National College of Ireland

Building an Integrated IoT Security Framework for Smart Homes

MSc Research Project
MSc IN CYBERSECURITY

SANDRA RAVI
Student ID: 23178302

School of Computing
National College of Ireland

Supervisor: Niall Heffernan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	SANDRA RAVI
Student ID:	23178302
Programme:	MSC IN CYBERSECURITY **Year:** 2024
Module:	MSc Research Project
Lecturer:	Niall Heffernan
Submission Due Date:	12/12/2024
Project Title:	BUILDING AN INTEGRATED IoT SECURITY FRAMEWORK FOR SMART HOMES
Word Count:	7125 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	SANDRA RAVI
Date:	12/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Building an Integrated IoT Security Framework for Smart Homes

SANDRA RAVI
23178302

Abstract

The analysis in this paper provides a new learning-based IDS that will support the security of smart homes with IoT technology. To execute the model, with the help of BoTNeTIoT_L01 and RT_IOT2022 datasets, the proposed approach presumed EDA where in it focuses its analysis on anomalies and imbalance of class labels of the target variable. To solve this problem, the application of SMOTE (Synthetic Minority Oversampling Technique) was implemented to balance the two classes hence allowing training to proceed fairly. I implemented six machine learning algorithms such as Decision Tree, Random Forest, Logistic Regression, LightGBM, XGBoost, and Multilayer Perceptron and used accuracy, precision, recall, F1 score, and ROC AUC to measure model performance. The best models were serialized using the joblib tool to fit into smart home systems to detect real-time intrusions. This pipeline provides an effective, expandable and realistic approach to developing solutions for emerging IoT security problems, translating research ideas and concepts into practice, while moving the state of the art of intrusion detection in smart homes forward.

Keywords: Smart Homes, Machine learning, Real Time Application, Streamlit, Data Analysis.

1. Introduction

Smart homes or households that are filled with automated, interconnected systems that also allow remote control and monitoring, have become very popular due to advancement in technology. Smart homes, by using the Internet of Things (IoT), households' routine activities such as lighting control, temperature regulation, security systems, appliances scheduling can be managed and controlled in an effective way (Malik et al., 2023). These homes managed mostly through mobile applications or through voice commands are a clear depiction of how IoT technologies can bring convenience and productivity in homes. Nevertheless, following the increasing number of IoT-based smart homes, concerns about the security and privacy of such interactive systems are growing as well (Gilani et al., 2024; Khanpara et al., 2023).

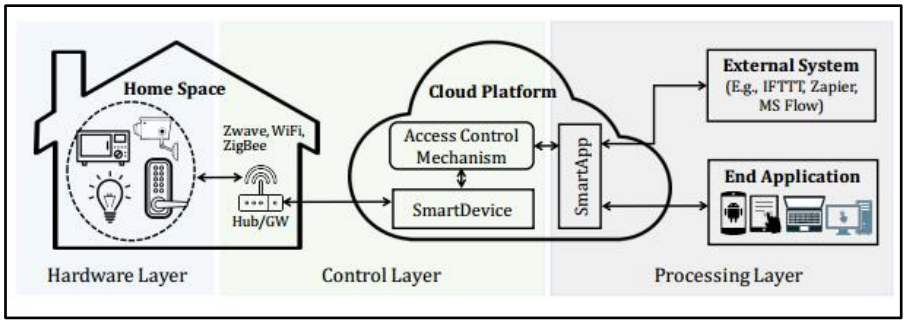


Figure 1: An example of applying the learning algorithms in the smart IoT system security (Source: American Institute of Mathematical Sciences)

1.1. Motivation

Smart homes derived from IoT have great benefits, but these houses have issues that traditional houses do not have. For example, connected devices are at risk of hack attacks, change of their configuration, destructive data attacks, and even physical control. These risks point to an urgent necessity for the implementation of effective security features of using IoT in SHS. Research addressing these challenges is expanding, focusing on several areas: from protecting communication of information and authorization to building of IDS and anoamics that can tell the security risks in real-time. For example, Khanpara et al. (2023) have suggested that implementing the details security protocols IoT can increase protection concerning suspected ability of perversion of protocols, control of devices, and the like and can increase efficiency and reliability as compared with traditional methods of protection.

In addition, more authors have further stressed developing elaborate IoT design frameworks to contain and optimize smart homes safety. This paper also envisioned a hierarchical SDN structure for nominal IoT performance and reliability Gilani et al., (2024), to reduce the level of packet loss for increased network reliability in smart home applications. Likewise, in their study titled The Internet of Things, Vardakis et. al., 2024 note that the use of a multi-tier IoT systems, where devices are categorized depending on their security level of operation provides a sturdier defense against threats and offer optimal safeguarding for information. These frameworks are designed to reduce likelihood of social engineers and hackers' getting into network or stealing devices' data by mapping out defensive barriers different from the traditional approach that guarantee security at both device and net levels.

