International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Study on Supervised Anomaly Detection Model For MQTT-Based lot Data for Dos Attacks

Mrs. Bhagyashri Hemant Katole[1], Dr. Tanuja R. Pattanshetti[2]

[1]Research Scholar, Department of Computer Engineering & Information Technology , COEP Technological University, Pune, India
[2]Assistant Professor, Department of Computer Engineering & Information Technology, COEP Technological University, Pune, India

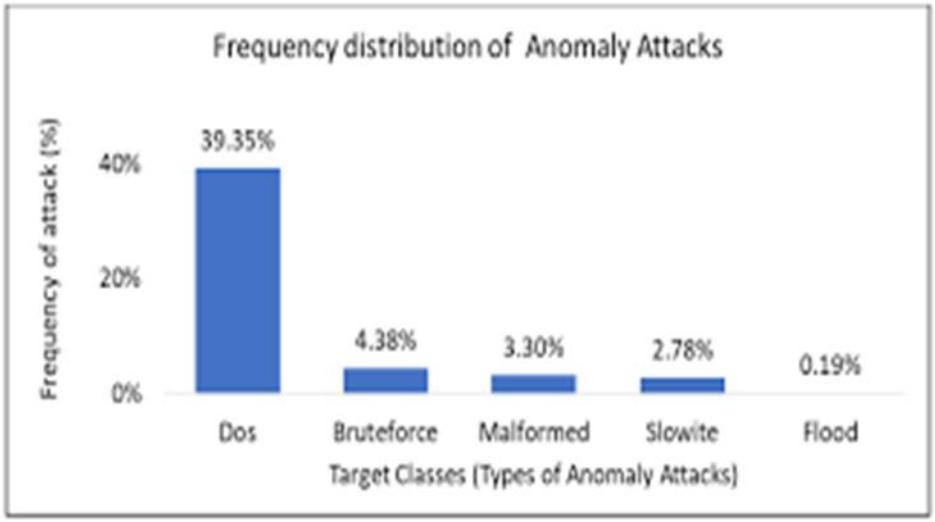
Abstract
This paper introduces a methodology for a generalized anomaly detection model for DoS attacks using supervised ML algorithms. This involves different MQTT-based IoT datasets using different MQTT brokers. Anomaly detection is identifying data points, events, or observations that deviate significantly from the expected pattern in a dataset. In IoT, anomaly detection monitors the health, performance, and security of devices and systems. It helps detect issues such as equipment malfunctions, security breaches, and inefficiencies, allowing for timely interventions and reducing the risk of major failures. Some of the anomalies are drop in signal strength, detection of unusual gateways and receiving messages without a data packet. Anomaly detection in MQTT-based IoT systems involves identifying unusual patterns or behaviours in the data being transmitted by IoT devices. MQTT is a lightweight messaging protocol used for sending data between IoT devices and servers. Data is usually transmitted in JSON or other lightweight formats. Examples for anomaly attacks are DoS, Brute-Force, Malformed, flood etc. The approach in paper will focus on generalized anomaly detection model for DoS attack.

Keywords: Internet of Things (IoT), MQTT (Message Queuing Telemetry Transport), Denial-of-service (DoS), Machine learning (ML)

1. INTRODUCTION

In the IoT, ensuring the reliability, efficiency, and security of connected devices is critical. As IoT devices generate massive amounts of data, detecting anomalies becomes increasingly important. Anomaly detection helps to identify potential issues before they escalate, providing businesses with valuable insights and the ability to improve operational efficiency if used correctly. DoS attack is a cyberattack that makes a device or computer unavailable to its intended users. It is having maximum percentage in frequency distribution of anomaly attack. The attacker usually achieves this by flooding the target with requests until it can no longer process normal traffic. Supervised learning is a type of ML where algorithms are trained on a labelled dataset. In supervised learning, we will train the model with data that includes both normal and anomalous examples.

