Issues in Information Systems
Volume 23, Issue 1, pp. 316-332, 2022

DOI: https://doi.org/10.48009/1_iis_2022_126

Malware detection framework in cyber-physical systems using artificial intelligence - machine learning

Temechu Girma Zewdie, *University of the District of Columbia, temechu.zewdie@udc.edu*
Dr. Anteneh Girma, *University of the District of Columbia, anteneh.girma@udc.edu*
Dr. Paul Cotae, *University of the District of Columbia, pcotae@udc.edu*

Abstract

Nowadays, in Artificial intelligence and machine learning research, better prediction modeling with better accuracy is a critical appetite for the domain. These prediction models struggle to capture the relationship between the concentration of given dataset features and their respective target resources. This research will examine various approaches and propose a framework that can use alternative machine learning algorithms to successfully differentiate between malware files and clean files in cyber-physical systems (CPS) while minimizing the number of false positives. Hence, the candidate machine learning algorithms we examined are Random Forest, Decision tree, K-nearest, Ada boost, SGD, Extra Tree, and Gaussian NB Classifier. After successfully testing with medium-size malware and clean file datasets, the proposed framework will submit to a scaling-up process that enables us to work with exceptionally large malware and clean file datasets (Alhowaide, Alsmadi, & Tang, 2019). Based on the captioned candidate algorithm, our experiments depict that Random Forest, Decision Tree, Ada Boost, and Extra Tree Classifier achieved 100% accuracy in detecting attacks with Zero False-positive and False-negative rates. Thus, in this research, we proposed captioned candidate Malware detection framework in cyber-physical systems using Artificial Intelligence - Machine Learning through the experiments. Finally, the proposed framework is based on a limited amount of data. But if such an experiment were performed with a high-volume dataset and/or different attack types, the result may differ. Therefore, such a limitation highlights further research will be required.

Keywords: Cyber-physical system (CPS), Malware Detection, Machine Learning, IoT, Security

Introduction

Cyber-physical systems (CPSs) are an umbrella term that includes systems of many sorts that, including robotics, machine automation, industrial control systems (ICSs), process control systems, supervisory control, and data acquisition (SCADA) systems, the Industrial Internet of things (Eliot), and the Internet of Things (IoT) (Houbing Song, Security and Privacy in Cyber-Physical Systems, 2018). Cyber-physical systems (CPS) are used on a large scale in the modern industrial system (Sharmeen, Huda, & Abawajy, 2019). Nowadays, the use of CPS, which is the interconnection of multiple devices, and connections of devices to humans, are showing rapid growth (Zewdie & Girma, 2020). Such an increase in interest, usage and dependency on CPS devices increases security threats exponentially. Moreover, the nature of the Cyber-Physical device and its' associated challenges and the threats and remediation provided are not going through the same speed for security fixes or updates are also additional threat doors in the CPS industry. Let us further describe the state-of-the-art cyber-physical systems in depth.

Cyber-physical system (CPS) Architecture

A cyber-physical system (CPS) combines computer-aided software components with mechanical and electronic parts via a data center where the Internet communicates (GHOSH, 2018). Nowadays the cyber-physical system (CPS) industry is growing from time to time to benefit from the advantages it has. For instance, CPS provides network integration such as media access control techniques and their effort on the system dynamics, middleware, and software that provides coordination over a network (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016).On the other hand, Situational awareness and human perception of the system are critical for decision-making. Some CPSs include humans as an integral part of the system, making the interaction easier because humans are usually difficult to model using standalone systems (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016).

As described (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016), Certainty is the process of providing proof that a design is valid and trustworthy. CPS is designed can evolve and operate in new and unreliable environments. CPS can demonstrate unknown system behavior to study further and evolve into a better system. Moreover, With the close interaction of sensors and cyberinfrastructure, CPS can provide better system performance in terms of feedback and automatic redesign. CPS can scale the system (Scalability) according to demand by utilizing the properties of Cloud Computing. Users can acquire the necessary infrastructure without investing additional resources. Due to having sensor-cloud integration, CPS can provide autonomy. CPS is a closed-loop system where sensors make measurements of physical dynamics (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016). The current CPS provides a more flexible service compared to the earlier research efforts in WSN and Cloud Computing alone, and Current sensors and cloud infrastructure offer large optimizations for various apps. This drives a great pathway for CPS to optimize the system to a wide extent (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016)

Finally, CPS can provide a faster response time due to sensors and cloud infrastructure's faster processing and communication capability. This Fast response time can facilitate the early detection of remote failure and proper utilization of shared resources such as bandwidth (Barbosa, Leitão, Trentesaux, Colombo, & Karnouskos, 2016). The efficiency of CPS matters with the design architecture of CPS.

CPS architecture

According to (Liang Hu, 2012), there are three tiers of CPS architecture. These are Environmental Tiers, Service Tiers, and Control Tiers. The following Picture demonstrates the three tiers of CPS architecture.

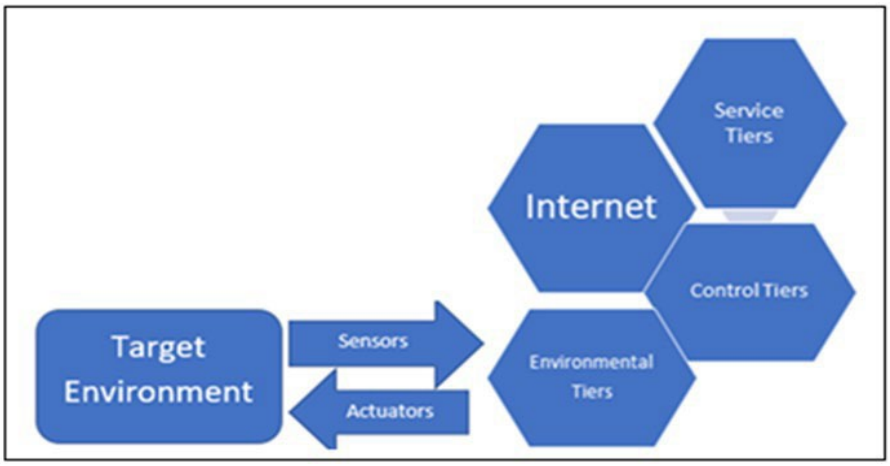

Figure 1 The three tiers of CPS architecture

Environmental Tiers: consists of physical devices and a target environment which includes end-users using the devices and their associated physical environment.