1.2. Research Objective

This study describes an integrated IoT security architecture tailored to smart homes based on the advancements. Our framework includes cutting-edge technologies, which aim to implement a certain level of security without materially sacrificing user ease. I offer a comprehensive approach that could address various levels related to IoT systems, administering security concurrently with integrity, data privacy, and the ability to protect the system from cyber-attacks. This contrasts with the current models, which are focused on addressing specific aspects of smart home security.

1.3. Research Aim

To create and implement an appropriate preventative and protective IoT security plan that can be installed in smart homes and increase the security of data, systems, and people while also preserving the integrated features of smart home services. This study suggests the use of modern technologies, including multi-tier IoT architectures, blockchain for secure data management, and machine learning-based IDS for real-time threat identification and response, to mitigate vulnerabilities, which include unauthorized access, data leakage, and device sabotage. Machine learning will be used to develop solutions for identifying abnormalities, potential threats, and intelligent authentication systems that may proactively defend systems and adjust their operating settings in response to dynamically evolving threats. With the help of these technologies, the framework is expected to provide IoT-based smart homes with a high level of security while also being user-friendly.

1.4. Research Questions

RQ1: In what way can machine learning algorithms improve live intrusion detection and anomalous behavior detection in smart home IoTs?
Justification: Smart homes are known to have centralized assets which are interconnected to run the home and exchange information constantly and therefore vulnerable to cyber risks. Real-time data can be analyzed as patterns of network traffic and show changes and illicit behavior with devices to be an effective countermeasure for securing networks.

RQ2: In deployment of IoT, which security protocols and frameworks are effective in keeping the data of smart homes secure for their respective users?
Justification: The scale of IoT device growth escalates the chance of hacks and privacy loss. The proper security measures, for example, blockchain and authenticated encryption, should be determined to protect the delivered information and to increase users' confidence in smart home systems.

RQ3: Under what circumstances can a multi-tier IoT architecture provide better protection against cyber threats in smart home systems, as well as relevance and agility?
Justification: A multi-tier architecture ranks the devices according to their security needs and provides the appropriate armor for that tier. Investigating this approach can guide the improvement of the simultaneously stringent and realistic requirements for practical, serious security and the corresponding priorities of system performance and end-user control.

In the following chapter we will discuss in detail and review different research papers related to this topic and machine learning. In the chapter on methodology, we will discuss in detail about the different methods and techniques in details. In the chapter on implementation, we will discuss in detail about the proposed framework and in the results and analysis we will compare the performances.

2. Related Work

2.1. Smart Home Frameworks and Architectures

Smart homes are described as homes in which home operations are controlled using automated appliances and their importance is likely to increase in future societies. However, the issues of reliability and stability are the main concerns that slow down the development of smart home services. In response to these, Gilani et al. (2024) presented a multi-level SDN framework incorporating two coordinating controllers for enhanced use of existing services and enabler for new service types. This study contrast different topologies where the authors determined that cloud local topology gives the least packet loss, equal to 1.4%. Furthermore, Khanpara et al. (2023) present IoT smart home security which is more efficient in terms of performance, cost, and convenience compared to traditional methods, including problems with various attacks, such as protocol manipulation and device kidnapping. Altogether, the present works demonstrate significant improvement in smart home architectural designs and security strategies (Gilani et al., 2024; Khanpara et al., 2023).

It is evidenced that within smart homes the IoT has spurred major advancements in elderly care. In the method proposed by Aziz et al. (2023), a human activity recognition system is designed based on the raw sensor data, median filter to eliminate the noise, and supervised learning transform. At the heart of this system is a model based on the GRU, which uses federated distillation so that edge devices cooperate and enhance the performance of local models. For the classification of activities and for emergency detection, the proposed system has an accuracy of 0.95 and an F1-score of 0.94.

2.2. IoT Security and Privacy Concerns

In the same vein, more concentration is given to the IoT enterprise in smart homes by Aldahmani et al. (2023) with reference to the security flaws. The security of smart home technologies, such as climate control systems, intelligent lighting systems and smart locks offers a true innovation: Data protection issues come up as well as accessibility to different kinds of attacks. Their work done do supports the need to have tiered IoT system to counter security risks, this can be easily achieved through countermeasures and standard.