Figure 1: Frequency distribution of anomaly attacks



MQTT is an OASIS standard messaging protocol for the IoT. It is designed as an extremely lightweight publish/subscribe messaging transport that is ideal for connecting remote devices with a small code footprint and minimal network bandwidth. A publish-subscribe messaging protocol designed for constrained devices communicating with low bandwidth (small overhead) over unreliable networks (e.g., high latency). Anomaly detection is the process of identifying data points, observations that deviate significantly from the expected pattern in a dataset. Some of available IoT Datasets along with MQTT broker

Table 1: MQTT brokers used in various IoT datasets

MQTT based IoT dataset	MQTT broker
MQTT-IoT-IDS2020	Mosquitto, HiveMQ, Eclipse Paho
MQTTset	Eclipse Mosquitto
MQTT-IoT	Mosquitto
BoT-IoT	Mosquitto, HiveMQ
DDoS-MQTT-I	Mosquitto, HiveMQ
SENMQTT-SET	Mosquitto

A ML model is a program that identify patterns or draw conclusions based on previously unknown data. A ML supervised algorithm need to be selected. Some ML classifiers are decision trees (DT), k-nearest neighbours (K-NN), kernel-support vector machines (k-SVM), logistic regression (LR), naïve Bayes (NB), random forest (RF), extreme gradient boosting (XGBoost), and artificial neural network (ANN). The analysis can be done using confusion matrix and accuracy. Some of considerations that we need to explore

- Data Imbalance: Anomalies are often rare, which can lead to imbalanced datasets. Techniques like oversampling, under sampling, or anomaly score adjustments can help.
- Scalability: Ensure that the system can handle the volume and velocity of incoming IoT data.
- Real-Time Constraints: Anomaly detection should be efficient enough to meet real-time processing requirements.
- Interpretability: Ensure that the model's predictions are understandable and actionable, especially in critical systems where understanding the cause of anomalies is important.

2. Literature review

Ongoing research on anomaly detection for the IoT is a rapidly expanding field. This growth necessitates an examination of application trends and current gaps. The vast majority of those publications are in areas such as network and infrastructure security, sensor monitoring, smart home, and smart city applications and are extending into even more sectors. Recent advancements in the field have increased the necessity to study the many IoT anomaly detection applications. Recent research has seen notable progress in the field of anomaly detection through the use of machine learning algorithms. A potential methodology involves the use of generative models and neural networks to convert the unsupervised task of anomaly detection into a supervised task. The most recent anomaly detection technique is Double Adversarial Activation Anomaly Detection (DA3D), which creates synthetic anomalies for training using adversarial autoencoders. This methodology generates synthetic anomalies by leveraging average data and outperforms existing cutting-edge techniques solely through data-driven means. In the authors' discussion of DA3D, adversarial autoencoders are used to create anomalous counterexamples based solely on normal data, enabling the identification of actual but unnoticed abnormalities. The method used includes adaptive auto-encoders and double adventuring activation anomaly detection. Supervised machine learning methods are currently employed as the predominant approach for anomaly detection in high-performance computing systems. The authors have presented a methodology for detecting anomalies in high-performance computing systems using ML techniques. The proposed methodology employs autoencoders as a means of detecting anomalies. Future objectives include conducting tests on a more extensive range of anomalies and integrating the findings into a functional prototype capable of real-time operation. The accuracy values show a range of 88% to 96%. Novel anomalies that have not been previously observed are identified. This article makes two significant contributions: first, it achieves a high level of accuracy (ranging from 88% to 96%) in detecting anomalies with precision; second, it identifies novel categories of anomalies. Another contribution that used autoencoder in its methodology suggests the use of federated learning within a neural network autoencoder framework to detect anomalies in financial transactions. This study explores the implementation of federated learning within the neural network autoencoder model for anomaly detection in a provided data stream. Federated learning facilitates the implementation of robust and efficient anomaly detection mechanisms while ensuring data security. The application of federated learning in the context of anomaly detection aims to mitigate online theft and scams through the use of ML techniques. In a supervised machine learning technique was used for anomaly detection. The study proposes a model that incorporates Naive Bayes and SVM algorithms for anomaly detection. In addition, an ensemble approach is suggested to address limitations and improve the accuracy of anomaly detection results. The findings indicate that the Naive Bayes classifier demonstrates favourable results. The ensemble approach shows promising results in the field of detection. Another assemble method was used. The study presents a novel approach to anomaly detection in multidimensional datasets containing both numerical and categorical features. The proposed method, known as Out-of-Bag anomaly detection, aims to improve the accuracy and reliability of ML systems by incorporating it as a data pre-processing step. To evaluate its effectiveness, a case study on home valuation is conducted. The method proposed in this study demonstrates exceptional performance on widely recognised benchmark datasets, surpassing existing approaches in the field. The proposed model enhances the precision and dependability of the machine learning system.

