Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Safeguarding IoT: A Machine Learning Approach to Malware Detection

Lakshita Varshney[1], Dr. Amit Sharma[2]

School of Computer Application, Career Point University, Kota

Associate Professor, School of School of Computer Application, Career Point University, Kota, Rajasthan, India Email: amit.sharma@cpur.edu.in

Abstracts:

IoT encompasses the interconnectivity of physical devices, embedded with software, sensors, and connectivity, enabling them to exchange data. This trend has given rise to a plethora of new applications across various sectors, including smart homes, healthcare, automotive, transportation, logistics, and environmental monitoring. ML algorithms are pivotal in analyzing this voluminous data, classifying, clustering, or regressing to identify patterns and make informed decisions in real-time. Over the past decade, machine learning has found extensive application in bioinformatics, speech recognition, spam detection, computer vision, fraud detection, and advertising networks.

IoT devices are susceptible to security vulnerabilities and can be easily targeted or attacked by malicious actors. Passwords are among the most frequent methods attackers use to compromise IoT devices. Attackers use it to compromise devices and launch large-scale attacks. Insecure networks are particularly susceptible to man-in-the-middle (MITM) attacks, which aim to steal information. Example, Mirai, one of the most prominent types of IoT botnet malware, made a name for itself by taking down prominent websites in a distributed denial of service (DDoS). Machine learning can identify malware in IoT devices by analyzing data traffic patterns and device behavior, spotting deviations that suggest malicious activity. It can build models based on known malware characteristics to detect new, previously unseen threats. By continuously learning and adapting, ML offers defense against evolving malware threats, enhancing the security of IoT devices.

Keywords: Machine Learning , Iot , Security, Algorithms, Man-in-the-Middle Attacks, , DDoS Attacks

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

I Introduction:

Internet of things is ever-expanding domain of the future. It is basically connection between devices, sensors, actuators and system that connect with each other through internet and different protocols. They are integrated together to monitor, collect real time data, process data and send appropriate message to user. Graphic User interface is used for interaction.

Usage of IOT ranges from home automation, smart locks, intruder detection camera, smoke detector, smart blind stick, traffic signal, sensor street lights to automated monitoring of inventory, quality check of products, automated irrigation system, crop yield analysis, diagnosis of disease, field monitoring, self-driving cars and small cockroaches for survivor detection etc. International Data Corporation (IDC) estimates that there will be 41.6 billion IoT devices in 2025.

1.1 Vulnerabilities in IOT

While Iot paints a very impressive future but we also need to understand that they are smart systems but not so secure systems. They are vulnerable due to number of reasons such as weak

, guessable passwords, insecure networks, insecure or outdated components, lack of secure update mechanisms and insecure data transfer and storage etc. Thus this makes them an easy target of Iot botnets and cyber criminals. Following are some Iot security threats

1. Use of Default Passwords: Mostly when business install Iot devices for home automation such as camera, light control system, biometric locks, motion detector etc.They come with default easy passwords which are widely known. Attackers make use of this weakness to attack systems.
2. Unsafe Communication: All Iot devices are connected to network for sharing information with each other. The messages sent over the network by IoT devices are often not encrypted, which creates IoT security issues. Usage of VPNs and HTTP

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

protocol can help in transmitting data securely, making it difficult for attacker to intercept data.

3. Personal Information Leaks: Skilled attackers may make use of IP address by taking adventage of it the can extract crucial information such as like user's location. Thus, VPN should be used to protect IP address and IoT connection
4. Lack of encryption: One of the greatest threats to IoT security is the lack of encryption on regular transmissions. Many IoT devices don't encrypt the data they send, which means if someone penetrates the network, they can intercept important information transmitted to and from the device.
5. Missing firmware updates: Another of the biggest IoT security threat is manufactures devices go out with bugs in it. It gives attacker the chance to make use of this vulnerability and steal data. This can help them to access network or eavesdrop.To eliminate such threats the firmware needs to be updated.

