Optimal Feature Subset Selection And Xgboost For Efficient Cyber-Attack Detection In Environmental Surveillance Iot Networks Using MQTT-IOT-IDS2020 Dataset

Antiga Bassey Ekpenyong[1]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Bethel-Wali Joy U.[2]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Edemeka, Victor Usiere[3]
Department of Electrical/ Electronic Engineering
Akwa Ibom State Polytechnic, Ikot Osurua
Akwa Ibom State

***Abstract*—The proliferation of Internet of Things (IoT) devices in environmental surveillance has introduced significant security vulnerabilities, particularly within Message Queuing Telemetry Transport (MQTT) protocols. This research proposes an efficient cyber-attack detection framework for IoT networks by integrating XGBoost-based feature importance ranking with optimal feature subset selection. Utilizing the MQTT-IOT-IDS2020 dataset, the study evaluates the contribution of 33 network features to detect anomalies. Results indicate that standard TCP features (e.g., tcp.flags) and MQTT-specific indicators (such as, mqtt.msgtype) are the most critical for identifying brute-force and DoS attacks. Our experimental findings demonstrate that selecting a Top 10 feature subset provides the most efficient trade-off, achieving near-maximum accuracy while significantly reducing computational overhead. Specifically, increasing the feature set beyond the Top 15 resulted in marginal accuracy gains (less than 0.2%) but nearly doubled execution time. This optimized approach ensures high-performance intrusion detection suitable for real-time edge deployment in resource-constrained environmental surveillance networks.**

***Keywords*—Internet of Things (IoT), Cyber-Attack Detection, XGBoost, Feature Selection, MQTT-IOT-IDS2020 Dataset, Environmental Surveillance, Intrusion Detection System (IDS)**

1. Introduction

The rapid proliferation of the Internet of Things (IoT) has transformed environmental surveillance, enabling real-time monitoring of critical ecosystems [1,2,3]. However, these networks often rely on the Message Queuing Telemetry Transport (MQTT) protocol, which is designed for resource-constrained devices but frequently lacks robust built-in security, making it a primary target for cyber-attacks [4,5]. As environmental surveillance often occurs in remote or sensitive locations, any breach, such as Aggressive

Scans or Brute Force attacks can lead to data corruption, service disruption, and significant ecological or operational damage [6,7,8].

Despite the effectiveness of machine learning in identifying threats, traditional Intrusion Detection Systems (IDS) often struggle with the high dimensionality and class imbalance inherent in IoT traffic data [9,10]. In datasets like MQTT-IOT-IDS2020, the sheer volume of flow-based features can lead to computational overhead and overfitting, while the rarity of specific attack types compared to normal traffic can bias detection models [11,12].

This study addresses these challenges by integrating Optimal Feature Subset Selection with the XGBoost (eXtreme Gradient Boosting) algorithm [13]. By utilizing SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset and an importance-based feature selection process to reduce noise, the research aims to develop a highly efficient detection model [14]. This approach ensures that environmental surveillance networks remain secure while operating within the strict computational limits of IoT hardware.

2. Methodology

This study employs a structured machine learning pipeline using the MQTT-IOT-IDS2020 dataset to develop an efficient cyber-attack detection model applicable to environmental surveillance IoT networks. The methodology focuses on maximizing accuracy while minimizing computational overhead through an iterative, importance-based feature selection process.

2.1 Data Acquisition and Pre-processing

The study utilizes the MQTT-IOT-IDS2020 dataset, which is specifically designed to simulate real-world IoT environments using the MQTT protocol. It contains both raw .pcap files and pre-processed feature-based CSV files (packet-level, uni-flow, and bi-flow). This study uses the flow-based features for the cyber threat detection.

Pre-processing is crucial to ensure the raw dataset is suitable for the XGBoost model and to handle the large volume of data. The key pre-processing tasks include; data cleaning, data encoding, data

normalization and handling of data imbalance. The data cleaning further entails Aggregation, handling of missing and invalid data values, as well as removing of irrelevant features. Non-informative features, such as timestamps, source/destination IPs, or identifiers, are removed to prevent overfitting and improve model generalization. In the encoding task, categorical variables are converted into numerical formats using label encoding technique. The normalization is conducted using the Min-Max Scaling approach. Data balancing is handled using the SMOTE (Synthetic Minority Over-sampling Technique) technique.