Table 3: Challenges described in anomaly detection for MQTT-based IoT data

Sr no	Title of the paper	Link	Summary	Challenges
1	SENMQTT-SET: An Intelligent Intrusion Detection in IoT-MQTT Networks Using Ensemble Multi Cascade Features	https://ieeexplore.ieee.org/document/9739734	The SENMQTT-SET dataset has been generated and analyzed for MQTT attack detection in IoT contexts. The intrusion detection testbed includes three scenarios: no attack, attack on a subscriber, and attack on a broker, which has been designed to record regular traffic and attack characteristics	The multi-context feature gene-ration from the raw dataset using algorithm [3]
2	Anomaly detection system for data	https://www.sciencedirect.com/science/article/pii/S2542660524000374	An analysis of the characteristics of the IoT	The model aims to perform

	quality assurance in IoT infrastructures based on machine learning		data packets has also been carried out choosing the parameters to be considered for anomaly analysis after the correlation study. The model aims to perform detections i.e to detect packets that are not correct, remove them from data injection flow, analyze them & extract the anomalies	broad spectrum detections that is to detect packets that are not correct, remove them from data ingestion in the applications that use IoT and then analyse them and extract the anomaly [26]
3	A Novel Network Intrusion Detection System Based on Semi-Supervised Approach for IoT	https://thesai.org/Downloads/Volume14No4/Paper_24-A_Novel_Network_Intrusion_Detection_System.pdf	MQTT-driven IoT systems that combine unsupervised and supervised learning techniques. This information creates supervised	Most of the existing IDS schemes are specific to a particular scenario and may not be adaptive

			multiple test cases on the IoT dataset MQTTIOT2020 that are conducted to demonstrate the potential of the proposed model	to detect unknown attacks and multi-class Intrusions [31]
4	Denial of service attack detection through ML for the IoT	https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1767484	A DoS attack detection framework for MQTT attack detection in IoT environment was proposed and evaluated. The attack detection testbed was designed to capture normal and attack traffic and count based statistical flow features	Features based on the TCP protocol analysis that do not provide sufficient information on the MQTT protocol parameters
5	Machine Learning Approaches for Anomaly	https://link.springer.com/article/10.1007/s11277-021-08994-z	The vast majority of those publications are in areas	The need of new perspectives for univariate

	Detection in IoT: An Overview and Future Research Directions		such as network and infrastructure security, sensor monitoring, smart home, and smart city applications and are extending into even more sectors. Recent advancements in the field have increased the necessity to study the many IoT anomaly detection applications	, multivariate, as well as for dealing with high-dimensional data. Anomaly detection tasks are separated from the sensor and sink nodes and assigned to distinct nodes in the network [2]
6	A Hybrid Approach for the Detection and Classification of MQTT-based IoT-Malware	https://ieeexplore.ieee.org/document/1010	A hybrid approach is applied to integrates two machine learning algorithms, namely multi-layer perceptron (MLPC) and K-Nearest Neighbour (KNN) for the detection of IoT	Anticipation of the attack to address the security risk of the MQTT protocol based on ML methods

			malware based on MQTT protocol communication, The results generated by the proposed model is 91% and 93% respectively	

