Securing The IoT: A Machine Learning Approach to Cyber Threat Analysis

Jani Shaik[1], Mogasala Bhanu[2*], Doddi Bhoomika[3], Burri Divya[4] , Kornu Himasagar[5], Prof Ashok Patel[6]

[1,2,3,4,5]Assistant Professor, Dept of CSE, Nadimpalli Satyanarayana Raju Institute of Technology Visakhapatnam, 531173, A.P.India
[6]University of Massachusetts Dartmouth, USA

skj.shaikjani@gmail.com, *mogasalabhanu@gmail.com, doddibhoomika@gmail.com

Abstract. The rise of Internet of Things (IoT) devices[1] has brought about numerous conveniences, but it has also introduced security challenges, notably the alarming threat of botnets. These malicious networks can compromise a large number of devices, posing significant risks to privacy, data integrity, and network availability. Recent years have seen the effectiveness of Machine Learning (ML) techniques in addressing this concern by identifying and mitigating IoT botnet attacks. The proposed system utilizes both supervised and unsupervised ML algorithms, such as classification, decision trees, and logistic regression, to identify compromised IoT devices. Trained on carefully labelled datasets, the system learns distinctive features and patterns associated with malicious activities, employing feature engineering to enhance accuracy. Real-time monitoring and anomaly detection are integrated to promptly respond to botnet-related activities. Ensemble learning methods further strengthen the system, resulting in high accuracy and minimized false alarms, contributing significantly to the security of IoT networks and paving the way for future advancements in IoT security.

Keywords: — IoT (Internet of Things), botnets, machine learning, Classification, Decision Tree, Logistic Regression.

1. Introduction

The surge in Internet of Things (IoT) attacks is a direct result of advancements in IoT devices[1]. Among these threats, IoT botnet attacks stand out as particularly hazardous, as they involve compromised IoT devices being used for cybercrimes[1]. Detecting security vulnerabilities and defending against these evolving threats poses significant challenges for IoT systems[1]. Although machine learning and deep learning methods[1] using time series data have shown improvement in detecting malware, early identification is crucial for effective IoT botnet response strategies. Timely recognition can mitigate the damage caused by potential attacks, providing essential data for deep learning and machine learning models in malware detection. In the context of a DDoS attack carried out by an IoT botnet[1], identification becomes more manageable once the attack has occurred. The use of bot malware and botnets extends beyond DDoS attacks, encompassing activities like spam, virus dissemination, click fraud, and other unwanted actions. The lifespan of an IoT botnet involves

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_123

an extended search and spreading phase, making it imperative to identify and isolate bots before they launch a full-scale attack, such as a DDoS. However, recognizing botnets, especially peer-to-peer botnets, can be challenging. Machine learning is a captivating aspect of artificial intelligence, successfully achieving the goal of learning from data with specific machine inputs. The process involves inputting training data into the chosen algorithm to initiate machine learning. The final machine learning algorithm is developed using this training data, whether known or unknown. To assess its effectiveness, the algorithm is tested with new input data, and the outcomes are cross-checked. If predictions do not align with results, the algorithm is iteratively retrained until the desired outcome is achieved. This self-training process enables the machine learning algorithm to continually improve, enhancing accuracy over time. A Machine Learning-based Classification and Prediction Technique for DDoS Attacks [2,5].

2. Literature Survey

Internet of things[1]. It is possible to observe that the most well-known problem is information theft by unauthorized users, as well as access to any information sources via unauthorized devices owned by private organizations. These, along with the client information that is thrifted, aid in the growth of financial frauds. Additionally, since the information thief or programmer is signed into the framework through devices and unapproved sources, it is impossible to track their exact location.
Types of attacks:
1. Vector-based computer virus
2. Data Breaching
3. DDOS
These are the most excellent illustrations of the ongoing cyberattacks that affect departmental network protection personnel worldwide. They obtain access to break the secret word and the username of their web servers by using any kind of outside device or secret programming.

Malware: It gives admittance to control any gadgets and assists with getting the information by utilizing an unapproved server or organization.
Botnet**:** programmers are mostly used when they are attempting to gain admittance to the client's account or the client's gadgets without requesting any consent from the client (Bansal et al., 2021). This is a sort of set of guidelines that assists with gaining admittance to the casualty's framework or the PCs, and the programmers can get the chance to introduce or infuse the secret programming inside the client's gadgets without asking for their consent.