Service Tiers: a typical computing environment with services in SOA and Cloud Computing.

Control Tiers: to receive monitored data which are gathered through sensors, to make controlling decisions, to find the right services by consulting service framework, and to let the services invoked on the physical device Such an architecture is not an absolute once for most CPS applications.
There are different technologies categorized under CPS. Internet of Things(IoT) based CPS is one of the significant ones. The following will discuss IoT as CPS.

Literature Review

This section describes details of the current state-of-the-art in different topics covered in the present work. Then, it reviews the usage of AI Machine Learning for Malware detection in cyber-physical systems precisely with IoT devices, with special consideration for IoT. Finally, it reviews the available datasets modeling cyberattacks on CPS, especially with IoT devices.

IoT as Cyber-Physical System

Internet of Things (IoT) as Cyber-Physical Systems is an emerging technology that enables the interaction of uniquely identifiable computing devices that can be embedded with other interfaces. For instance, machines and humans are linked via wired and wireless networks. It has been exposed to capture contextual data and create an information network to provide new functionalities and digital business models. (Chaudhuri, 2019). It will lead to efficient mechanisms with high scalability and interoperability features among the things or objects. IoT is a reality that is progressing day by day, connecting billions of people and things to form a vast global network. IoT has applications in various domains like agriculture, industry, military, and personal spaces. Potential research challenges and issues in IoT act as a hurdle in the complete exploration of IoT in real-time implementation. (Jain, Choudhari, & Srivastava, 2021). The Following Figure 2 describes a common IoT system's high-level architecture divided into three layers: perception, network, and applications. How components are grouped in the three layers of a generic IoT system.

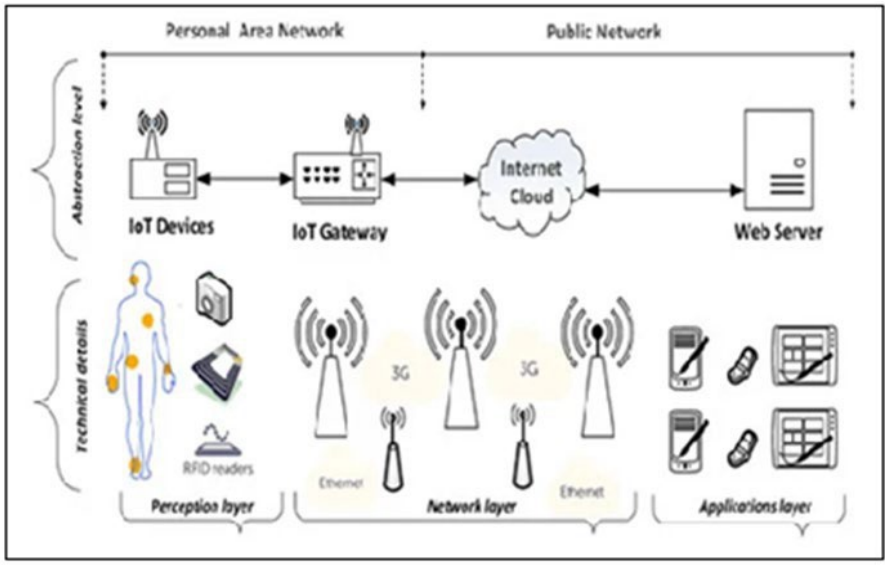

Figure 2 A High-level architecture of an IoT System (Ali & Awad, 2018)

Security Issue in Cyber-Physical Systems

Security Issues of cyber-physical systems that includes identify the possible vulnerabilities, attack issues, adversaries' characteristics and a set of challenges that need to be addressed. Security and privacy in CPSs more complex than materialized today. One of the complexities of CPSs is when they are invisibly connected to the Internet. The extent of CPS devices' security and privacy boundaries may suddenly become global in scope (Houbing Song, Security and Privacy in Cyber-Physical Systems).

Security in IT system is all about CIA (Confidentiality, Integrity, and Availability). But IT systems and CPS systems differ in system and security requirements and challenges. CPS security paradigm is the same as the IT System (CIA). But Paradigm is in an inverted order, AIC, namely called Availability, integrity, and Confidentiality of a CPS system. Let us discuss these from an AIC perspective.

Availability provides timely and reliable access to and use of information to authorized persons. A few examples of threats that affect Availability include Denial of Service (DoS) or Distributed Denial of Service (DDoS).

Integrity ensures information non-repudiation and authenticity, guards against improper information modification or destruction. Salami attacks, Data diddling attacks, Session hijacking, and Man-in-the-middle (MITM) attacks are typical examples that CPS Hackers used.

Confidentiality is ensuring Confidentiality means keeping secret information protected and away from unauthorized disclosure. By ensuring Passwords and data encryption with standard techniques, we can keep the Confidentiality of the CPS System. Moreover, solutions grounded in cryptography, such as those that use *Transport Layer Security (TLS)*, Hash-based message authentication (HMACs), or other authentication used to ensure Confidentiality but due to limitations in CPS devices and relative computational cost of such protocols. The following section will discuss the critical challenges of the CPS system.

Challenges of Cyber-Physical Systems

Incorrect access control, overly large attack surface, Outdated software, Lack of encryption, Application vulnerabilities, Lack of Trusted Execution Environment, Vendor security posture, and Insufficient privacy protection are the most common Security challenge with CPS Devices (Langkemper, n.d.). Restricting access to devices, connecting resources in a cloud environment in a secured way, and auditing device usage are the necessary protection mechanism that should be in place to keep data confidential and private (Zewdie & Girma, 2020). Security experts state that most CPS devices lack safeguards and become easy targets for attackers. Even though it is not the case with all objects connected in the IoT network, identifier-specific codes such as identification codes for particular devices, like the IMEI number for mobile phones, are another security challenge. Access to vulnerable devices could provide easy access for cybercriminals, and they could quickly gain access to other connected systems in the network. IoT Security needs artificial intelligence to play a significant role as a security tool to mitigate these challenges.