3. Research gaps

Here are some potential research gaps in the field of supervised anomaly detection models for MQTT-based IoT data targeting DoS attacks:

- Limited Scalability of Supervised Models [3]
- Dataset Challenges - The lack of publicly available, comprehensive datasets that include labeled DoS attack instances specific to MQTT protocols
- Feature Selection and Engineering - Insufficient exploration of MQTT-specific features that can effectively differentiate between normal and attack behaviors. Dynamic feature selection methods tailored for MQTT traffic patterns are underdeveloped
- Model Generalization- Models often face difficulties in generalizing to different IoT environments or unseen types of DoS attacks, leading to poor performance in cross-domain applications [33]
- Scarlet of labelled data – Insufficient exploration of MQTT-specific features that can effectively differentiate between normal and attack behaviors. It provides labelled IoT overcoming the issue of unlabelled data [23]
- The nurture of data in IoT is rapid fast and the data is formed as rows without classes. The unlabeled data in IoT represents a significant challenge, due to the cost of manual labelling, which requires time and experts
- IoT benchmarking - The term ‘benchmarking’ is used in machine learning to evaluate and compare different solutions based on different criteria, such as performance or computational time, with publicly available datasets [24]

4. Proposed methodology

The possible DoS attack model involves attacks that can overwhelm server resources and to deny access by legitimate clients. According to Little's Law (Little & Graves, Citation2008), the average number of items in a queuing system can be defined as: [3]

$L=\lambda*W(1)$

where λ is the arrival rate of items into the system

W is the average time spent by an item in the system

DoS attacks aim to fill-up the system queue, thus denying service to legitimate clients. These attacks can either increase arrival rate of packets or increase the per-packet processing time by forcing complex computing operations at the victim device. An attacker without valid credentials can only vary the parameters of a CONNECT packet as clients cannot publish or subscribe without successful connection to the broker. However, after a successful connection using valid credentials but without valid authorization to publish and subscribe to topics in MQTT, the attacker can vary PUBLISH or SUBSCRIBE control packet parameters. Various attack metrics were measured to assess the impact of DoS attacks against the MQTT brokers. Possible DoS attack metrices may include CPU utilization, bandwidth and memory utilization

- Process CPU (pCPU): Measured using bash script fetching the CPU utilization associated with broker process ID using top Linux command
- Bandwidth: Total bandwidth consumed during the attack (kbytes)
- Memory: Percentage of memory consumed during the attack