In the existing system, there was no appropriate method that could identify the IOT attacks which are created by hackers in IOT devices [5]. All the existing systems suffer from this problem to trace the attacks that are present inside the IOT networks. Also, there were a lot of hacking issues for the IOT devices for getting valid information. IoT gadgets are the most generally utilized gadgets across the world, yet there are a few hubs that are feeble as far as security, and programmers are looking for those sorts of hubs.
.
In the realm of cybersecurity, researchers are perpetually challenged by the evolving landscape of online threats. As they strive to bolster defenses, cybercriminals relentlessly innovate, devising new strategies to identify and exploit security vulnerabilities. The transmission of malware, in particular, has seen a proliferation of novel and sophisticated methods. Once malware infiltrates systems, compromised machines become potent tools for executing further attacks. These may range from data theft to orchestrating denial of service attacks, amplifying the impact and scope of cyber threats. This perpetual cat-and-mouse game underscores the dynamic nature of cybersecurity and the ongoing need for vigilance, adaptation, and innovation in defense strategies.

The functionality and accessibility of Internet of Things (IoT) services and applications[1] have led to a major growth in their general usage. From simple personal devices to more sophisticated applications like smartwatches, smart manufacturing, smart mining, and driverless autonomous automobiles, businesses have created a wide range of Internet services. However, the widespread adoption of IoT has also drawn potential hackers who want to commit cyberattacks and data theft. The Internet of Things has brought up extremely pressing cybersecurity challenges. The main goal is to present a novel model for botnet detection and prevention that is based on machine learning algorithms [7].Identify IoT Botnet Threats,Explore Theories and Models,Develop Integrated Methods

3 Proposed methodology

It focuses solely on utilizing Plotly with Python to mitigate security threats posed to various networking systems by Botnet attacks. It emphasizes both categorical and numerical variables relevant to the project, with Python employed for basic data visualization, particularly for numerical values. Initially, a dataset encompassing diverse threats within the Internet of Things (IoT) Botnet ecosystem is created, considering the spectrum of threats arising from this domain. Thorough observation of this dataset is crucial for the development of an integrated and advanced application aimed at tackling IoT Botnet threats effectively. Following dataset training and observation, variables are categorized to facilitate easier threat identification. Subsequent data cleaning ensures accurate results by removing irrelevant data. Implementation of the cleaned dataset in the research work follows suit. Despite obtaining results, the accuracy remains undetermined. Implementing the system in the real world without accurate results renders the endeavor futile. Therefore, assessing precision, recall, and accuracy becomes imperative to ensure the effectiveness of the integrated application in addressing IoT Botnet threats. If these metrics fall short, the research's relevance diminishes, hindering its ability to resolve issues stemming from IoT Botnet threats.

3.1 Dataset Description

The models employed in this project are designed and trained to classify botnet attacks, utilizing the UNSW-NB15 dataset sourced from Kaggle. This publicly accessible dataset encompasses nine types of assaults, such as worms, fuzzes, analysis, backdoors, denial-of-service, exploits, generic, reconnaissance, and shellcode.

```
['id' 'dur' 'proto' 'service' 'state' 'spkts' 'dpkts' 'sbytes' 'dbytes'
 'rate' 'sttl' 'dttl' 'sload' 'dload' 'sloss' 'dloss' 'sinpkt' 'dinpkt'
 'sjit' 'djit' 'swin' 'stcpb' 'dtcpb' 'dwin' 'tcprtt' 'synack' 'ackdat'
 'smean' 'dmean' 'trans_depth' 'response_body_len' 'ct_srv_src'
 'ct_state_ttl' 'ct_dst_ltm' 'ct_src_dport_ltm' 'ct_dst_sport_ltm'
 'ct_dst_src_ltm' 'is_ftp_login' 'ct_ftp_cmd' 'ct_flw_http_mthd'
 'ct_src_ltm' 'ct_srv_dst' 'is_sm_ips_ports' 'attack_cat' 'label']
['id' 'dur' 'proto' 'service' 'state' 'spkts' 'dpkts' 'sbytes' 'dbytes'
 'rate' 'sttl' 'dttl' 'sload' 'dload' 'sloss' 'dloss' 'sinpkt' 'dinpkt'
 'sjit' 'djit' 'swin' 'stcpb' 'dtcpb' 'dwin' 'tcprtt' 'synack' 'ackdat'
 'smean' 'dmean' 'trans_depth' 'response_body_len' 'ct_srv_src'
 'ct_state_ttl' 'ct_dst_ltm' 'ct_src_dport_ltm' 'ct_dst_sport_ltm'
 'ct_dst_src_ltm' 'is_ftp_login' 'ct_ftp_cmd' 'ct_flw_http_mthd'
 'ct_src_ltm' 'ct_srv_dst' 'is_sm_ips_ports' 'attack_cat' 'label']
```