Cyber threats of CPS

A cyber threat is any action that could result in an unwanted impact on IT infrastructures. Such a threat is a malicious act that strives to disrupt or steal an information technology asset, or to gain unauthorized access, damage, a computer network, intellectual property, or any other form of sensitive data Malware, Web-based Attacks. Phishing. Web Application Attacks, SPAM, Distributed Denial of Service (DDoS), Identity Theft, Data Breach, Insider Threat, Botnets, Physical Manipulation, Damage, Theft and Loss,

Information Leakage, Ransomware, Cyber Espionage, Crypt jacking, and other attack vectors are common types of Cyber threats (OWASP, 2022) (ENISA, 2022). Among listed threats, this research will focus on the detection of malware with AI/ML that can affect CPS precisely in IoT devices based on the IoT dataset for Intrusion Detection Systems (IDS).

Malware is malicious software that is designed to harm computers and computer users by stealing information, corrupting files, or doing mischievous activities to annoy users. Currently, Malware is widely spreading, and increase in security incidents on computers. There are different types of Malwares, and they are categorized in the following Class. I.e., viruses, worms, Trojan horses, Rootkit, Spyware, Adware, Cookies, Sniffers, Botnet, Keyloggers, Spam, and Ransomware (Tahir, 2018).

According to a Statista report, currently, the Cumulative detections of newly developed malware applications worldwide from 2015 to March 2020 are strictly increased. Figure 2 shows how the malware attack grows from time to time and affects the cyber information.

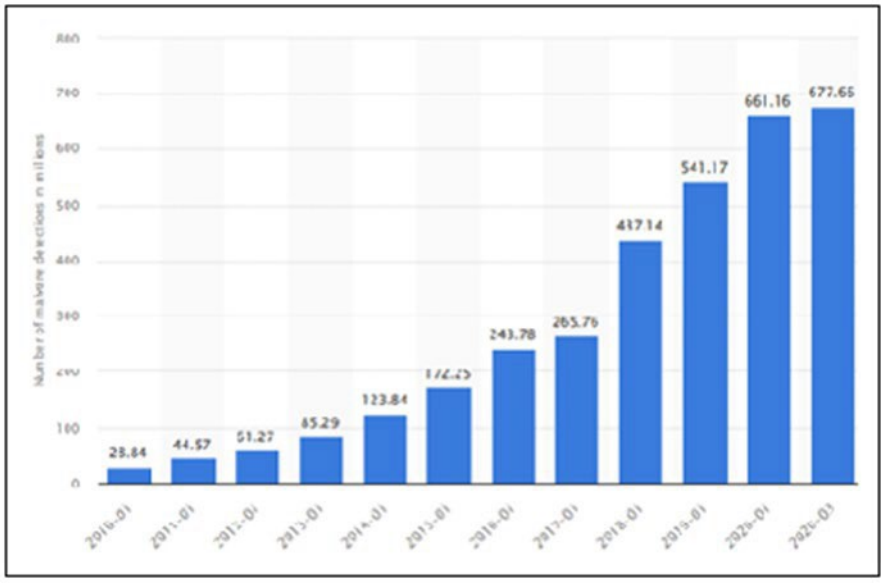

Figure 3 Cumulative detections of newly developed malware applications (in millions)

Therefore, Malware creates a severe problem for industries and users. As mentioned in the above graph, in 2020, 677,660,000 million malwares have been detected. As a result, it impacted a significant economic loss on industries and affects an individual. That is why, advanced malware protection i.e., AI ML solution is required.

Methodology

Dataset used for this study

The Malware dataset, the most recent dataset, contains nine IoT devices traffic sniffed using Wireshark in a local network using a central switch. It includes two Botnet attacks (Mirai and Gafgyt). The dataset contains twenty-three statistically engineered features extracted from the .pcap files. Seven statistical measures were computed (mean, variance, count, magnitude, radius, covariance, correlation coefficient) over the time window of 10 sec with decay factor equals 0.1 (Kaggle). For our research, we used the IoT dataset for Intrusion Detection Systems (IDS) from Kaggle.

Data Description

BoTNeTIoT-L01 is a data set integrated with all the IoT device's data files from the detection of IoT botnet attacks (BoTNeTIoT) data set. This latest version reduced the redundancy of the original dataset by choosing the features of 10 seconds time window only. In the dataset class label, 0 stands for attacks, and 1 stand for normal samples.

Preprocessing Steps

BoTNeTIoT dataset are contains nine IoT devices traffic sniffed using Wireshark in a local network before applying the preprocessing steps. The following data preprocessing steps have been done on the captioned dataset:

1. Add Feature Names: The feature names are added to each dataset column.
2. Dropping of Empty columns: Six empty column features: Unnecessary columns are deleted from the dataset.
3. Replace Empty Ports with 0: Two NA values were filled with 0.
4. Dropping non-required features: Features not required for the ML model are removed from the dataset.
5. Encoding object to Categorical Value: Three features, "HpHp_L0.1_pcc", "Attack," and "label," are factorized to encode the object as a categorical variable.

The preprocessing steps implemented on the BoTNeTIoT dataset reduced the number of features to twenty-five. Such preprocessed data is then used to train the ML model.

Feature Selection and extraction

Feature selection is also called variable selection or attribute selection. It is the automatic selection of attributes in a dataset such as columns in a tabular data that are most relevant to the predictive modeling problem you are working on (Brownlee, Data Preparation for Machine Learning, 2020). Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them.

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features (Ippolito, 2019).