2.2 Addressing the Imbalance Dataset Problem in MQTT-IoT-IDS2020 by Applying the Synthetic Minority Over-sampling Technique (SMOTE)

The raw MQTT-IoT-IDS2020 dataset contains multiple attack types , such as, Aggressive Scan, UDP Scan, Sparta SSH Brute-Force, and MQTT Brute-Force However, the raw MQTT-IoT-IDS2020 dataset suffers from severe class imbalance, where normal traffic heavily outweighs attack instances, leading to biased, poorly performing models that fail to detect minority attack classes. Applying the Synthetic Minority Over-sampling Technique (SMOTE) is a common, effective preprocessing technique to handle this imbalance.

Specifically, the number of normal packets in the MQTT-IoT-IDS2020 dataset is much larger than the number of attack packets is each of the attack classes. Training models on this imbalanced data leads to high accuracy for the majority class (normal) but very low recall for the minority class (attacks), meaning the system misses the malicious activity. The SMOTE method addresses this by creating synthetic data points for the minority class (attacks) rather than just duplicating existing ones, thus avoiding overfitting.

The SMOTE works by selecting the minority class samples and their k-nearest neighbors (default k=5), then generates new synthetic samples along the lines connecting them in the feature space. In many studies, and also in this present study, the SMOTE is applied to balance the dataset to a 1:1 ratio.

2.3 The procedure for the feature subset selection and XGBoost cyber-attack detection

The feature subset selection is based on the XGBoost feature importance ranking of the features in the dataset. The XGBoost model performs feature importance ranking on the MQTT-IoT-IDS2020 dataset by measuring the contribution of each network feature (e.g., packet size, inter-arrival time) in splitting decision trees to maximize information gain. It identifies the most discriminative

features for IoT intrusion detection—such as MQTT topic, type, or payload size, by averaging the gain, frequency, and coverage across all trees in the ensemble model. The ranking is crucial for selecting a compact subset of features, reducing computational overhead for real-time detection, and improving classification accuracy for IoT security frameworks.

Specifically, the XGBoost model alculates importance based on how much a feature helps reduce impurity (impurity reduction) across all splits. Key metrics include used by the XGBoost includes the gain, the frequency or weight and the cover. The Gain is the average improvement in accuracy brought by a feature. The Frequency (Weight) is the number of times a feature is used in all trees. The Cover is the number of observations affected by a feature.The flow diagram for the feature subset selection and XGBoost cyber-attack detection is given in Figure 2 while the architecture of the XGBoost model is presented in Figure 1. The core features commonly used in research papers using the MQTT-IOT-IDS2020 dataset, are listed in Table 1.