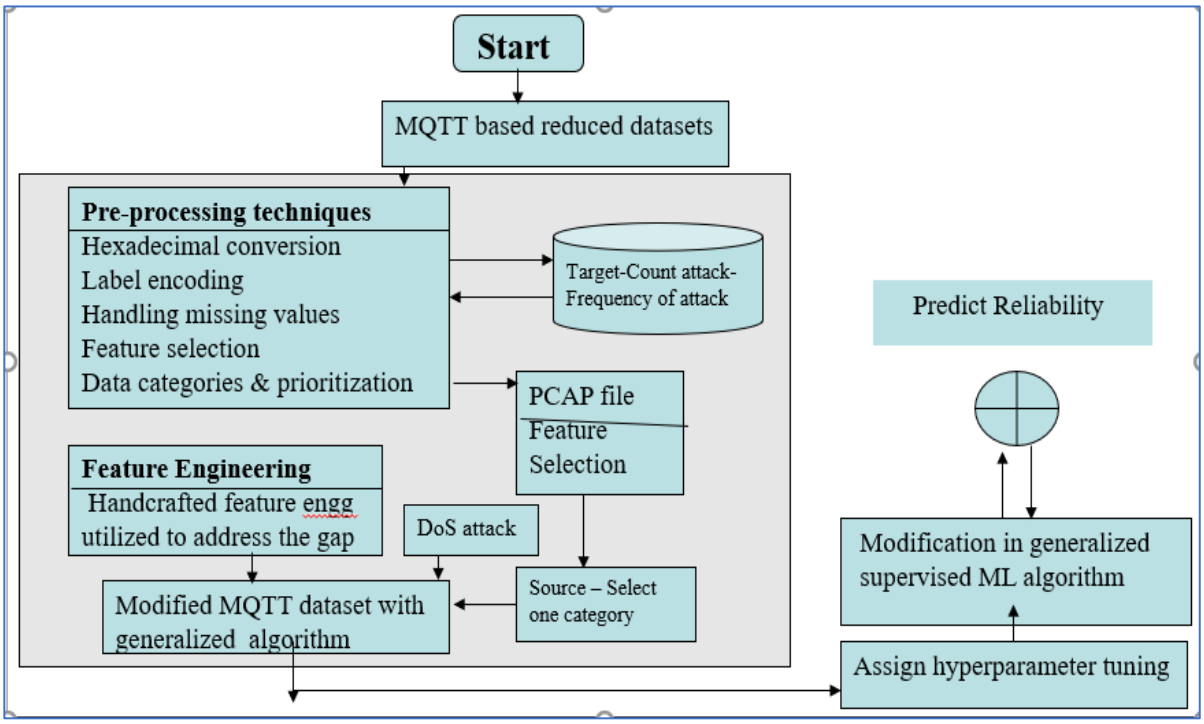
The implementation steps involved in proposed methodology are as follows

- For supervised anomaly detection labeled data will be used. This would involve manual labeling or using domain knowledge to identify constituents. This labelled data includes both TCP and MQTT traffic data
- Preprocess the data and feature extraction – Extract relevant features from MQTT messages. This might involve time-series data, categorical variables or other type of feature. [Tools like wireshark , tcpdump can be used]
- Normalization - The min-max normalization, also known as feature scaling techniques can be used to linearly transform data to a range of 0 to 1.
- Data splitting – Split your data into training and test sets
- Investigate ML classifier suitable for MQTT-based IoT data
- Modelling of DoS attacks for MQTT brokers
- Evaluate the model – Evaluate the model using matrices like accuracy, precision, recall, F1 score and confusion matrix
- Deploy the model and monitor the status of the model

The supervised machine learning algorithms offer effective means of detecting, identifying, and classifying anomalies [16]. The benefit of this approach is the capability of learning from data that is freely and widely available in the environment of the IoT, though class labels (normal or anomalous) for IoT data are rarely available. Thus, it is a challenge to construct an anomaly detection model that is wholly reliant volume of IoT data that is expanding at a rapid rate and the model must be able to predict anomalies.

s

Figure 2 – Flowchart of anomaly detection model for DoS attacks



5. Conclusion

This paper proposed a methodology of generalized anomaly detection model for DoS attacks specific to MQTT-based IoT data. Various supervised ML algorithms need to be explored and studied to provide a generalised ML model for DoS attack. These supervised ML algorithms need to be analyzed and compared.[31] The new future research directions have been proposed and discussed from the perspective of the different MQTT-based IoT datasets. Overall, the proposed approach can be a valuable addition to the existing DoS detection techniques, providing a more efficient and effective solution to combat DoS attacks.