Data Classification and prediction process

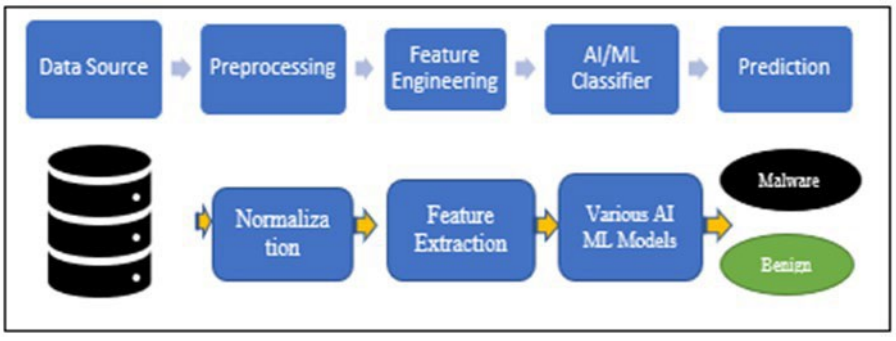

Figure 4 Data Classification and prediction process

Proposed Malware Detection using AI Machine Learning Algorithm

This research used and profoundly analyzes the following models to get maximum accuracy results.

Random Forest Classifier

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. This classifier is more accurate than the decision tree algorithm. Moreover, it provides an effective way of handling missing data and can produce a reasonable prediction without hyper-parameter tuning. RF solves the issue of overfitting in decision trees. Finally, in every random forest tree, a subset of features is selected randomly at the node's splitting point (Mbaabu, 2020).

Decision Tree Classifier

The main advantage of the decision tree classifier is its ability to use different feature subsets and decision rules at different stages of classification (Szczerbicki, 2008).

K-Neighbors Classifier

We used KNN in order benefits from other algorithms is that KNN can be used for multiclass classification. Therefore, if the data consists of more than two labels or in simple words if you are required to classify the data in more than two categories then KNN can be a suitable algorithm (Kulkarni, 2020).

AdaBoost Classifier

AdaBoost is best used to boost the performance of decision trees on binary classification problems. AdaBoost can be used to boost the performance of any machine learning algorithm. These are models that achieve accuracy just above random chance on a classification problem.
The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps. Each instance in the training dataset is weighted. The initial weight is set to:

$$\text{Weight}(xi) = 1/n$$

Where x_i is the i^{th} training instance and n is the number of training instances (Brownlee, Boosting and AdaBoost for Machine Learning, 2020).

Stochastic Gradient Descent (SGD) Classifier

In this research we use stochastic Gradient descent for the following two reasons. Th first one is, as we can read from the previous text, SGD allows minibatch (online/out-of-core) learning. Therefore, it makes sense to use SGD for large scale problems where it's very efficient, and the second one is, SVM or logistic regression will not work if you cannot keep the record in RAM. However, SGD Classifier continues to work (Fuchs, 2019).

Extra Trees Classifier

In this research we use an extra-trees classifier. This Classifier implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (scikit-learn developers , 2022).

Gaussian Naïve Bayes

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It is specifically used when the features have continuous values. It is also assumed that all the features are following a Gaussian distribution i.e., normal distribution. Bayes' theorem is based on conditional probability. The conditional probability helps us calculate the probability that something will happen, given that something else has already happened.

$$P(A/B) = P(B/A) * P(A) / P(B)$$ (Zewdie & Girma, 2022)

Performance Evaluation Model and Scoring

We have implemented the algorithm in python and evaluated the performance of our proposed model. Known malware and benign ware are served as labelled data in the algorithm. We have measured the accuracy, recall, and f1_score as the performance metrics.

Accuracy is the percentage of correctly identified apps. Accuracy can be defined using the following equation, in other word, accuracy is the number of correct prediction (TP and TN) divided by the number of all samples.

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$ (Zewdie & Girma, 2022)

Precision is used as a performance metric when the goal is to limit the number of false positives.

$$Precision = TP/(TP + FP)$$ (Zewdie & Girma, 2022)

Recall is the number of correct results divided by the number of results that should have been returned.

$$Recall = TP/(TP + FN)$$ (Zewdie & Girma, 2022)

Accuracy alone cannot define the model. ROC_AUC Curve defines capability of distinguishing between classes. Higher ROC_AUC curve score is more desirable. F-measure or f1_score illustrate the balance between recall and precision and can be calculated using the following equation.

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall)$$ (Zewdie & Girma, 2022)

Our proposed model has been achieved a higher accuracy level and f1_score than supervised AI/ML learning method. Semi-supervised model with deep learning has shown better performance than the AI Machine learning model.

Results and Discussion

In this research, we are the necessary Importing Libraries and load the right dataset, the IoT dataset for Intrusion Detection Systems (IDS), to preprocess data such as checking whether a value is NaN (Not a Number) or not. After cleaning data processing and getting structured data, we used to Learn an algorithm to identify the right candidate model. The following high-level process architecture shows the high-level process to identify the right candidate framework.

Figure 5 High Level prosses Architecture

As mentioned above, we propose a versatile framework in which one can employ different machine learning algorithms to successfully distinguish between malware files and benign files while aiming to minimize the number of false positives. After successfully testing on medium-size datasets of malware and benign files, the ideas behind this framework were submitted to a scaling-up process that enables us to work with large datasets of malware and benign files. Mirai and Gafgyt have been the go-to IoT malware for many years now in cybercrime circles: their versions have successfully infected millions of vulnerable IoT devices. (Zsigovits, 2021). The following picture shows the classification of malware and their respective count on both malware and benign files

The following figure shows the classification of the data after preprocessing. Thus, 78.8% of the files are malware (0) attack, and the remaining 21.2% of the entire dataset are benign (1).