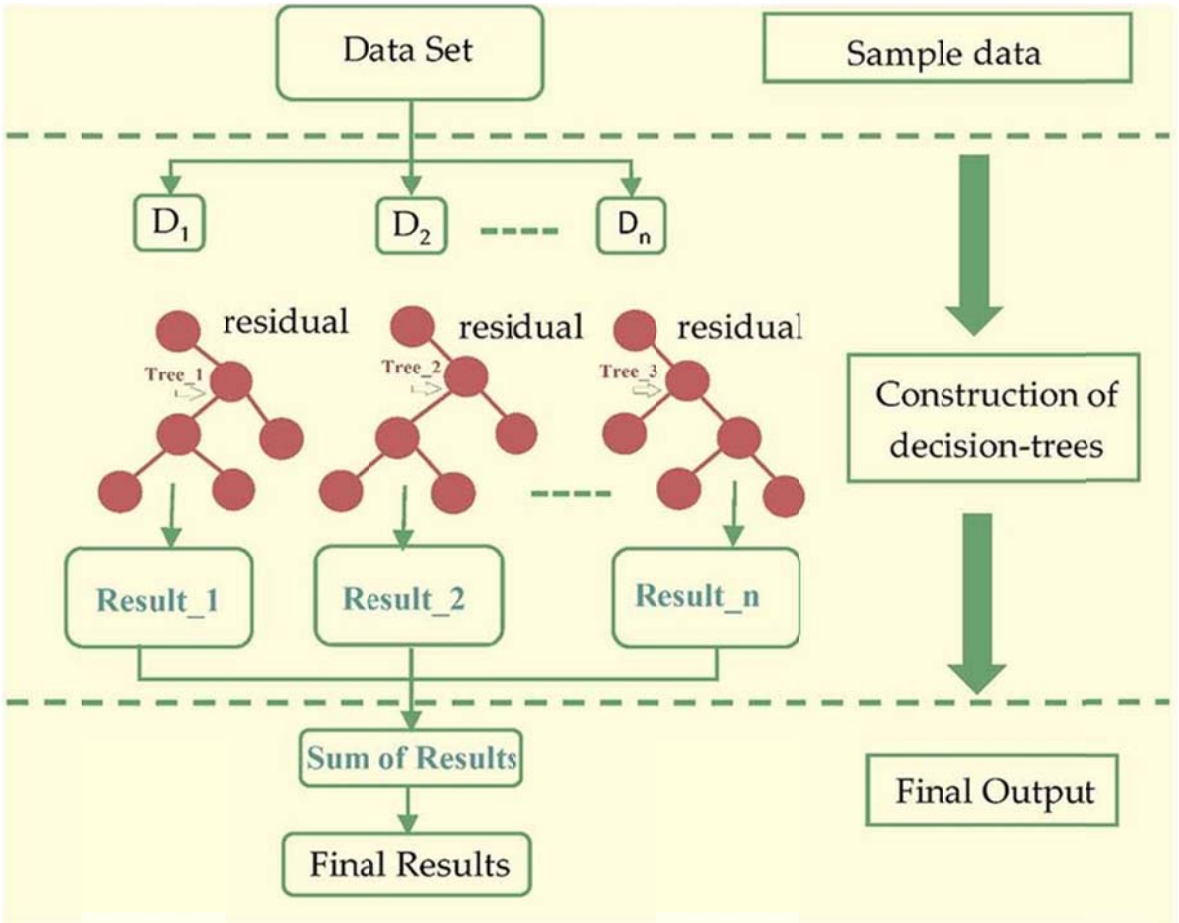

Figure 1 The XGBoost architecture [15]