References

1. Luis M. Camarinha-Matos & Srinivas Katkoori , Challenges in IoT Applications and Research. link.springer.com/chapter/10.1007/978-3-030-96466-5_1, IFIPIoT 2021
2. Praveen Kumar Donta a c, Satish Narayana Srirama b c, Tarachand Amgoth a, Chandra Sekhara Rao Annavarapu, a Survey on recent advances in IoT application layer protocols and machine learning scope for research directions https://www.researchgate.net/publication/355268771_Survey_on_recent_advances_in_IoT_application_layer_protocols_and_machine_learning_scope_for_research_directions, 2022
3. Hariprasad siddharthan, Deepa T., Prabhu chandhar, SENMQTT-SET: An Intelligent Intrusion Detection in IoT-MQTT Networks Using Ensemble Multi Cascade Features, https://ieeexplore.ieee.org/document/9739734, 2022
4. Jeddou Sidna, Baina Amine, Najid Abdallah, Hassan El , Analysis and evaluation of communication Protocols for IoT Applications https://dl.acm.org/doi/10.1145/3419604.3419754,2020
5. José A. Barriga, Pedro J. Clemente, Miguel A. Pérez-Toledano, Elena Jurado-Málaga, Juan HernándezDesign, code generation and simulation of IoT environments with mobility devices by using

model-driven development:SimulateIoT-Mobile https://www.sciencedirect.com/science/article/pii/S1574119223000093?via%3Dihub
6. Ángel Luis Muñoz Castañeda, José Antonio Aveleira Mata & Héctor Aláiz-Moretón, Characterization of threats in IoT from an MQTT protocol-oriented dataset , https://link.springer.com/article/10.1007/s40747-023-01000-y.,2023
7. Mustafa Al-Fayoumi, Qasem Abu Al-Haija , Capturing low-rate DDoS attack based on MQTT protocol in software Defined-IoT environment sciencedirect.com/science/article/pii/S2590005623000413 ,2023
8. Steve Chesney; Kaushik Roy, AI Empowered Intrusion Detection for MQTT Networks ,https://ieeexplore.ieee.org/document/985612 , 2022
9. Ali Abdullah M. Alzahrani, Theyazn H. H. Aldhyani, Artificial Intelligence Algorithms for Detecting and Classifying MQTT Protocol Internet of Things Attacks,
10. https://ieeexplore.ieee.org/document/10104820, 2023
11. Ali Alzahrani, Theyazn H. H. Aldhyani , Artificial Intelligence Algorithms for Detecting and Classifying MQTT Protocol Internet of Things Attacks https://www.researchgate.net/publication/365645786_Artificial_Intelligence_Algorithms_for_Detecting_and_Classifying_MQTT_Protocol_Internet_of_Things_Attacks, 2022
12. Ibrahim Kok, Suat Ozdemir, DeepMDP: A Novel Deep-Learning-Based Missing Data Prediction Protocol for IoT https://ieeexplore.ieee.org/document/9121973 , 2020
13. Mengmeng Ge; Xiping Fu; Naeem Syed; Zubair Baig; Gideon Teo; Antonio Robles-Kelly ,Deep Learning-based Intrusion Detection for IoT Networks https://ieeexplore.ieee.org/document/8952154 , 2019
14. B. Haritha Lakshmi, M. Navyasri, M. Prasanna, M. Manasa, Machine Learning Model for Message Queuing Telemetry Transport Data Analytics https://www.researchgate.net/publication/365645786_Artificial_Intelligence_Algorithms_for_Detecting_and_Classifying_MQTT_Protocol_Internet_of_Things_Attacks , 2023
15. Fatemeh Mosaiyebzadeh, Luis Gustavo Araujo Rodriguez, Network Intrusion Detection System using Deep Learning against MQTT attacks in IOT https://ieeexplore.ieee.org/document/9647850, 2021
16. Andrii Shalaginov, Oleksandr Semeniuta, Mamoun Alazab , MEML: Resource-aware MQTT-based Machine Learning for Network Attacks Detection on IoT Edge Devices https://www.researchgate.net/publication/337699877_MEML_Resource-aware_MQTTbased_Machine_Learning_for_Network_Attacks_Detection_on_IoT_Edge_Devices ,2019
17. Andrii Shalaginov , Giovanni Chiola, Maurizio Aiello, Maurizio Mongelli, MQTTset a new dataset for machine learning techniques in MQTT, https://www.mdpi.com/1424-8220/20/22/6578 , 2020
18. José Roldán-Gómez , Javier Carrillo-Mondéjar , Juan Manuel Castelo Gómez, Sergio Ruiz-Villafranca , Analysis of the MQTT-SN Protocol for the Internet of Things, https://www.mdpi.com/2076-3417/12/21/10991, 2022
19. Alaa Alatram, Leslie F. Sikos, Mike Johnstone, Patryk Szewczyk, James Jin Kang , DoS/DDoS-MQTT- IoT: A dataset for evaluating intrusions in IoT networks using the MQTT protocol https://www.sciencedirect.com/science/article/pii/S1389128623002542 , 2023
20. Hanan Hindy, Ethan Bayne, Miroslav Bures, Robert Atkinson, Christos Tachtatzis & Xavier Bellekens , Machine Learning Based IoT Intrusion Detection System: An MQTT CaseStudy