Fig. 1. Dataset Attributes and columns

By examining and comparing the column names between the training and testing datasets, one can discern similarities or differences in the features available for model training and evaluation. This step is essential for ensuring consistency in data representation across different sets, as models are trained on the features present in the training data and subsequently tested on similar features in the testing data. Furthermore, knowledge of the column names facilitates the selection of relevant features for analysis or model training, aiding in the identification of input variables and target variables. This insight into the dataset's structure is valuable for preprocessing steps, feature

engineering, and overall understanding of the data's characteristics, contributing to the development of accurate ML models.This research utilizes a dataset centered around threats in IoT botnets. The dataset encompasses various cases directly associated with the threats prevalent in IoT botnets. To conduct this study, two distinct datasets have been employed, each containing instances related to the identified threats in IoT botnets.

Identifying the Threats

Addressing multiple important topics is necessary for machine learning-based IoT device security. Machine learning helps with anomaly detection in the context of access control and authentication, which are essential for preventing unwanted access. Machine learning can identify unusual activities, and firmware security and device integrity are critical. Secure communication protocols and intrusion detection technologies are used to combat network security. Machine learning-based monitoring improves device lifecycle management. By merging security measures with machine learning for anomalous event detection, physical security concerns are solved. Machine learning enhances supply chain security by spotting irregularities in workflow. Machine learning patterns help detect Denial of Service (DoS) attacks and reduce their impact. Machine learning-driven behavioral analytics defines baseline behavior and spots anomalies. Respecting regulatory compliance

Methods:

Decision Tree Classifier

Because logistic regression works well for binary classification problems, it is useful for identifying IoT botnet threats. Logistic regression is used in threat detection to model the likelihood that an event falls into the threat or non-threat category. This method offers insights into the traits linked to botnet threats and is especially useful when working with datasets that contain a large number of attributes. Logistic regression is a practical option for comprehending and recognizing patterns suggestive of malicious activity in Internet of Things environments since it is lightweight, computationally efficient, and interpretable. Its effective use in IoT botnet threat detection can be attributed to its simplicity and capacity to manage feature sets of a moderate size.

Logistic Regression Tree

Decision trees are effective in the detection of IoT botnet threats, offering a transparent and intuitive approach to classification. In threat detection, decision trees analyze the features of IoT devices to create a hierarchical structure of decision rules. This allows for the identification of patterns associated with botnet activities. Decision trees are particularly adept at handling both numerical and categorical data, making them well-suited for diverse IoT datasets. Their interpretability is a key advantage, enabling security analysts to understand the criteria leading to threat predictions. Additionally, decision trees can automatically identify important features in the detection process, contributing to the efficient recognition of IoT botnet threats based on their distinctive characteristics.

4. Results

We attempt to use Python as a programming language and then attempt to create the present task in order to demonstrate the performance of our suggested application. After deploying the present application, we will receive the following categories.

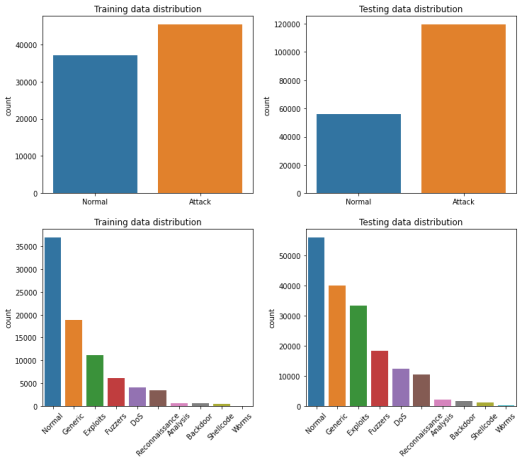

Fig.2. Graphical representation of the Training and Testing Data distribution

These lengths are crucial metrics in understanding the size and scale of the datasets, which is often essential in machine learning and data analysis tasks. To elaborate further, the lengths of the datasets indicate the number of instances or observations available for training and testing purposes. In machine learning, the training dataset is typically used to train a model, allowing it to learn patterns and relationships within the data. The testing dataset, on the other hand, is used to evaluate the model's performance on unseen data, providing an indication of its generalization capabilities.By printing and comparing the lengths of the training and testing datasets, one can gain an initial understanding of the data distribution and the ratio between training and testing samples. This information is valuable for assessing the adequacy of the dataset for building and evaluating machine learning models, as a well-balanced and representative dataset is crucial for achieving reliable and accurate results.