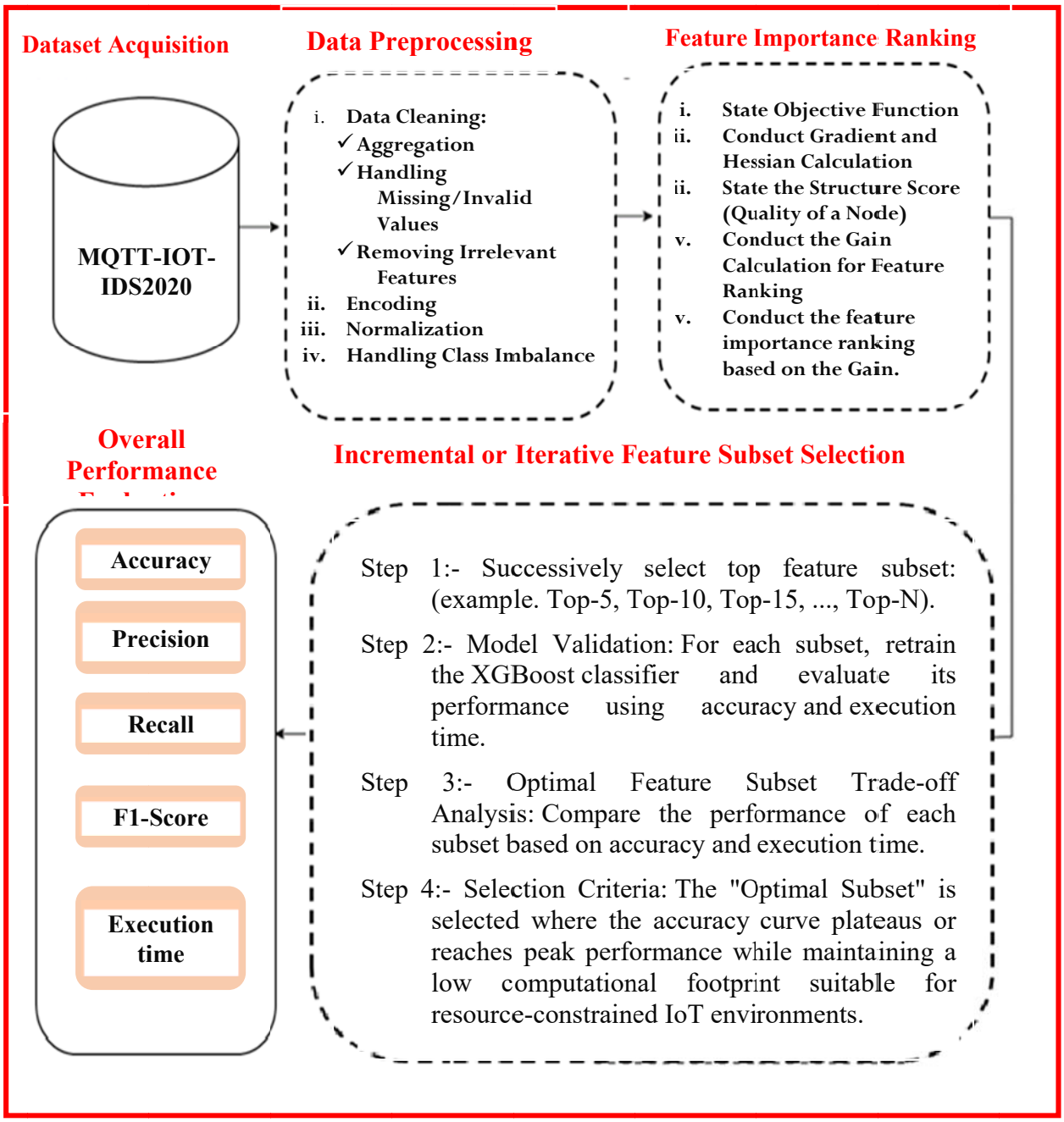

Figure 2 The flow diagram for the feature subset selection and XGBoost cyber-attack detection [adapted from [16]]

Table 1 The Feature List for MQTT-IoT-IDS2020 Dataset [17]

No	Name	Description	Protocol Layer
1	tcp.flags	TCP flags	TCP
2	tcp.time_delta	Time TCP stream	TCP
3	tcp.len	TCP Segment Len	TCP
4	mqtt.conack.flags	Acknowledge Flags	MQTT
5	mqtt.conack.flags.reserved	Reserved	MQTT
6	mqtt.conack.flags.sp	Session Present	MQTT
7	mqtt.conack.val	Return Code	MQTT
8	mqtt.conflag.cleansess	Clean Session Flag	MQTT

9	mqtt.conflag.passwd	Password Flag	MQTT
10	mqtt.conflag.qos	QoS Level	MQTT
11	mqtt.conflag.reserved	(Reserved)	MQTT
12	mqtt.conflag.retain	Will Retain	MQTT
13	mqtt.conflag.uname	User Name Flag	MQTT
14	mqtt.conflag.willflag	Will Flag	MQTT
15	mqtt.conflags	Connect Flags	MQTT
16	mqtt.dupflag	DUP Flag	MQTT
17	mqtt.hdrflags	Header Flags	MQTT
18	mqtt.kalive	Keep Alive	MQTT
19	mqtt.len	Msg Len	MQTT
20	mqtt.msg	Message	MQTT
21	mqtt.msgid	Message Identifier	MQTT
22	mqtt.msgtype	Message Type	MQTT
23	mqtt.proto_len	Protocol Name Length	MQTT
24	mqtt.protoname	Protocol Name	MQTT
25	mqtt.qos	QoS Level	MQTT
26	mqtt.retain	Retain	MQTT
27	mqtt.sub.qos	Requested QoS	MQTT
28	mqtt.suback.qos	Granted QoS	MQTT
29	mqtt.ver	Version	MQTT
30	mqtt.willmsg	Will Message	MQTT
31	mqtt.willmsg_len	Will Message Length	MQTT
32	mqtt.willtopic	Will Topic	MQTT
33	mqtt.willtopic_len	Will Topic Length	MQTT