https://www.researchgate.net/publication/348206258_Machine_Learning_Based_IoT_Intrusion_Detection_System_An_MQTT_Case_Study_MQTT-IoT-IDS2020_Dataset , 2022
21. Naveen Saran a, Nishtha Kesswani, A comparative study of supervised Machine Learning classifiers for Intrusion Detection in Internet of Things https://www.sciencedirect.com/science/article/pii/S1877050923001813, 2023
22. Megat Farez Azril Zuhairi, Improving Reliability for Detecting Anomalies in the MQTT Network by Applying Correlation Analysis for Feature Selection Using Machine Learning Techniques https://www.mdpi.com/2076-3417/13/11/6753, 2023
23. Milica Matic , Marija Antic , Istvan PAPP, Sandra IVANOVIC, Optimization of MQTT communication between microservices in the IoT cloud , https://ieeexplore.ieee.org/document/9427602, 2021
24. Nusaybah Alghanmi, Reem Alotaibi, Seyed M. Buhari , Machine Learning Approaches for Anomaly Detection in IoT: An Overview and Future Research Directions https://link.springer.com/article/10.1007/s11277-021-08994-z, 2022
25. Mukherjee, N. K. Sahu, S. Sahan , Simulation and Modeling for Anomaly Detection in IoT Network Using Machine Learning, https://link.springer.com/article/10.1007/s10776-021-00542-7 ,2022
26. Lucia Arnau Muñoz, José Vicente Berná Martínez, Francisco Maciá Pérez, Iren Lorenzo Fonseca , Anamoly detection system for data quality assurance in IoT infrastructure based on machine learning https://www.sciencedirect.com/science/article/pii/S2542660524000374, 2024
27. Ishaani Priyadarshini, Anamoly detection of iot data for smart city https://www.snap4city.org/download/video/AnomalyDetection2020.pdf , 2023
28. Daniel ZINCA, Virgil DOBROTA , DDoS attack detection using supervised machine algorithm over the CIDDOS2019 dataset https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5010134, 2023
29. I. Mukherjee, N. K. Sahu, S. Sahan , Simulation and Modeling for Anomaly Detection in IoT Network Using Machine Learning, https://link.springer.com/article/10.1007/s10776-021-00542-7 ,2022
30. Pierfrancesco Bellini; Daniele Cenni; Paolo Nesi; Mirco Soderi, Anamoly detection of iot data for smart city https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1767484 , 2020
31. Durga Bhavani A, Neha Mangla, A Novel Network Intrusion Detection System Based on Semi-Supervised Approach for IoT
32. https://thesai.org/Downloads/Volume14No4/Paper_24A_Novel_Network_Intrusion_Detection_System.pdf , 2023
33. Abdul Qadir Khan, Saad El Jaouhari, Nouredine Tamani, Lina Mroueh, Knowledge- based anomaly detection: Survey, challenges, and future directions https://www.sciencedirect.com/science/article/abs/pii/S0952197624011540?via%3Dihub , 2024
34. https://docs.google.com/presentation/d/1OWGy1D1MyokhJUCFuUAedIGohKsrOuT9/.p18_QEu2qNY1ck