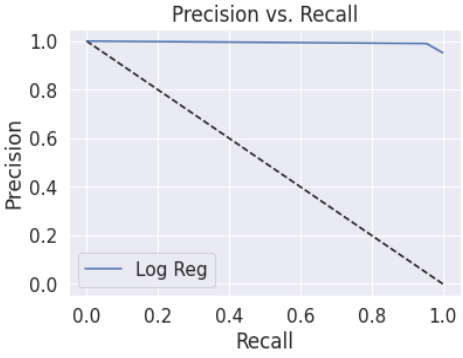

Fig.3. Binary Classification and Graphical Representation of Precision vs Recall

The trade-off between precision and recall across various threshold values used to classify cases is graphically represented by this curve. In classification tasks, recall and precision are two crucial measures. Precision is a metric used to quantify the accuracy of positive predictions. It shows the percentage of correctly predicted positive instances out of all positive forecasts. Conversely, recall measures the percentage of accurately predicted positive instances among all true positive instances, hence evaluating the model's capacity to capture all genuine positive cases. When dealing with imbalanced datasets—where

one class considerably outnumbers the other—the precision-recall curve is especially helpful. It sheds light on how the model performs differently at various decision thresholds. The precision-recall curve's points are represented by the threshold values in the lr. Analyzing the trade-off between recall and precision at various decision thresholds is necessary for understanding the precision-recall curve. A model with high recall and high precision across a range of threshold values is indicated by a curve that remains closer to the upper-right corner, which denotes superior overall performance. Depending on the particular needs of the application, this visual representation is helpful in choosing the right threshold based on the intended balance between precision and recall.

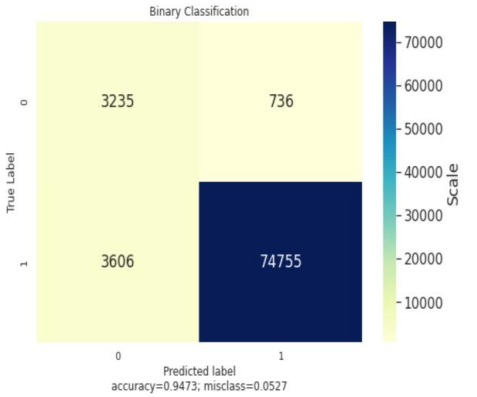

Fig.4. Machine Learning Analysis of Logistic Regression

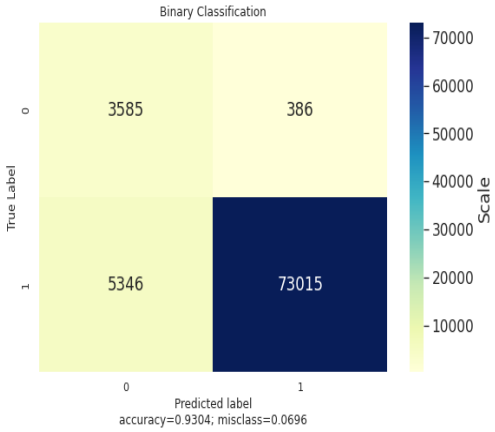

Fig.5. Machine Learning Analysis of Decision Tree Classifier

It Integrates key evaluation metrics (F1-score and confusion matrix) and a custom visualized function to assess the performance of a Decision Tree classifier in a binary classification scenario. The printed output aims to provide a

concise summary of the model's effectiveness, incorporating both quantitative and potentially visual information for a comprehensive assessment.

	precision	recall	f1-score	support
0	0.47	0.81	0.60	3971
1	0.99	0.95	0.97	78361
accuracy			0.95	82332
macro avg	0.73	0.88	0.79	82332
weighted avg	0.97	0.95	0.95	82332

Fig.6. Performance of Logistic regression

Accuracy 0.9999409589307727

	precision	recall	f1-score	support
0	1.00	1.00	1.00	771574
1	1.00	1.00	1.00	769726
accuracy			1.00	1541300
macro avg	1.00	1.00	1.00	1541300
weighted avg	1.00	1.00	1.00	1541300

AUC Score:
0.9999408880562694

Fig.7. Performance of Decision Tree Classifier

The classification report serves as a vital tool for practitioners to evaluate the performance of their models comprehensively. By scrutinizing this report, practitioners gain insights into the model's strengths and weaknesses across various classes. This allows them to discern which classes the model excels in and where it encounters challenges, providing valuable guidance for refining the model. Moreover, the report aids in identifying potential biases inherent in the model's predictions, enabling practitioners to address these biases effectively. Armed with this information, practitioners can make informed decisions about deploying the model, tailoring its application to specific requirements. Overall, the classification report offers a succinct yet insightful summary of the classifier's performance, facilitating the interpretation and assessment of its effectiveness across different classes.

