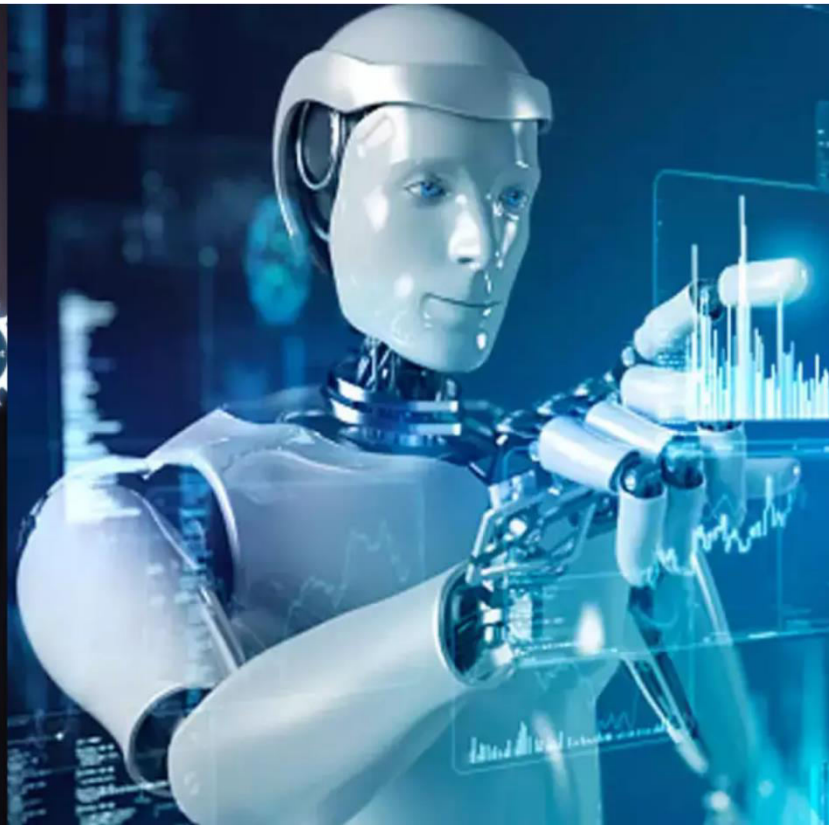
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406183|

Protecting Smart-Home IOT Devices from the MQTT Attacs: An Empirical Study of ML-Based IDS

Dr. S J R K Padminivalli V, Kalava Sneha, Randhi Nikhita Syam, Gunturu Sairam

Associate Professor, Department of CSE, R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

B. Tech Student, R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

B. Tech Student, R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

B. Tech Student, R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

ABSTRACT: Smart homes are gaining popularity globally, largely driven by Internet of Things (IoT) technologies that enable their functionality. However, due to the limited computational power and resources of IoT devices, implementing robust security measures can be challenging. As a result, intrusion detection systems (IDS) have emerged as a practical solution. In this study, we propose an optimized, high-performance model for detecting intrusions in IoT networks based on the Message Queue Telemetry Transport (MQTT) protocol, commonly used in smart home environments. Our approach involves evaluating 7 machine learning (ML) algorithms using an extended two-stage evaluation framework. This framework incorporates multiple aspects to optimize and validate performance, aiming to identify the most effective model. Through empirical analysis, the GRADIANT BOOST (GBT) classifier, combined with a smote-sampling technique, achieved the best detection performance, demonstrating accuracy and an f-score of 91.4% and enhanced IoT Network Intrusion Detection System (IDS) integrates two major innovations Time-based Analysis and IP Address Intelligence to evolve from static detection into a context-aware threat intelligence platform.. These results surpass those reported in previous studies.Additionally, we explored the impact of automatic feature engineering techniques on algorithm performance. The findings revealed that applying automatic feature engineering improved performance and reduced the time required to classify attacks by up to 67.7%. This highlights the significant role of automatic feature engineering in enhancing both the efficiency and effectiveness of intrusion detection systems in IoT-based smart home networks.