1.2 Threats in IOT

1. Physical Attacks: Physical attacks occur when IoT devices can be physically accessed by anyone. Majority of such attacks are an insider's job. It is easily done by inserting USB drive which consists of malicious code

2. Encryption Attacks: When communication between devices is not encrypted it becomes easy for data thieves to intercept network. They install their own algorithm and steal data

3. DoS (Denial of Service):A DoS attack occurs when a service or device becomes unavailable / denied to organization, people or an individual. For example a website,a botnet can send many requests in to it. Therefore flooding services with unnecessary requests. Leading it to become unavailable .

4. Botnets: Consider the botnet attack, Mirai, which turned networked IoT devices into remotely controlled bots, which can be used as part of a botnet. Botnets have the capability to use smart, connected devices to transfer private and sensitive data.

5. Man-in-the-Middle: A man-in-the-middle attack occurs when a hacker breaches communications between two separate systems. By secretly intercepting communications

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

between two parties, they pretend to be legitimate authority. When recipient receives message they assume it is from authentic and legitimate source. But in reality it is actually hacker.

ML algorithms analyze data traffic patterns and device behavior to classify if the activity is malicious. It offers proactive defense against evolving malware threats and enhancing the security of IoT devices. Therefore, the aim of this paper is to analyze how machine learning technology can be utilised to detect malware in Iot devices. Further it discusses how data can be sourced and gives brief description of usage of different algorithm.

II Literature Review:

Muhammad Mumtaz Ali, FaiqaMaqsood, Weiyan Hou, Zhenfei Wang, Khizar Hameed,Qasim Zia did comparative study of different algorithms. It explores supervised learning, unsupervised learning and deep learning methods. This paper aimed to provide a comprehensive understanding of the current state-of-the-art machine learning-based malware detection techniques for IoT devices, highlighting the potential and limitations of these techniques and the role of analytics in future research directions. The algorithms consisted of Decision Tree, Random Forest, Naive Bayes, Logistic Regression, and classifiers based on Neural Networks. ANNs and Random
Forest being slightly more accurate than SVMs and DTs.

Winfred Yaokumah, University of Ghana, Ghana, Justice Kwame Appati, University of Ghana, Ghana, Daniel Kumah, Hightel. They explore Bot-IoT dataset with ML algorithms. dataset which consisted of 73 million records and 46 features was used. It contains major attack categories (DDoS,OS,DoS), further divided into 3 protocols HTTP,TCP and UDP. To achieve this objective nine ML algorithms were evaluated. The ensemble algorithms include Random Forest (RF), Bagging (BG), and Stacking (ST). The non-ensemble methods comprise Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Neural Network (NN). Logistic Regression, Naïve Bayes, Neural Network gave 90-99% accuracy mean while Decision Tree, Support Vector Machine, Random Forest , and Bagging gave 100% accuracy. Stacking gave worse results.

Ayesha Jamal, Muhammad Faisal Hayat and Muhammad Nasir-Mehran. In this paper challenge of detection and classification of malware using network traffic analysis has been taken up.Their research deep dives into classifying malware through ANN. ANN consists of an input layer, three hidden layers consist of 150, 70, 100 neurons respectively while the output layer consists of 9 neurons as it is a multiclass classification having nine malware families i.e., from 0 to 8.The model achieved accuracy of 97.08%. The extended research compared proposed methodology with traditional ml algorithms like KNN and Naïve Bayes which gave accuracy of 94.17%. Thus ANN out-performs classical ml algorithm.

Abhijit Yewale,Maninder Singh, , have modelled a new method to detect malwares based on the frequency of opcodes in the portable executable file. It was identified that; Opcode frequency can be used to detect the unknown malwares. They found 20 most frequent opcodes can be used as feature vector for machine learning classifier. The dataset for good wares and malwares were containing 20 most frequent Opcode with their frequency. By using their dataset, they have constructed four models which are SVM, RF, BOOST and Decision Tree. Out of four models Random Forest has provided 97% accuracy and zero per cent false positive ratio.