5 Conclusion

In the ever-changing landscape of technology, devices are continually advancing, especially in the realm of IoT. However, this progress brings a heightened risk of hacking through IoT Botnets. Prioritizing the security of users' personal data is paramount in the development of advanced system applications for IoT devices. To counteract threats from IoT Botnets, it is crucial to regularly test the application's effectiveness using diverse threat datasets in Python coding systems. Continuous improvement involves the use of various threat datasets, regularly updated to address evolving risks. The coding process and Python software should undergo periodic updates for enhanced functionality. Testing methods also require frequent updates to ensure quicker and more accurate results. Improving sensors and nodes is

vital for swift detection of malware or hacking activities, allowing for immediate user notification and response.

References:

1. AL-Hawawreh, Nour Moustafa, Elena Sitnikova. "Identification of malicious activities in industrial internet of things based on deep learning models", Journal of Information Security and Applications, 2018.

2. A. Ismail, Muhammad Ismail Mohmand, Hameed Hussain, Ubaid Ullah et al. "A Machine Learning based Classification and Prediction Technique for DDoS Attacks", IEEE Access, 2022..

3. McDermott, C.D., Majdani, F. and Petrovski, A.V., 2018, July. Botnet detection in the internet of things using deep learning approaches. In 2018 international joint conference on neural networks(IJCNN) (pp. 1-8). IEEE.

4. Kaska, K., Beckvard, H. and Minarik, T., 2019. Huawei, 5G and China as a security threat. NATO Cooperative Cyber Defence Center for Excellence (CCDCOE), 28.

5. Alieyan, K., Almomani, A., Abdullah, R., Almutairi, B. and Alauthman, M., 2021. Botnet and Internet of Things (IoT s): A Definition, Taxonomy, Challenges, and Future Directions. In Research Anthology on Combating Denial-of-Service Attacks (pp. 138-150). IGI Global.

6. Sagirlar, G., Carminati, B. and Ferrari, E., 2018, October. AutoBotCatcher: blockchain-based P2P botnet detection for the internet of things. In 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC).

7. Letteri, I., Penna, G.D. and Gasperis, G.D., 2019. Security in the internet of things: botnet detection software-defined networks by deep learning techniques. International Journal of High Performance Computing and Networking.

8. Avanija, J., K. E. Kumar, Ch Usha Kumari, G. Naga Jyothi, K. Srujan Raju, and K. Reddy Madhavi. "Enhancing Network Forensic and Deep Learning Mechanism for Internet of Things Networks." (2023).9 AL-Hawawreh, Nour Moustafa, Elena Sitnikova. "Identification of malicious activities in industrial internet of things based on deep learning models", Journal of Information Security and Applications, 2018.

10. T. Kalsoom, N. Ramzan, S. Ahmed, and M. Ur-Rehman, “Advances in sensor technologies in the era of smart factory and industry 4.0,” *Sensors*, vol. 20, p. 6783, 2020.

11. A. Husain, a. Salem, c. Jim, and g. Dimitoglou, “development of an efficient network intrusion detection model using extreme gradient boosting (xgboost) on the unsw-nb15 dataset,” in *proceedings of the 2019 ieee international symposium on signal processing and information technology (isspit)*, pp. 1–7, ajman, uae, december 2019.

12. Subba Rao Polamuri, Dr. Kudipudi Srinivas, Dr. A. Krishna Mohan, Multi-Model Generative Adversarial Network Hybrid Prediction Algorithm (MMGAN-HPA) for stock market prices prediction, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 9, 2022, Pages 7433-7444,https://doi.org/10.1016/j.jksuci.2021.07.00113. https://www.kaggle.com/siddarthml1698/ddos-botnet-attack-on-iot-devices

14. https://www.kaggle.com/mrwellsdavid/unsw-nb15

15. Abomhara, M.; Køien, G.M. Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks. *J. Cyber Secur. Mobil.* **2015**, *4*, 65–88.