KEYWORDS: Smart homes, Internet of Things (IoT), Intrusion Detection System (IDS), Machine Learning (ML), Message Queue Telemetry Transport (MQTT), Generalized Linear Model (GLM), Random Over-Sampling, Automatic Feature Engineering, Performance Optimization, Attack Classification.

I. INTRODUCTION

Smart homes have gained significant popularity in recent years due to their convenience and ease of control. According to recent statistics, there are currently **360.72 million smart homes worldwide**, and this number is projected to increase by **86.74% by 2027** . Typically, smart home devices are connected to a central hub or "gateway," which allows users to control and monitor their smart home systems either through a unified interface or via dedicated apps for individual devices Smart homes represent a specialized branch of the **Internet of Things (IoT)** that focuses on household appliances and leverages IoT technologies to enable seamless functionality. These IoT-enabled devices include smart thermostats, lighting systems, security cameras, door locks, and kitchen appliances, all designed to enhance convenience, energy efficiency, and safety for homeowners.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406183|

II. LITERATURE REVIEW

Several studies have focused on designing Intrusion Detection Systems (IDS) tailored specifically for IoT networks. Machine Learning (ML) algorithms have been widely applied in this domain, yielding promising results. For example, Liu et al. [12] developed an ML model utilizing the Decision Tree (DT) classifier, which achieved an accuracy of 91 when trained and tested alongside three other classifiers. However, the results appear to be biased due to the limited size of the dataset, which contained only 480 records, with a mere 95 records used for testing. Such a small sample size likely contributed to the inflated accuracy metrics.In another study, the authors of introduced a three-layer IDS architecture designed to determine whether an attack occurred, identify its location, and classify the type of attack. Among the classifiers tested, the J48 algorithm demonstrated the best performance, achieving an F-measure of 88.8% for detecting attacks on unseen data.A common trend among researchers using ML for intrusion detection in IoT networks is the reliance on datasets originally generated from traditional IT networks, which typically include conventional attack types. These datasets are then applied in the context of IoT networks. However, IoT network traffic involves protocols that differ significantly from those in traditional IT networks, resulting in distinct feature sets. To address this, the authors in utilized well-known datasets such as CIDDS-001, UNSW-NB15, and NSL-KDD to protect IoT networks from Denial of Service (DoS) attacks, despite the inherent differences in network characteristics.

III. METHODOLOGY

In this study, we propose an enhanced intrusion detection model for MQTT-based IoT networks in smart homes using an extended two-stage approach to select the best classifier and sampling techniques. Uses the MQTT-IoT-IDS2020 dataset, a recent dataset published in 2020 that captures traffic from a real MQTT-based IoT network, including normal operations and four attack scenarios: aggressive scan, UDP scan, Sparta SSH brute-force, and MQTT brute-force attacks. For our model, we utilized the packet-based features dataset, which contains the most comprehensive set of features. Experiments were conducted on a Windows 10 laptop with an Intel Core i7 processor and 16GB RAM, using tools like Anaconda Navigator, Jupyter Notebook, Python, and RapidMiner Studio, while implementing the GBT algorithm via H2O with specific parameters. Data preprocessing involved removing columns like IP addresses and timestamps to ensure model independence from specific network conurations, replacing missing values, and normalizing numerical columns. The dataset was split into 80% for training and 20% for testing unseen data. To improve model performance, we employed automatic feature engineering, a novel approach in MQTT-related attack studies, which selects or builds new features automatically, reducing feature space and training time while enhancing accuracy and the system now analyzes temporal patterns in network traffic, identifying high-risk hours and adjusting alert sensitivity accordingly, which enables proactive scheduling of security measures and reduces false positives during low-risk period. The model was trained with and without this technique to evaluate its impact on performance.