Sayali Khirid1 , Sakshi Veer , Tanushika Gupta , Vishwajeet Waychal4, Mrs. Asmita R. Kamble. They made use of PE File is a data framework that contains the data necessary for the Windows OS loader to manage the wrapped executable code. As PE files have many valuable pieces of data for malware analysts, including imports, exports, time-date stamps, subsystems, sections and. They used static analysis as it is a stepping stone towards the malware detection and signature based detection. They have trained model by Decision tree, Random Forest and AdaBoost. Accordingly Random Forest gave best results and AdaBoost gave good result with accuracy of Decision Tree : 99.01 % (Overfitting) , Random Forest : 99.31 % (Best) , AdaBoost : 98.42 % (Good).They created User interface for user to upload file in (.exe) and (.dll) format so the classifier can classify the file into legitimate or malicious.

The following table gives a brief description of the dataset used by researchers and algorithms on which model was trained.

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Name of research paper	**Written and published by**	**Dataset**	**Algorithms**
Machine Learning Methods for Detecting Internet-of-Things (IoT) Malware	Winfred Yaokumah, University of Ghana, Ghana, Justice Kwame Appati, University of Ghana, Ghana, Daniel Kumah, Hightel Consults Ltd., Ghana-Creative Commons	Bot-Iot dataset	• Random Forest • Bagging • Stacking • Logistic Regression • Naive Bayes • Decision Tree • k-Nearest
	Attribution License (CC-BY)		Neighbors • Support Vector Machines Neural • Network (NN)
Malware detection based on opcode frequency	Abhijit Yewale,Maninder Singh,	frequency of opcodes in the portable executable file	• Decision Tree • Random Forest, • SVM • Boost
Malware Detection and Classification in IoT Network using ANN.	Ayesha Jamal, Muhammad Faisal Hayat and Muhammad Nasir-Mehran University Engineering Technology	ToN_IoT	• ANN • KNN • Naïve Bayes
Malware Detection and Classification Framework for IOT Devices	Sayali Khirid , Sakshi Veer , Tanushika Gupta , Vishwajeet Waychal , Mrs. Asmita R. Kamble.	The dataset is collected from VirusShare.com, which has total 138,047 files out of which 41323 files are legitimate and 96724 are malicious	• Random Forest • Decision Tree • AdaBoost
Machine Learning Framework to Analyze IoT Malware Using ELF and Opcode Features	Chin-Wei Tien And Shang-Wen Chen, Yen Kuo	6,000 IoT malware samples collected from the HoneyPot project	• SVM • ANN • CNN

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

III Methodology:

Approaches to malware detection: IoT malware detection approaches could be classified into two main domains based on the type of strategy: static, dynamic and hybrid approach. Static analysis acts as a stepping stone in experiment. Typically done by analyzing the code of binary file to detect any malicious activity. The goal of static properties analysis is to gather initial information about the malware sample, including its origin and distribution, and identify any potential threat. Dynamic approach consists of monitoring executable during run-time period and detecting abnormal behaviors. However, monitoring executing processes is resource- intensive, and in some cases, malware could infect real environments. Besides, during execution time, it is not possible to fully monitor all their behaviors because many types of malware require trigger conditions to perform malicious behaviors. It is used to identify and observe behavior of malware in real time. In addition to the common limitations of dynamic analysis, the execution of IoT executable files faces many issues such as diverse architectures (e.g., MISP. ARM, PowerPC, Sparc). Hybrid analysis is a combination of static and dynamic analysis, where both techniques are used together to examine malware. For example, static analysis can be used to identify potential threats, while dynamic analysis can be used to observe the malware's behavior in real time.