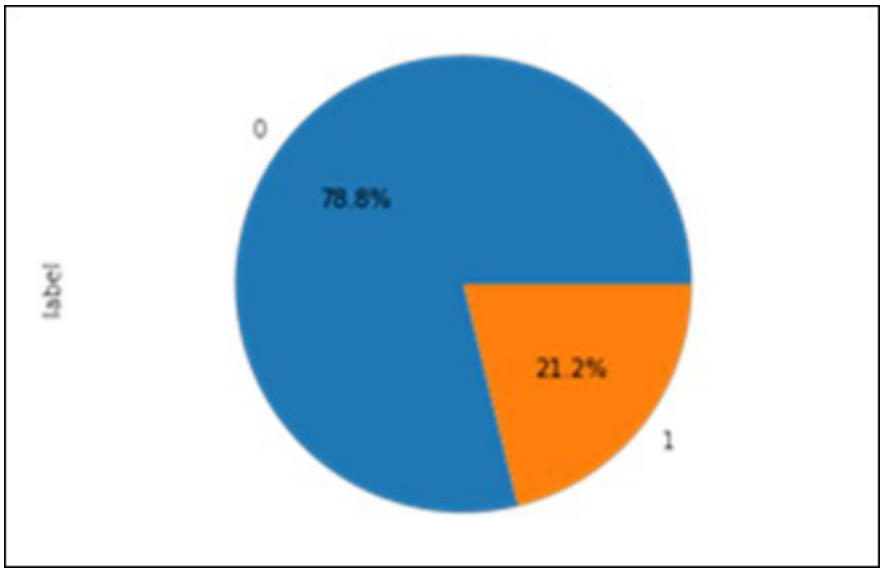

Figure 5 Classification of malware and benign files

Model Accuracy result

Our experiment, with a given dataset, which is 2,426 573 columns with 25 columns, we got the following accuracy result. Thus, Random Forest, Decision tree, Ada boost and Extra Tree classifier has 100% accuracy. The remaining classifier such as K- Neighbors Classifier has 99.64% , SGD Classifier - 91.24%, and Gaussian NB Classifier has 78.87% accuracy.

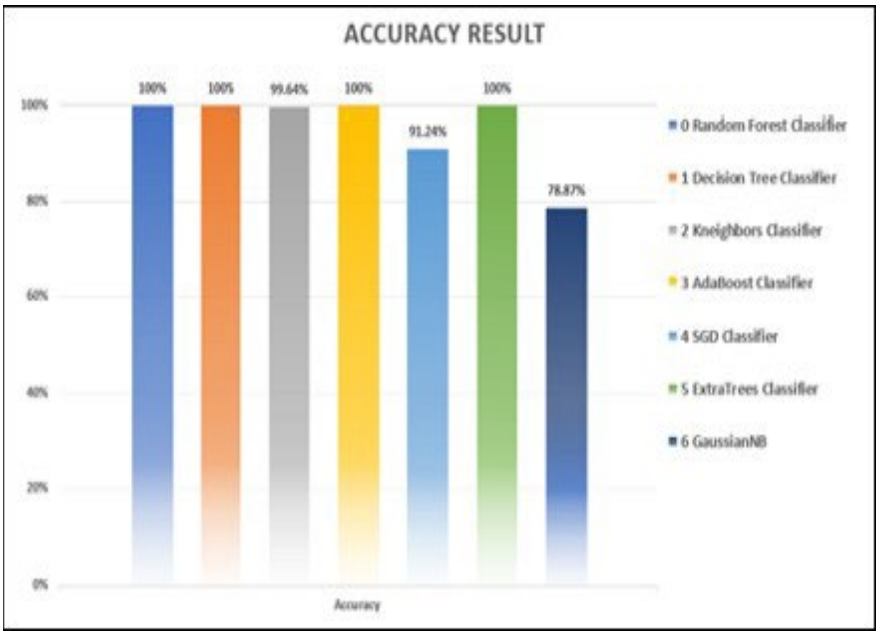

Figure 6 Accuracy Result

Evaluation Results

Random Forest

A random forest classifier was proposed by Breiman and Schapire (Breiman, L., & Schapire, E. (2001), and it consists of many individual classification trees, where each tree is a classifier by itself that is given a certain weight for its classification output. A Random Forest Classifiers were successful in many classification problems. In this work, the random forest classifier was the one that produced the best performance.
The reported averages include macro average (averaging the unweighted mean per label), weighted average (averaging the support-weighted mean per label), and sample average (only for multilabel classification). Micro average (averaging the total true positives, false negatives, and false positives) is only shown for multilabel or multi-class with a subset of classes because it corresponds to accuracy otherwise and would be the same for all metrics. Figure 8 shows that the RF Classifier model misidentified only Zero fraudulent cases.

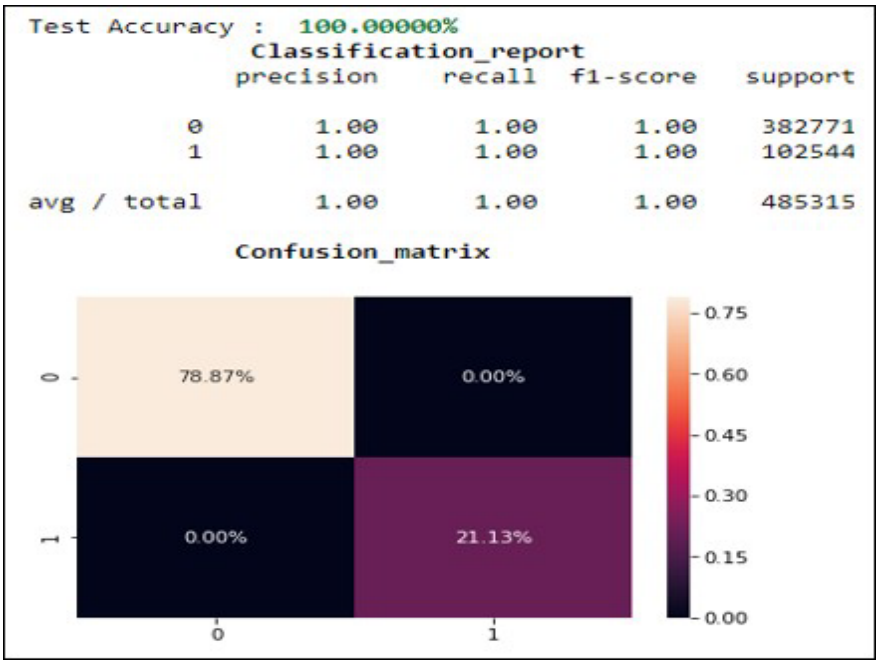

Figure 7 Random Forest Classifier Classification Report

Decision Tree Classifier

Figure 9 shows the Decision Tree Classifier model. In this model, there are no misidentified cases. This model seems perfect since there are no fraudulent cases in the examined mode, like a random forest.