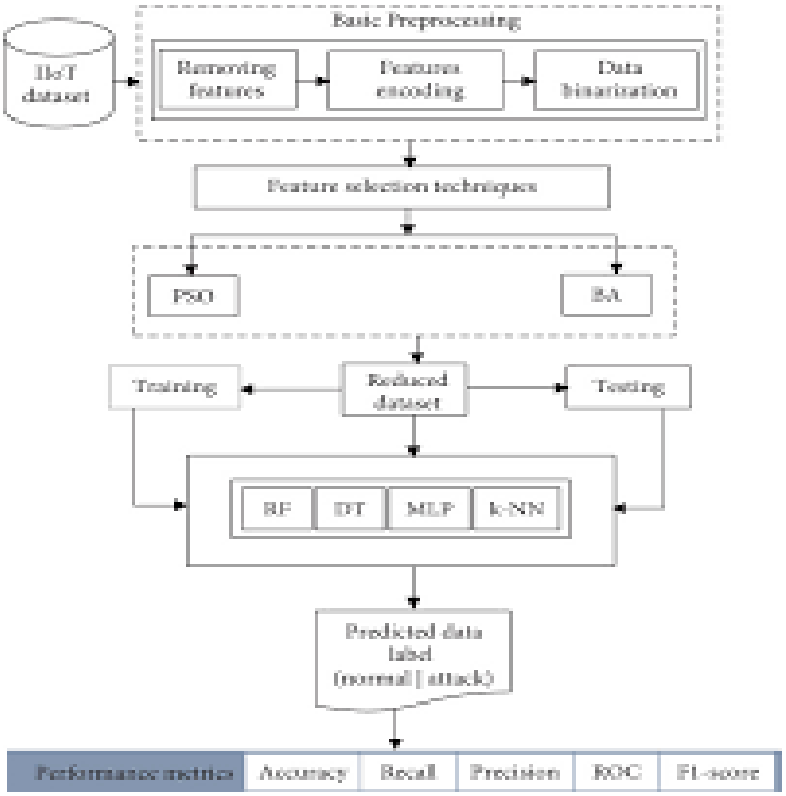
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406183|



IV. IMPLEMENTATION

classified as attacks out of the total actual attack traffic. Equation (3) shows the formula for calculating Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F-measure, or F1-score, is the harmonic mean of Precision and Recall, providing a balance between the two metrics. It is particularly useful when dealing with imbalanced datasets. Equation (4) shows the formula for calculating the F-measure:

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The False Alarm Rate (FAR), also known as the False Positive Rate (FPR), measures the percentage of normal traffic incorrectly classified as attacks. Equation (5) shows the formula for calculating FAR:

$$\text{FAR} = \frac{FP}{FP + TN}$$

To evaluate the efficiency of the models, we also measured the time spent on training and testing in milliseconds. These metrics were computed based on the confusion matrix, which organizes classification outcomes into four categories: True Positives (TP), False Negatives (FN), True Negatives (TN), and False Positives (FP). The confusion matrix serves as the foundation for deriving all performance metrics, ensuring a comprehensive assessment of each classifier's effectiveness. In our study, these evaluation metrics were calculated for each model across both stages to ensure a robust comparison of their performance. Additionally, we employed a 7-fold cross-validation technique to rigorously validate the models, averaging the results from seven iterations to produce reliable and unbiased performance estimates. This approach ensures that the selected model achieves optimal accuracy, precision, recall, and F-measure while minimizing false alarms and computational overhead.

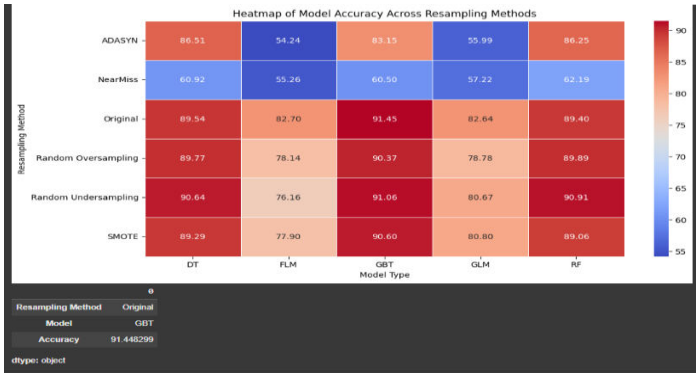
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406183|