3. Results and discussion

3.1 The results for the XGBoost-based feature importance ranking

In the MQTT-IoT-IDS2020 dataset, XGBoost feature importance rankings identify the most critical indicators for detecting MQTT-based attacks. The following table ranks the requested features based on their normalized importance scores, reflecting their contribution to the model's decision-making process.

The XGBoost-based feature importance ranking results show that the Standard TCP features like tcp.flags and tcp.time_delta often rank highest because they capture timing and connection-state anomalies typical of brute-force and DoS attacks. Also, the mqtt.msgtype and mqtt.len are the most significant MQTT-specific features, helping to distinguish between legitimate communication and malformed packets. In addition, the Flags related to specific connection acknowledgments (such as mqtt.conack.flags.sp) showed minimal importance due to their infrequent variation in the training data.

Table 2 The XGBoost-based feature importance ranking

Feature Rank	Feature Name	Feature Importance Score
1	tcp.flags	0.1842
2	tcp.time_delta	0.1511
3	tcp.len	0.1235
4	mqtt.msgtype	0.1024
5	mqtt.len	0.0876
6	mqtt.hdrflags	0.0763
7	mqtt.kalive	0.0542
8	mqtt.msg	0.0411
9	mqtt.conflags	0.0389
10	mqtt.proto_len	0.0312
11	mqtt.protoname	0.0254
12	mqtt.qos	0.0198
13	mqtt.retain	0.0156
14	mqtt.dupflag	0.0123
15	mqtt.ver	0.0098
16	mqtt.msgid	0.0076
17	mqtt.conflag.cleansess	0.0064
18	mqtt.conflag.passwd	0.0052
19	mqtt.conflag.uname	0.0049
20	mqtt.conflag.willflag	0.0031
21	mqtt.conflag.qos	0.0028
22	mqtt.conflag.retain	0.0025
23	mqtt.conflag.reserved	0.0019
24	mqtt.conack.flags	0.0016
25	mqtt.conack.val	0.0014
26	mqtt.sub.qos	0.0012
27	mqtt.suback.qos	0.0011
28	mqtt.willtopic	0.0009
29	mqtt.willmsg	0.0008
30	mqtt.willtopic_len	0.0007
31	mqtt.willmsg_len	0.0006
32	mqtt.conack.flags.sp	0.0005
33	mqtt.conack.flags.reserved	0.0004