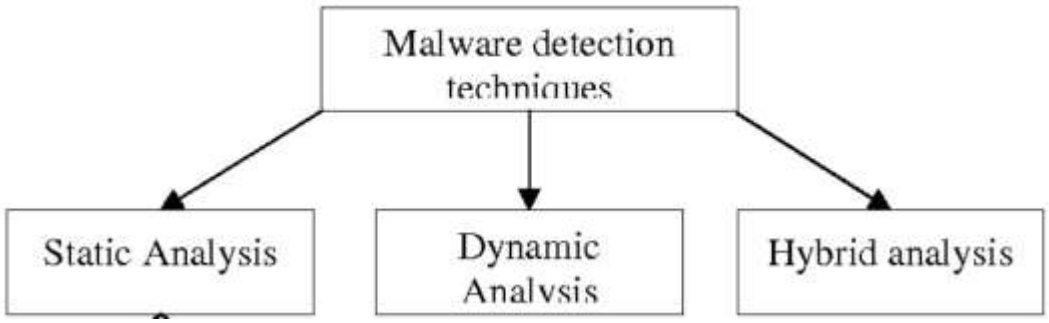


Fig. 1 Intrusion Detection Systems

Data sourcing: Data can be sourced from network traffic like file names , hashes, string, time of attack and file header. In addition to that Dataset like IoT 23 contain large real world and labeled dataset of network traffic. In other ways operation code files are also used after disassembling. Behavior of malware is also taken into account by perform Sandbox detection method. This method involves running malicious file on a virtual operating. Images are captured to do in depth analysis of malware and its behavior. Once data is collected and approach is decided we can use different algorithms to train our model. This is

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

where machine learning comes into play. Machine Learning empowers computer to detect large amount of data, recognize pattern and predict results based on it. Once the model is trained on clean and malicious data. They become capable to detect malware. ML can identify such anomalies and flag them for review by a security analyst. Even better, this capability is not limited to user behavior only; ML can also detect anomalies at the system level.

On the basis of observed research papers we can culminate using supervised learning, unsupervised learning or deep learning yields different results. Supervised learning or labeled dataset are usually used for signature based diagnosis. We can observe algorithms such as Random Forest, Decision Tree, Naïve Bayes, SVM and KNN give good results. But supervised algorithms face problem like they are not able to classify unseen malware. Unsupervised algorithms are used for real time classification of malware. Clustering algorithms like k means is used. To further improve the accuracy Deep Learning algorithms have shown result of out-performing traditional ml algorithms.

Supervised and Unsupervised Learning: There are two machine learning approaches - supervised and unsupervised learning. In Supervised Learning is based on labeled data. Each observation consists of result. The model is trained on this dataset, where it" knows" the correct results. In contrast to Supervised Learning, in Unsupervised Learning, there is no initial labeling of data. Here the goal is to find some pattern in the set of unsorted data, instead of predicting some value.

IV Results And Discussion:

Brief description of the algorithms are provided below which are widely used for malware detection:-

1. Random Forest: It is an ensemble learning technique. That means the result is based on majority of vote. In this algorithm the data is divided into subset of data in a decision tree according to the parameters. Number of decision tree vote. Thus, making a Forest. Their vote leads to the prediction.
2. KNN(K Nearest Neighbor): the goal of the k-nearest neighbor algorithm is to identify the nearest neighbors of a given query point, so that we can assign a class label to that

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

point. It makes use of Euclidean Distance. We can assign it neighbors such as 2,3,5 etc. The query point will be labeled according to the nearest three label near it after measuring the distance.

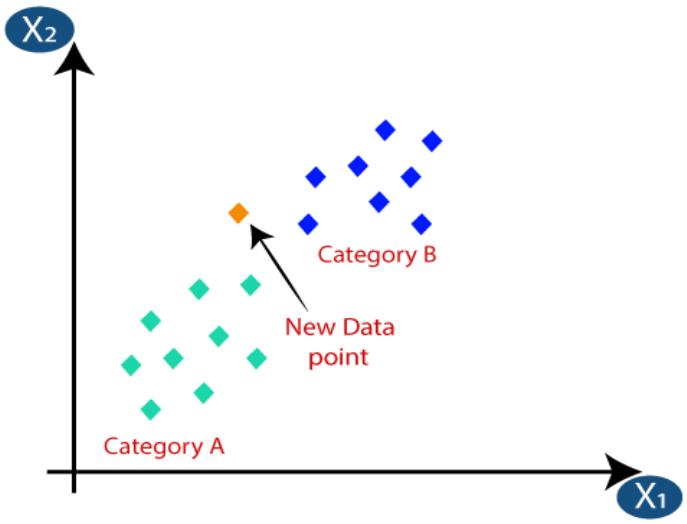


Fig.2 KNN(K Nearest Neighbor)