V. RESULTS AND CONCLUSION

This study addresses the critical need to secure MQTT-based IoT networks in smart homes, which are vulnerable to cyber-attacks due to the protocol's weak security measures. Using an extended two-stage approach, we compared ML algorithms to identify the best-performing model for intrusion detection, considering factors like automatic feature engineering, hyperparameter tuning, cross-validation, and resampling techniques. In the first stage, we evaluated seven classifiers with and without automatic feature engineering, finding that it significantly improved performance, reducing training and testing times while increasing accuracy by up to 38.9% (e.g., GLM accuracy rose from 71.5% to 82%) and precision for RF from 79.5% to 89%. The FLM, RF, and GBT algorithms showed the best results, achieving near-perfect scores across metrics . In the second stage, these top classifiers were combined with five resampling techniques to address data imbalance, revealing that GBT with random over-sampling achieved the highest performance, scoring 91.4% across all evaluation metrics with a low FAR of 4.26% and efficient testing time. However, care was taken to interpret results where high precision and low recall indicated potential over-classification of positive cases, as seen with RF and DT using random over-sampling. Analysis of confusion matrices highlighted that GBT correctly classified most attack types but faced challenges distinguishing MQTT brute-force attacks from legitimate traffic due to overlapping commands. This study demonstrates the effectiveness of GBT-SMOTE resampling for detecting intrusions in MQTT-based networks, achieving optimal performance on the MQTT-IoT-IDS2020 dataset. Additionally, the extracted features from automatic feature engineering can streamline IDS development by focusing on the most relevant traffic attributes. Future work will expand evaluations to additional datasets, incorporate advanced ML/DL algorithms (e.g., CNN, RNN, LightGBM), explore novel resampling and feature engineering methods (e.g., ADASYN, DFS), and apply multiclass classification to identify specific attack types, further enhancing IoT network security. Furthermore, the study's findings underscore the importance of addressing data imbalance and leveraging advanced techniques like automatic feature engineering to improve model accuracy and efficiency. By demonstrating the superiority of GBT with random over-sampling, this research provides a robust framework for developing intrusion detection systems tailored to IoT environments. However, challenges remain in distinguishing certain attack types, such as MQTT brute-force attacks, due to their reliance on legitimate protocol commands like publish and subscribe, which complicates differentiation from normal traffic. The enhanced IoT Network Intrusion Detection System (IDS) integrates two major innovations Time-based Analysis and IP Address Intelligence to evolve from static detection into a context-aware threat intelligence platform. The system now analyzes temporal patterns in network traffic, identifying high-risk hours and adjusting alert sensitivity accordingly, which enables proactive scheduling of security measures and reduces false positives during low-risk periods. Simultaneously, it builds a dynamic IP reputation system, classifying IP addresses by behavior, network class, and traffic volume, and automatically blacklisting high-risk sources for integration with firewalls. These combined capabilities allow the IDS to assign contextual risk scores by evaluating both the time of attack and source reputation, significantly boosting detection accuracy. By leveraging historical data for predictive defense, optimizing resource allocation, and enabling real-time automated responses, the system becomes not just reactive but strategically intelligent, making it a powerful asset in modern cybersecurity operations.To overcome these limitations, future work will focus on integrating more sophisticated algorithms, such as K-nearest neighbor (KNN), light gradient-boosting machine (LightGBM), convolutional neural networks (CNN), and recurrent neural networks (RNN), alongside novel optimization methods like Elephant Herding Optimization (EHO) and Slime Mold Algorithm (SMA). Additionally, adaptive resampling techniques like Adaptive Synthetic (ADASYN) and One-Sided Selection, as well as advanced feature engineering approaches such as Deep Feature Synthesis (DFS) and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406183|

SULOV with Recursive XGBoost, will be explored to further enhance detection capabilities. Multiclass classification will also be implemented to not only detect but also classify attack types, enabling more granular threat analysis. These advancements aim to create a comprehensive, scalable, and highly accurate IDS capable of addressing the evolving security demands of smart home IoT networks, ultimately safeguarding user privacy and ensuring resilient cybersecurity infrastructure.