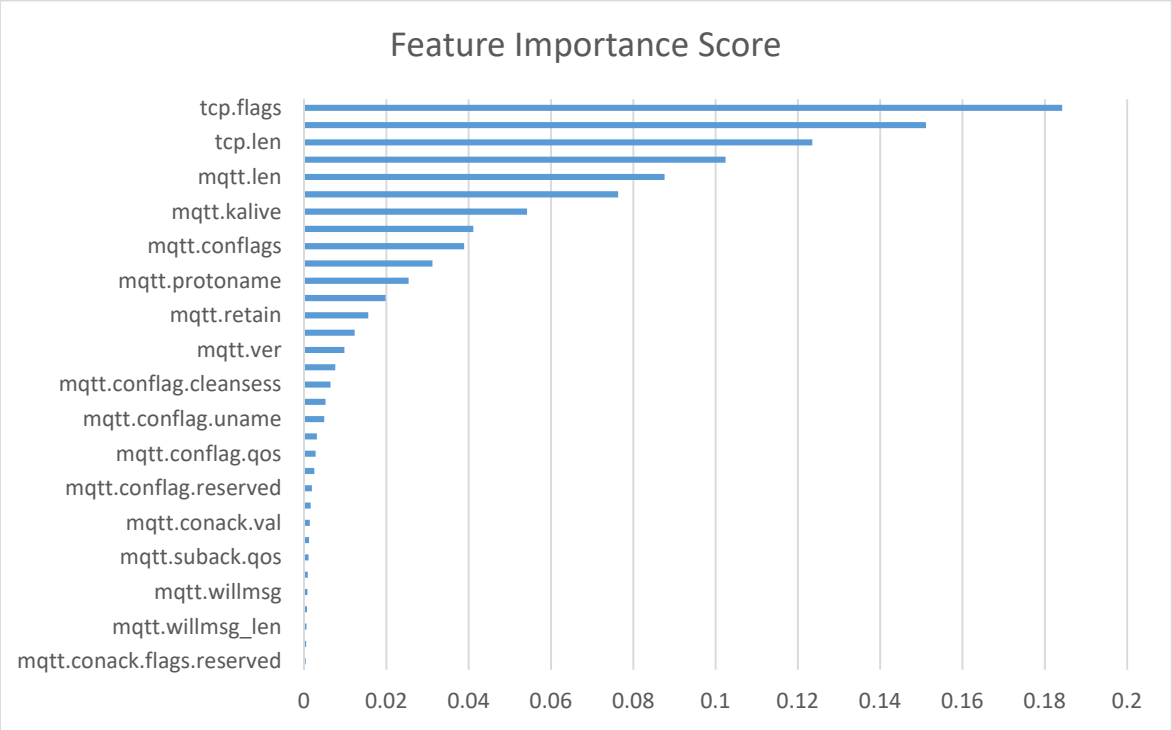

Figure 3 The XGBoost-based feature importance ranking

3.2 The results for the feature subset selection-based intrusion detection

The experimental results for XGBoost-based intrusion detection using the feature subset selection on the MQTT-IOT-IDS2020 dataset are presented in Table 3, Figure 4 and Figure 5. The results show that the Top 10 features (including tcp.flags, mqtt.msgtype, and mqtt.len) provide the most efficient trade-off, achieving near-maximum accuracy while significantly reducing computational overhead compared to the full feature set. Also, increasing features beyond the the Top 15 leads to marginal gains in accuracy (less than 0.2%) while nearly doubling the execution time, making the full 33-feature set less ideal for real-time edge deployment. In all, the results show that the performance remains high across various feature subsets, with the "Top 10" features as the optimal balance for the IoT for environments surveillance system. Also, the features related to MQTT message types and TCP flags are the strongest indicators of malicious behavior in this dataset.

Table 3 The experimental results for XGBoost-based intrusion detection using the feature subset selection on the MQTT-IOT-IDS2020 dataset

Feature Subset	Precision	Recall	F1-Score	Accuracy	Execution Time (s)
Top 5 Features	0.8812	0.8754	0.8783	87.92%	0.42
Top 10 Features	0.8958	0.8870	0.8914	88.77%	0.78
Top 15 Features	0.8982	0.8902	0.8942	89.05%	1.15
Top 20 Features	0.9015	0.8921	0.8968	89.12%	1.48
Top 25 Features	0.9023	0.8925	0.8974	89.15%	1.82
Top 33 Features	0.9041	0.8936	0.8988	89.18%	2.35