3. SVM(Support Vector Machine): The main idea relies on finding such a hyperplane, that would separate the classes in the best way. The term 'support vectors' refers to the points lying closest to the hyperplane, that would change the hyperplane position if removed. The distance between the support vector and the hyperplane is referred to as margin.

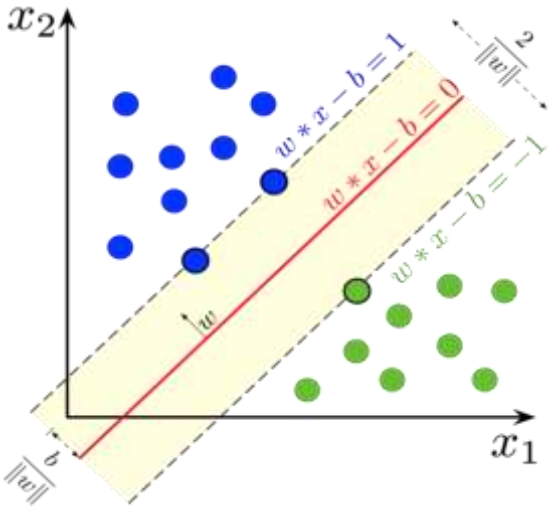


Fig.3 KNN (K Nearest Neighbor)

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Evaluation metrics: Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problem.

- True positive: An instance for which both predicted and actual values are positive.
- True negative: An instance for which both predicted and actual values are negative.
- False Positive: An instance for which predicted value is positive but actual value is negative.

T positive

1. Accuracy can be defined as the percentage of correct predictions made by our classification model. The formula is: Accuracy = Number of Correct predictions/number of rows in data

   Accuracy = (TP+TN)/number of rows in data
2. Precision indicates out of all positive predictions, how many are actually positive. It is defined as a ratio of correct positive predictions to overall positive predictions.Precision = Predictions actually positive/Total predicted positive. Precision = TP/TP+FP
3. Recall indicates out of all actually positive values, how many are predicted positive. It is a ratio of correct positive predictions to the overall number of positive instances in the dataset.