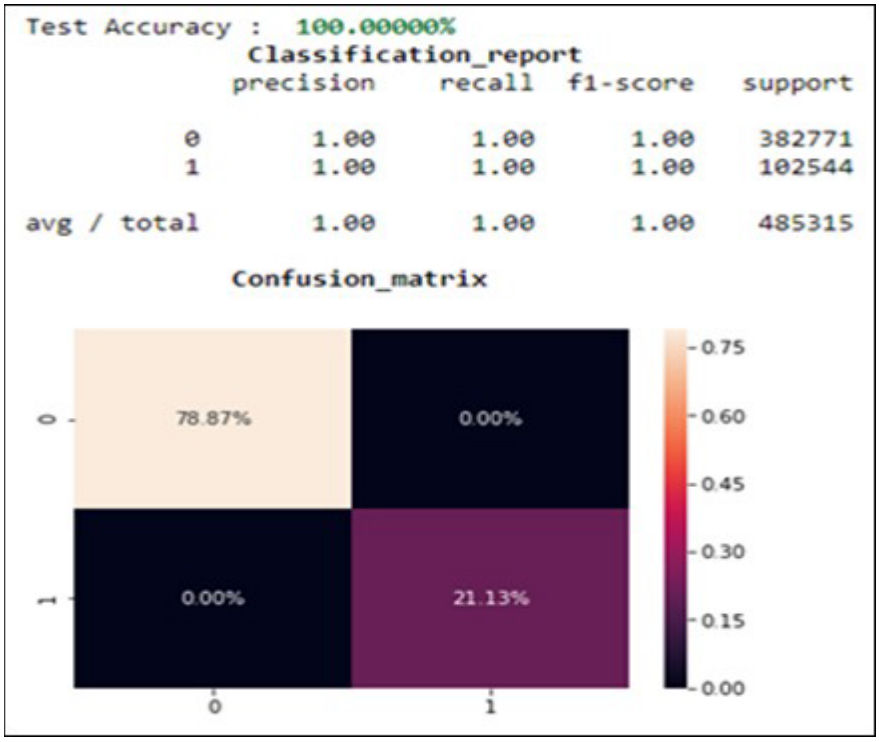

Figure 9 Decision Tree Classifier Classification Report

K - Neighbors Classifier

Figure 10 shows the K - Neighbors Classifier model. In this model, there are 8735 misidentified cases. 0.08% (698 cases) are False Negatives, and 0.28% (8036 cases) are false positives. This model needs further improvement.

Figure 10 K - Neighbors Classifier Classification report

AdaBoost Classifier

Figure 11 shows the Ada Boost Classifier model. In this model, there are no misidentified cases. The following model is perfect since there are no fraudulent cases in the examined mode.

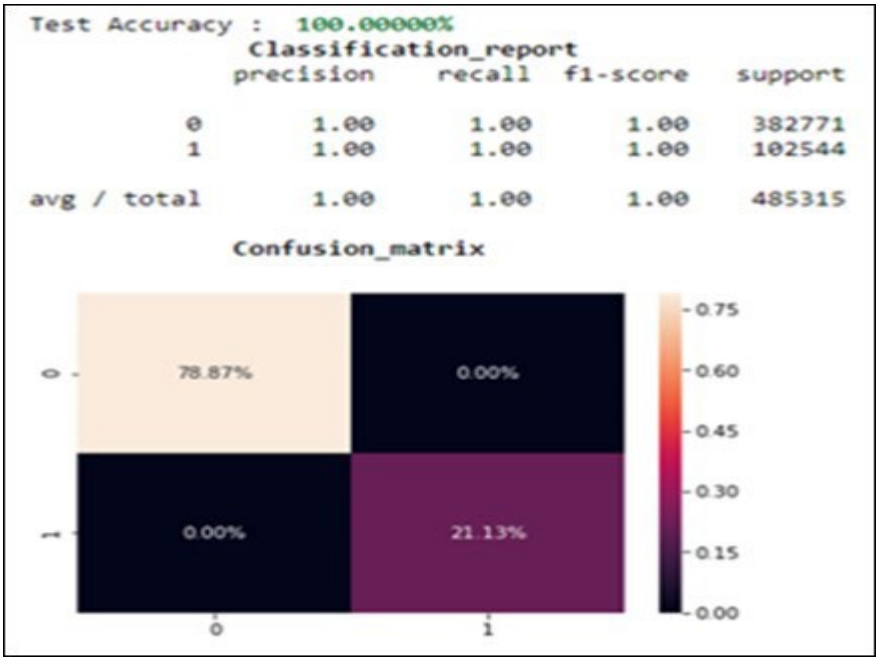

Figure 11 AdaBoost Classifier Classification report

SGD Classifier

Figure 12 shows the test accuracy of the SGD Classifier model looks 91.24424%, which is low when we compared to the other mentioned models above. In this model, there are 212,568 misidentified cases which are significant. Hence, this model needs further improvement.