VI. LIMITATIONS AND FUTURE WORK

The proposed intrusion detection system (IDS) for MQTT-based IoT networks demonstrates significant advancements in detecting cyber-attacks within smart home environments. However, several limitations highlight opportunities for future improvements. Firstly, the system's reliance on the MQTT-IoT-IDS2020 dataset restricts its ability to generalize across diverse attack types and network configurations, as this dataset primarily focuses on four specific attack scenarios. Future work will incorporate additional datasets with varied attack types, such as zero-day exploits, ransomware, and advanced persistent threats (APTs), to enhance model robustness.Secondly, while automatic feature engineering improved performance, the current approach may overlook latent features that could further refine attack detection. Integrating advanced feature selection techniques like Deep Feature Synthesis (DFS) or Recursive XGBoost could uncover more nuanced patterns in network traffic. Additionally, the system faces challenges in distinguishing certain attacks, such as MQTT brute-force attempts, due to their reliance on legitimate protocol commands like publish and subscribe. Incorporating contextual metadata, such as device behavior analytics or time-series analysis, could help differentiate malicious activities from normal operations.Another limitation is the absence of real-time integration with IoT network infrastructure, which delays actionable insights. Future enhancements will focus on integrating the IDS with cloud-based platforms, enabling remote monitoring and alert notifications via APIs or dashboards. Similarly, extending the system to support multi-class classification will allow it to identify and categorize attack types, providing granular insights into threat severity and progression.The computational efficiency of the system can also be improved by exploring lightweight deep learning models, such as MobileNet or TinyML, optimized for resource-constrained IoT devices. Furthermore, incorporating adaptive resampling techniques like Adaptive Synthetic (ADASYN) or One-Sided Selection could address data imbalance more effectively than traditional methods. Novel optimization algorithms, such as Elephant Herding Optimization (EHO) or Slime Mold Algorithm (SMA), could also be explored to fine-tune hyperparameters and improve model accuracy.Finally, the system's deployment in real-world settings necessitates enhanced security measures, including biometric authentication for user verification and blockchain-based mechanisms to ensure tamper-proof logging of detected threats. Multilingual report generation, voice-based feedback, and guided instructions for non-technical users will further increase accessibility. By addressing these limitations and pursuing these future directions, the proposed IDS can evolve into a comprehensive, scalable, and intelligent solution capable of safeguarding modern IoT ecosystems across diverse environments.

REFERENCES

1.B.Thormundsson. (Nov. 2022). Smart Home Security Market Value Worldwide2018–2026.[Online].Available:https://www.statista.com/statistics/1056057/worldwide- smart-home-security-market-value/

2. F. Arat and S. Akleylek, ''Attack path detection for IIoT enabled cyber physical systems: Revisited,'' Comput. Secur., vol. 128, May 2023, Art. no. 103174, doi:10.1016/j.cose.2023.103174.

3. F. Arat and S. Akleylek, ''A new method for vulnerability and risk assessment of IoT,'' Comput.Netw.,vol. 237, Dec. 2023, Art. no. 110046, doi:10.1016/j.comnet.2023.110046.

4.A.J.Hintaw,S.Manickam, M. F. Aboalmaaly, and S. Karuppayah,''MQTT vulnerabilities, attack vectors and solutions in the Internet of Things (IoT),'' IETE J. Res., vol.69, no. 6, pp. 3368–3397,Aug. 2023.

5.A. Verma and V. Ranga, ''Machine learning based intrusion detection systems for IoT applications,'' Wireless Pers. Commun., vol. 111, no. 4,pp. 2287–2310, Apr. 2020.

6. E.Anthi,L.Williams, M. Slowinska, G. Theodorakopoulos, and P. Burnap,''A supervised intrusion detection system for smart home IoT devices,''IEEE Internet Things J., vol. 6, no. 5,pp. 9042–9053, Oct. 2019.

7. I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso,''MQTTset, a new dataset for machine learning techniques on MQTT,''Sensors, vol. 20, no. 22, p. 6578, Nov. 2020.

8. S. Ullah, J. Ahmad, M. A. Khan, M. S. Alshehri, W. Boulila, A. Koubaa,S. U. Jan, and M. M. Iqbal Ch, ''TNN-IDS: Transformer neural network-based intrusion detection system for MQTT-enabled IoT networks,'' Comput. Netw., vol. 237, Dec. 2023, Art. no. 110072, doi:10.1016/j.comnet.2023.110072.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details