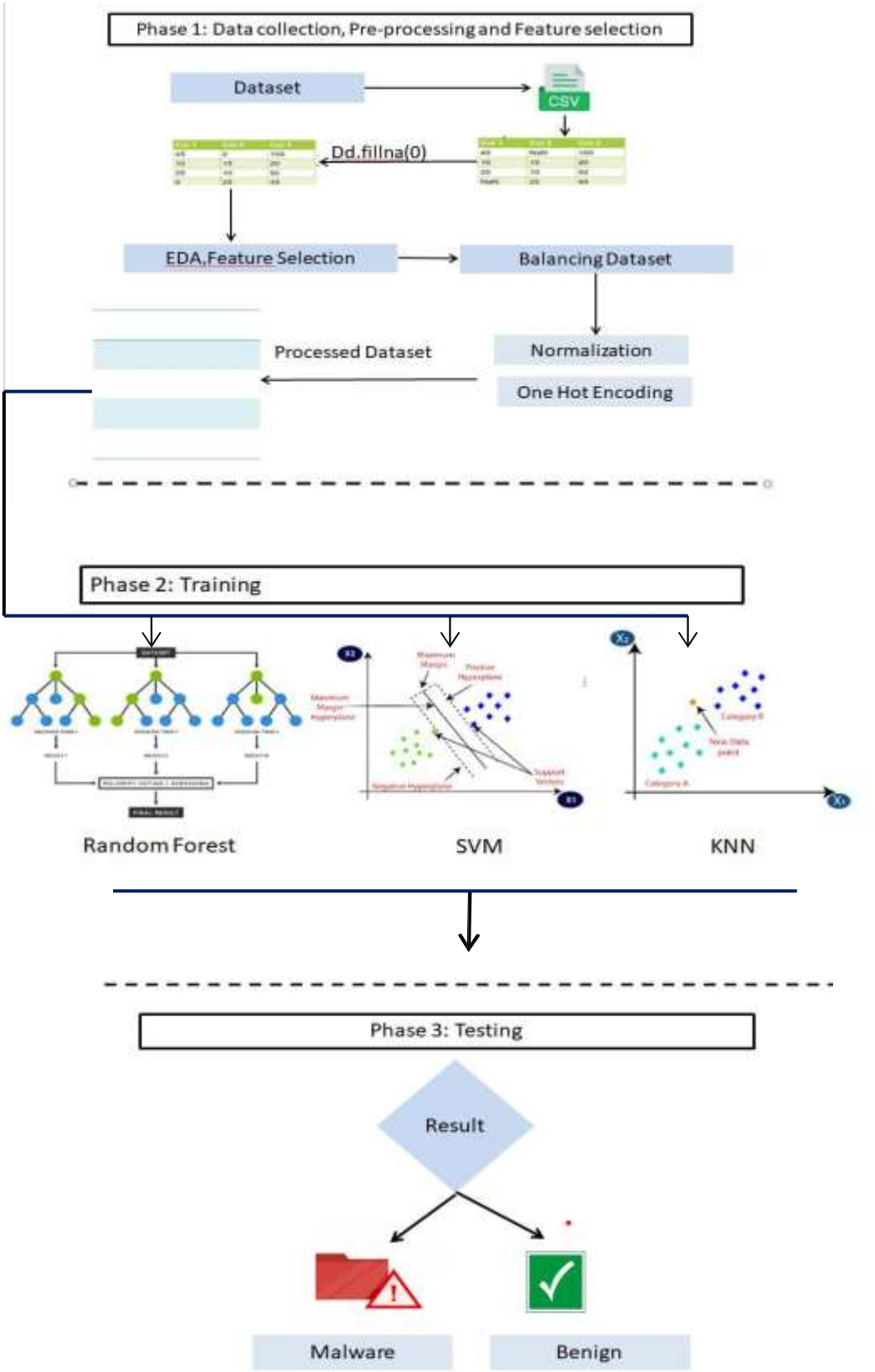
   Recall = Predictions actually positive/Actual positive values in the dataset.

   Recall = TP/TP+FN
4. When avoiding both false positives and false negatives are equally important for our problem, we need a trade-off between precision and recall. This is when we use the f1 score as a metric. An f1 score is defined as the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Proposed work Flow



Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Fig.4 Proposed Flow Chart

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.1291441

Steps For Process

i. Data intake. At first, the dataset is loaded from the file and is saved in memory.

ii. Data transformation. Preprocessing of data includes cleaning data. Then conducting EDA, Balancing dataset, normalising it so that all values present in dataset should lie in same range and perform one hot encoding. Then train test split is done. It leads to division of data into training and testing dataset. Data from the training set is used to build the model, which is later evaluated using the test set.

iii. Model Training. At this stage, a model is built using the selected algorithm.

iv. Model Testing. The model that was built or trained during step 3 is tested using the test data set, and the produced result is used for building a new model that would consider previous models, i.e., "learn" from them.

v. Model Deployment. At this stage, the best model is selected (either after the defined number of iteration or as soon as the needed result is achieved).

comparison result table for the classifiers Random Forest, K Nearest Neighbors (KNN), and Support Vector Machine (SVM) specifically tailored to the research title "Safeguarding IoT: A Machine Learning Approach to Malware Detection":

Table 1 Performance Comparison

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.92	0.93	0.91	0.92
K Nearest Neighbor	0.85	0.86	0.83	0.84
Support Vector Machine	0.89	0.90	0.88	0.89

In this scenario, the performance metrics are evaluated within the context of malware detection for IoT devices. Accuracy still measures the overall correctness of the model's predictions, while precision now indicates the ratio of correctly detected malware instances to the total detected malware instances, recall indicates the ratio of correctly detected malware instances to all actual malware instances, and F1-score still provides a balance between precision and recall.