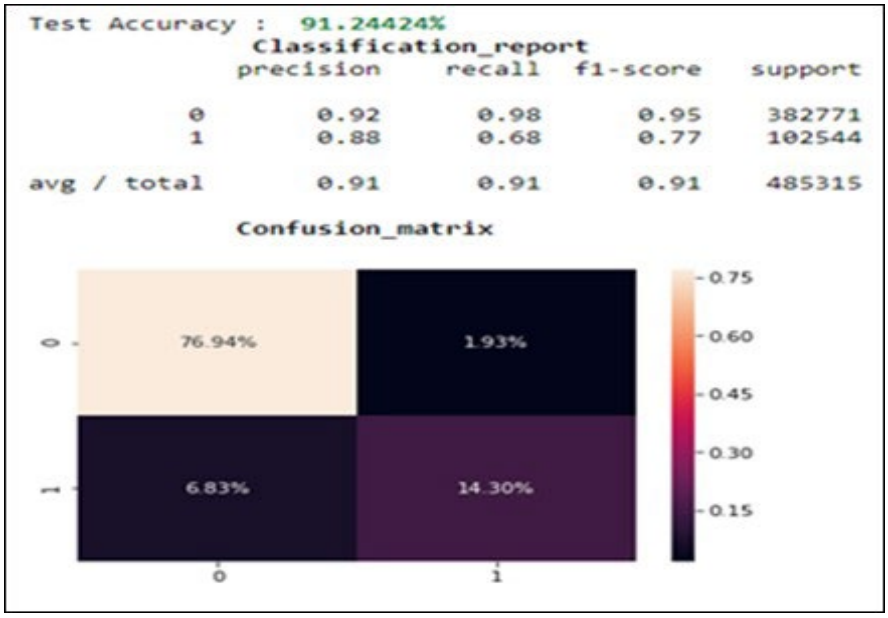

Figure 8 SGD Classifier Classification report

Extra Trees Classifier

Figure 13 shows the Extra Tree Classifier model. In this model, there are no misidentified cases. The captioned model is perfect since there are no fraudulent cases in the examined mode.

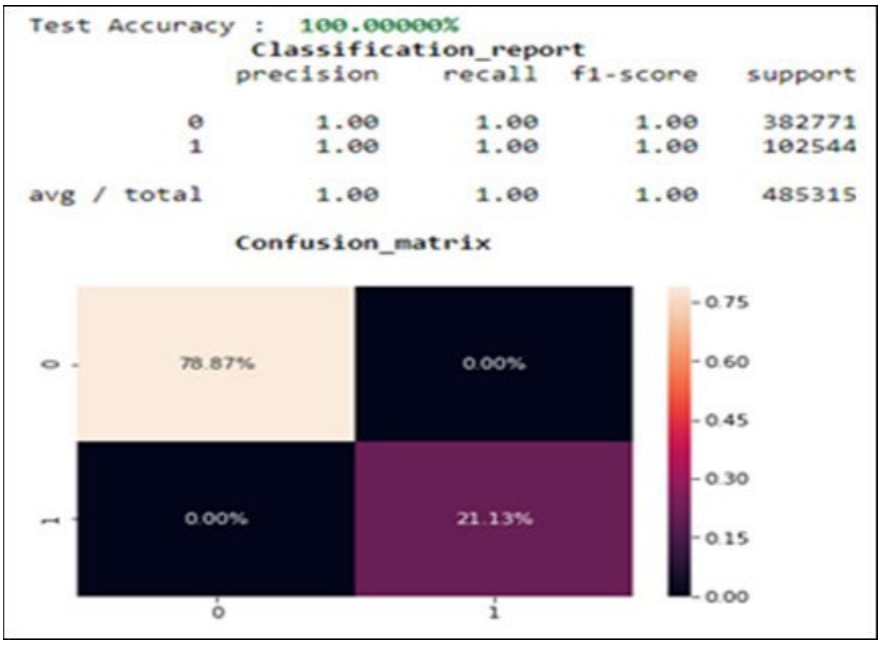

Figure 13 Extra Trees Classifier Classification report

Gaussian NB Classifier

Figure 14 shows the test accuracy of the Gaussian NB Classifier model looks 78.87063%, which is lowest classifier among the candidate models above. In this model, there are 512,735 misidentified cases which are significant. Hence, this model needs further improvement among all the candidate models.

Figure 14 Gaussian NB Classifier Classification Report

Conclusion and Future Work

Nowadays, attacks on CPS result in making the system out of function. As a result, it leads to a chain of problems for individuals, organizations, and various serious institutions for significant financial and data losses. IoT and CPS have shared similar architecture. Therefore, security issues are raised on IoT, similar to the CPS security issue.

To examine and propose the right candidate AI/ML framework in this research, we used the BoTNeTIoT-L01 dataset (Kaggle). A botnet attack is one type of attack that is infected with Malware. Unlike the conventional solution, we examine various AI/ machine learning models such as Random Forest, Decision tree, K-nearest, Ada boost, SGD, Extra Tree, and Gaussian NB Classifier to detect and classify an attack, whether it is a malware or benign on CPS.
Based on the captioned candidate algorithm, our experiments depict that Random Forest, Decision Tree, Ada Boost, and Extra Tree Classifier achieved 100% accuracy in detecting most attacks with Zero False-positive and False-negative rates. Thus, in this research, we proposed captioned candidate Malware detection framework in cyber-physical systems using Artificial Intelligence - Machine Learning through the experiments. Moreover, in this paper, we investigate the security challenges and issues of Cyber-Physical Systems that are limited to a similar state of the arts. In this regard, we hope that these CPS security

challenges and issues, precisely the detection and classification of attacks with AI ML models, bring enough motivation for future discussions and interests in the academic research work.

Finally, the proposed framework has been done based on the limited amount of data or a mid-size dataset, precisely 2,426,573 cases with twenty-five columns of data. Thus, if such an experiment were performed on a high-volume dataset or different attack types, the result may differ.

Acknowledgment

ARLIS (Applied Research Lab for Intelligence and Security) Grant supports this research work.

References

Alhowaide, A., Alsmadi, I., & Tang, J. (2019). Features Quality Impact on Cyber Physical Security Systems. *International Conference and Workshop on Computing and Communication (IEMCON).* Vancouver, BC, Canada. Retrieved from https://www.kaggle.com/datasets/azalhowaide/iot-dataset-for-intrusion-detection-systems-ids?resource=download

Ali, B., & Awad, A. I. (2018). Cyber and Physical Security Vulnerability Assessment for IoT-Based Smart Homes. *Sensors, 18*(3), 817.