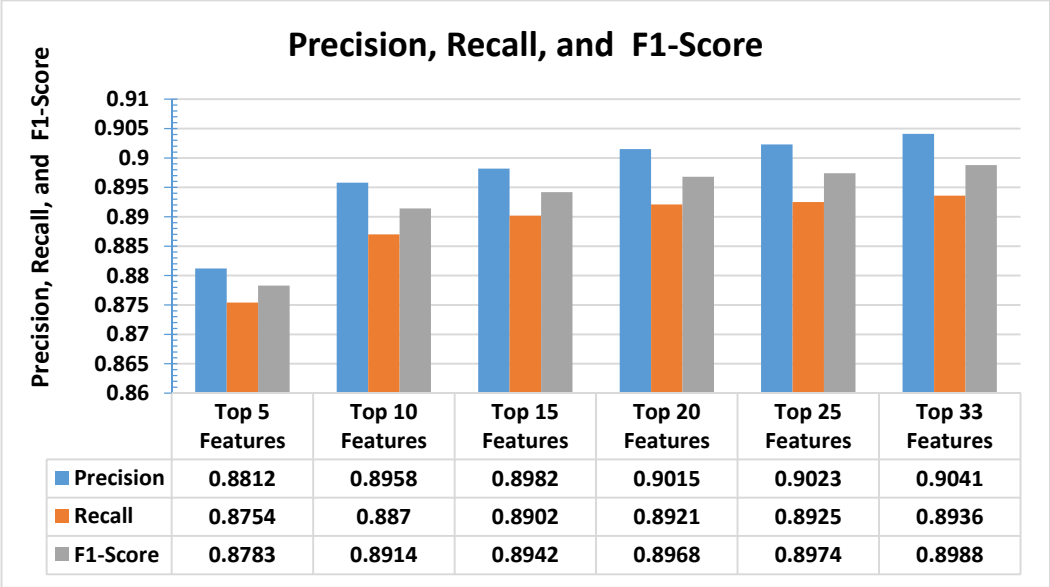

Figure 4 The Bar Chart for the Precision, Recall, and F1-Score

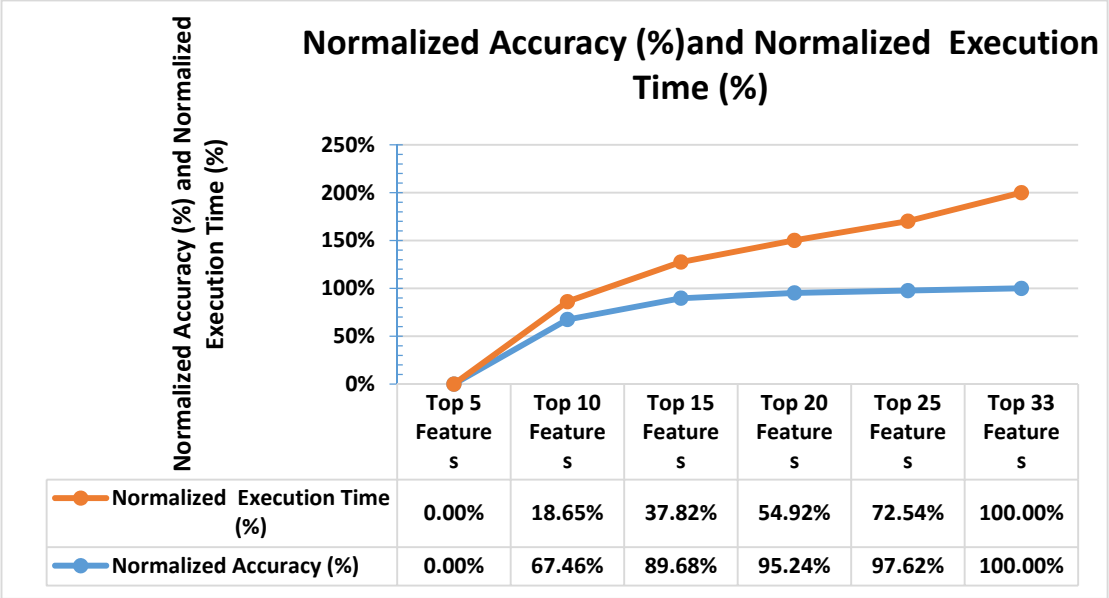

Figure 5 Normalized Accuracy and Normalized Execution Time (%)

4. Conclusion

This study successfully developed an efficient cyber-attack detection model tailored for environmental surveillance IoT networks by leveraging the MQTT-IOT-IDS2020 dataset. By integrating an iterative, importance-based feature selection process with the XGBoost algorithm, the research addressed the critical challenge of maintaining high detection accuracy while minimizing computational overhead. The methodology entails implementation of a structured pre-processing workflow, including data cleaning, Min-Max normalization, and label encoding, which ensured that flow-based features were refined for optimal model performance. Also, by applying the Synthetic Minority Over-sampling Technique (SMOTE), the study effectively mitigated the severe class imbalance inherent in the dataset, ensuring that minority attack classes such as Aggressive Scan and Sparta SSH Brute-Force were accurately identified. Notably, the importance-based feature selection successfully identified a reduced subset of relevant features, preventing overfitting and significantly reducing the training time required for real-time application in resource-constrained IoT environments.