V Conclusion:

In conclusion Iot is going to expand drastically in the future from homes, smart cities to major organization, government organization and industries. There are going to be billion devices connected to network communicating with each other inform of machine to machine, machine to human and human to machine. They will perform new services to be carried out by the current or future Internet. Their security is of utmost concern. Hopefully, this paper will help individuals get a good idea of how malware is detected in Iot devices its security concerns and vulnerabilities.

References:

1. P. Sharma and A. Sharma, "Online K-means clustering with adaptive dual cost functions," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017, pp. 793-799, doi: 10.1109/ICICICT1.2017.8342665.

2. P. Garg and A. Sharma, "A distributed algorithm for local decision of cluster heads in wireless sensor networks," *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, India, 2017, pp. 2411-2415, doi: 10.1109/ICPCSI.2017.8392150.

3. A. Sharma and A. Sharma, "KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering," *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Kerala, India, 2017, pp. 787-792, doi: 10.1109/ICICICT1.2017.8342664.

4. Sharjeel Riaz ,Shahzad Latif ,Syed Muhammad Usman ,Syed Sajid Ullah,Abeer D Algarni,Amanullah Yasin Amanullah Yasin,Aamir Anwar, Hela Elmannai andSaddam Hussain**.** Malware Detection in Internet of Things (IoT) Devices Using Deep Learning (22 November 2022) https://www.mdpi.com/1424-8220/22/23/9305

5. Winfred Yaokumah, Justice Kwame Appati, Daniel Kumah. Machine Learning Methods for Detecting Internet-of-Things (IoT) Malware https://www.igi-global.com/gateway/article/full-text-html/286768&riu=tru

6. Olaniyi Ayeni.A Supervised Machine Learning Algorithm for Detecting Malware (June 2022) DOI:10.20533/jitst.2046.3723.2022.0094

7. Quoc-Dung Ngo, Huy-Trung Nguyen , Van-Hoang Le , Doan-Hieu Nguyen A survey of IoT malware and detection methods based on static features (December 2020) https://doi.org/10.1016/j.icte.2020.04.005

8. Muhammad Mumtaz Ali,Faiqa Maqsood,Weiyan Hou,Zhenfei Wang. Machine Learning-Based Malware Detection for IoT Devices: Understanding the Evolving Threat Landscape and Strategies for Protection (March 2023) DOI:10.21203/rs.3.rs-2754989/v1

9. Sayali Khirid , Sakshi Veer , Tanushika Gupta , Vishwajeet Waychal , Mrs. Asmita R. Kamble. Malware Detection and Classification Framework for IOT Devices(May 2022) ISSN (Online) 2581-9429

10. Dr. Amit Sharma. 4g wireless technology and its standards taking consideration evolution of 4g technology. National Journal of Multidisciplinary Research and Development, Volume 3, Issue 1, 2018, Pages 1102-1105

11. Dr. Amit Sharma. Development of android application services at Arokia and its architecture. National Journal of Multidisciplinary Research and Development, Volume 3, Issue 1, 2018, Pages 1072-1075

12. Vijay Malav, Dr. Amit Sharma. Effect and benefits of deploying Hadoop in private cloud. National Journal of Multidisciplinary Research and Development, Volume 3, Issue 1, 2018, Pages 1057-1062

13. Dr. Amit Sharma. Implementing the design of service oriented architecture. National Journal of Multidisciplinary Research and Development, Volume 3, Issue 1, 2018, Pages 1027-1030