Barbosa, J., Leitão, P., Trentesaux, D., Colombo, A. W., & Karnouskos, S. (2016). Cross benefits from cyber-physical systems and intelligent products for future smart industries. *IEEE 14th International Conference on Industrial Informatics (INDIN).* Poitiers, France.

Boehmke, B., & Greenwell, B. (2020, 02 01). *Hands-On Machine Learning with R.* Retrieved 29 10, 2021, from https://bradleyboehmke.github.io/HOML/DT.html

Breiman, L., & Schapire, E. (2001). Random forests. *Statistics Department, University of California, Berkeley, 45*(Machine Learning), 5 - 32.

Brownlee, J. (2020, 08 15). *Boosting and AdaBoost for Machine Learning*. Retrieved 05 04, 2022, from https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/#

Brownlee, J. (2020). Data Preparation for Machine Learning.

Chatterjee, M. (2020, 02 03). *Great Learning.* Retrieved 11 08, 2021, from https://www.mygreatlearning.com/blog/knn-algorithm-introduction/

Chaudhuri, A. (2019). *Internet of Things for Things, and by Things.* Boca Raton: aylor & Francis Group, LLC.

cisco. (2022). *What Is Malware?* Retrieved 03 25, 2022, from https://www.cisco.com/c/en/us/products/security/advanced-malware-protection/what-is-malware.html#~7-types-of-malware

Dobson, R. (2020, 10 14). *Implement K Nearest Neighbor Solution with T-SQL.* Retrieved 11 13, 2021, from https://www.mssqltips.com/sqlservertip/6596/k-nearest-neighbor-tsql-example/

ENISA. (2022). *ENISA Threat Landscape 2020*. Retrieved 03 14, 2022, from https://www.enisa.europa.eu/news/enisa-news/enisa-threat-landscape-2020

Fuchs, M. (2019, 11 11). *Introduction to SGD Classifier*. Retrieved 05 01, 2022, from https://michael-fuchs-python.netlify.app/2019/11/11/introduction-to-sgd-classifier/#:

Gandhi, R. (2018, 5 5). *Naive Bayes Classifier.* Retrieved 10 31, 2021, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

GHOSH, A. (2018, 04 25). *What is Cyber-Physical System (CPS)?* Retrieved 03 16, 2022, from https://thecustomizewindows.com/2018/01/cyber-physical-system-cps/

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications.*

Harrison, O. (2018, 09 10). *towardsdatascience.* Retrieved 25 10, 2021, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Houbing Song, G. A. (2018). Security and Privacy in Cyber-Physical Systems. Hoboken,: JohnWiley & Sons Ltd.

Houbing Song, G. A. (n.d.). Security and Privacy in Cyber-Physical Systems. In *2018* (p. 35). Hoboken,: JohnWiley & Sons Ltd.

Ippolito, P. P. (2019, 10 10). *Feature Extraction Techniques*. Retrieved 04 15, 2022, from https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be#:

Jain, S., Choudhari, P., & Srivastava, A. (2021). The fundamentals of Internet of Things: architectures, enabling technologies, and applications. *Healthcare Paradigms in the Internet of Things Ecosystem.*

Kaggle. (n.d.). *IoT dataset for Intrusion Detection Systems (IDS).* Retrieved 03 26, 2022, from https://www.kaggle.com/datasets/azalhowaide/iot-dataset-for-intrusion-detection-systems-ids?resource=download&select=BotNeTIoT-L01_label_NoDuplicates.csv

Kaur, R. (2021). Naive Bayes: A Text Classifier based on Machine Learning. *International Journal of Research Publication and Reviews*, pp. 260-266.

Kulkarni, R. (2020, 05 23). *Summary of KNN algorithm when used for classification*. Retrieved 05 01, 2022, from https://medium.com/analytics-vidhya/summary-of-knn-algorithm-when-used-for-classification-4934a1040983

Langkemper, S. (n.d.). *Security Problems IoT Devices*. (eurofins) Retrieved 03 22, 2022, from https://www.eurofins-cybersecurity.com/news/security-problems-iot-devices/

Liang Hu, N. X. (2012). Review of Cyber-Physical System Architecture. *International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops.* Changchun.

Mbaabu, O. (2020, 12 11). *Introduction to Random Forest in Machine Learning*. Retrieved 04 25, 2022, from https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/#

OWASP. (2022). (OWASP Top Ten) Retrieved 02 28, 2022, from https://owasp.org/www-project-top-ten/

Scikit-Learn. (2021). *Decision Trees.* Retrieved 11 1, 2021, from https://scikit-learn.org/stable/modules/tree.html

scikit-learn developers . (2022). *ExtraTreesClassifier*. Retrieved 05 04, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html#

Sharmeen, S., Huda, S., & Abawajy, J. (2019). International Conference on Smart Power & Internet Energy Systems. *IOP Conference Series : Earth and Environmental Science.*

Szczerbicki, E. (2008). *Decision Tree Classifier*. (ScienceDirect) Retrieved 04 22, 2022, from https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier#

Tahir, R. (2018). A Study on Malware and Malware Detection Techniques. *I.J. Education and Management Engineering, 2*, 20-30.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *IJISET - International Journal of Innovative Science, Engineering & Technology, 2*(9).

Zewdie, T. G., & Girma, A. (2022). An Evaluation Framework for Machine Learning Methods in Detection of DoS and DDoS Intrusion. *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).* Jeju Island, Korea.

Zewdie, T., & Girma, A. (2020). IoT security and the role of AI/ML to combat emerging Cyber threats in Cloud Computing Environment. *Information Systems Journal, 21*(4), 253-263.

Zsigovits, A. (2021, 09 07). *CUJOAI.* Retrieved 04 04, 2022, from https://cujo.com/mirai-gafgyt-with-new-ddos-modules-discovered/