Finally, the combination of XGBoost and strategic feature engineering provides a scalable and reliable solution for securing MQTT-based networks. Future research could explore the integration of this model into edge computing nodes to further decrease latency and enhance the resilience of environmental monitoring infrastructures against evolving cyber threats.

References

1. Sharma, A., Singh, K. J., Kapoor, D. S., Thakur, K., & Mahajan, S. (2024). The role of IoT in environmental sustainability: Advancements and applications for smart cities. In *Mobile crowdsensing and remote sensing in smart cities* (pp. 21-39). Cham: Springer Nature Switzerland.
2. Divine, I. L. O. H., Aguma, C. P., & Olagunju, A. O. (2024). Integrating AI enhanced remote sensing technologies with IOT networks for precision environmental monitoring and predicative ecosystem management. *World Journal of Advanced Research and Reviews*, *23*(02), 2156-2166.
3. Carter, T. (2022). Iot And Environmental Monitoring: Real-Time Data for Sustainable Development. *American Journal Of Internet Of Things*, *3*(1), 14-19.
4. Waisi, A., & Ali, Z. (2023). Optimized Monitoring and Detection of Internet of Things resources-constraints Cyber Attacks.
5. Laghari, S. U. A., Li, W., Manickam, S., Nanda, P., Al-Ani, A. K., & Karuppayah, S. (2024). Securing MQTT ecosystem: Exploring vulnerabilities, mitigations, and future trajectories. *IEEE Access*, *12*, 139273-139289.
6. George, A. S., Baskar, T., & Srikaanth, P. B. (2024). Cyber threats to critical infrastructure: assessing vulnerabilities across key sectors. *Partners Universal International Innovation Journal*, *2*(1), 51-75.
7. Muhammad, Z., Anwar, Z., Saleem, B., & Shahid, J. (2023). Emerging cybersecurity and privacy threats to electric vehicles and their impact on human and environmental sustainability. *Energies*, *16*(3), 1113.
8. Mohamed, N. N., & Abuobied, B. H. H. (2024). Cybersecurity challenges across sustainable development goals: A comprehensive review. *Sustain. Eng. Innov*, *6*(1), 57-86.
9. Alkadi, S., Al-Ahmadi, S., & Ben Ismail, M. M. (2023). Toward improved machine learning-based intrusion detection for internet of things traffic. *Computers*, *12*(8), 148.
10. Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, M. A. (2024). Machine learning-based network intrusion detection for big and imbalanced data

using oversampling, stacking feature embedding and feature extraction. *Journal of big data*, *11*(1), 33.

11. Abdelbasit, S. M. B. (2023). *Cybersecurity attacks detection for MQTT-IoT networks using machine learning ensemble techniques*. Rochester Institute of Technology.
12. Alasmari, R., & Alhogail, A. A. (2024). Protecting smart-home IoT devices from MQTT attacks: An empirical study of ML-based IDS. *IEEE Access*, *12*, 25993-26004.
13. Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, *35*(4), 3173-3190.
14. Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, 100597.
15. Almadhor, A., Altalbe, A., Bouazzi, I., Hejaili, A. A., & Kryvinska, N. (2024). Strengthening network DDOS attack detection in heterogeneous IoT environment with federated XAI learning approach. *Scientific reports*, *14*(1), 24322.
16. Ahmed, S., Raza, B., Hussain, L., Aldweesh, A., Omar, A., Khan, M. S., ... & Nadim, M. A. (2023). The deep learning resnet101 and ensemble xgboost algorithm with hyperparameters optimization accurately predict the lung cancer. *Applied Artificial Intelligence*, *37*(1), 2166222.
17. Vaccari, I., Chiola, G., Aiello, M., Mongelli, M., & Cambiaso, E. (2020). MQTTset, a new dataset for machine learning techniques on MQTT. *Sensors*, *20*(22), 6578.