security, Sep 2021

[12] Kalinin, Maxim O., V. M. Krundyshev, and B. G. Sinyapkin. "Development of the Intrusion Detection System for the Internet of Things Based on a Sequence Alignment Algorithm." Automatic Control and Computer Sciences 54.8, Dec 2020

[13] Daniel K. Andrews; Rajeev K. Agrawal; Suzanne J. Matthews; Alexander S. Mentis. "Comparing Machine Learning Techniques for Zeek Log Analysis", IEEE Xplore, January 2022

[14] Steffen Haas, Robin Sommer & Mathias Fischer. "Zeek-Osquery: Host-Network Correlation for Advanced Monitoring and Intrusion Detection", IFIPAICT, Sep 2020

[15] Gustavsson, Vilhelm. “Machine Learning for a Network-based Intrusion Detection System : An application using Zeek and the CICIDS2017 dataset.”, 2019.

[16] Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. "Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions", Mobile Networks and Applications, 1-17, 2022

[17] Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir, S. J. "Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review", Electronics, 11(2), 198, 2022