Internet of Things (IoT) is still one of the primary forces driving home automation given the industry 4.0 environment. Sayeduzzaman et al. (2024) present an innovative IoT smart home system that can perform home chores autonomously and increase security. This system enables users to manage and oversee home appliances with aid of Android devices and voice control. Some of these features are based on a touch keypad door lock, CCTV cameras, fire or gas detect unit that informs the homeowner. The objective of the system is to improve IoT security by integrating distant control of associated devices, energy conservation, and boosting safety for the elderly, children, and disabled people.

Technologies such as smart home automation systems are on the increase thanks to a proliferation of connected devices, serverless computing and MQTT protocol. Esposito at al.

(2023) present a framework using MQTT and serverless cloud functions to handle home appliances, where the emphasis is given to voice command interfaces. Their devised smart kitchen fan that includes the NB-IoT for the transmission of messages had acceptable performance with low packet loss. In a similar vein, Hasan et al. (2023) presented an IoT smart home automation system that has implemented security options such as automatic door locks and a gas detection system and remote control through the Blyn virtual application. This system results to increased security, efficiency in consumption of energy, and ability to give interval alerts thus enhancing usability of home automation.

2.3. Advancements in Intrusion Detection and Authentication

It is being witnessed that use of ML is gradually rising to develop better IDS for IoT setting. In the existing work by Rani et al. (2023), the different ensemble ML algorithms have been examined to create an optimal and timely IDS for IoT. This study also conducted a classification and prediction of the network attack based on the network traffic data via logistic regression, Random Forest, XGB, LGBM data analysis techniques. The differences between the models analyzed in the experiment were minimal in accuracy but were marked by a significantly shorter processing time and lower false positive rates of the proposed model – LGB-IDS based on LGBM.

In contrast, Popoola et al. (2023) discussed advantages and the effectiveness of applying blockchain to solve security and privacy issues in the IoT-based smart homes, with a focus on the health care field. They put forward a PoA-based blockchain system that ensures the secure, tamper-free storage of data and enforced novel models of authority, which allows more refined access control and decision-making for such data. This model focuses on key issues of IoT and presents a strong approach to designing adequate protection for private information.

Smart homes and buildings are incorporating the Internet of Things (IoT) technology in doing away with traditional living spaces as the IoT expands. However, this rise in interconnectivity introduces great security threats such as unauthorized access of the devices, data leakage, and device sabotage. Vardakis et al. (2024) gives an outline of these vulnerabilities and reviews the current protection methods like encryption and intrusion detection systems; however, the user's awareness is the critical defensive factor expected to be advocated. Alasmary and Tanveer (2023) extend an efficient solution called "ESCI-AKA", for preserving the information reliability of smart homes in resource-limited conditions. This framework involves an authenticated encryption scheme and hash function for the identification of users and secure transmission of messages using session keys. Comparisons made with other security frameworks reveal that ESCI-AKA is secure and lightweight, and thus may provide an ideal solution to IoT-based smart homes.

Smart homes have revolutionized modern living by enhancing convenience and quality of life through the integration of IoT, AI, and cloud technologies. Malik et al. (2023) outline how IoT-enabled smart homes employ sensors and actuators connected to a central unit for controlling devices like lights, locks, and appliances, often via smartphones. However, the rapid adoption of IoT introduces critical security challenges. As Bhardwaj et al. (2023) highlight, IoT camera

firmware is particularly vulnerable, and effective firmware security detection tools are becoming essential to protect against potential threats. Uppuluri and Lakshmeeswari (2023) propose a secure user authentication protocol using Modified Honey Encryption (MHE-IS-CPMT) and Elliptic Curve Cryptography (ECC), ensuring robust device access control and communication between users and smart home devices. By addressing these issues, the future of IoT-enabled smart homes lies in balancing convenience with enhanced security protocols to safeguard privacy and data integrity.

The IoT technology in smart homes has benefited from increased convenience but with high risks in privacy and security. To this end, Malik et al. (2023) proposed and investigated deep learning models for anomaly detection and face recognition in smart home IoT devices. In this paper, six models were considered, the result of which shows that LR-HGBC-CNN has been accurate (94%) in terms of detection of anomalies and 88% face recognition rate. These results indicate that the models work well but underscore the importance of future studies regarding their generality and methods of preserving the user's privacy.

Network intrusion detection systems (NIDS) have been discussed by Uppuluri and Lakshmeeswari (2023) in another study, conducted in the year 2023. They developed a Transformer-based NIDS using network traffic and IoT telemetry data with accuracy of 98.39% on the ToN_IoT dataset. The incorporation of telemetry data with traffic data can therefore solve the issues of how best to secure IoT devices in this hybrid access network, and thereby serves to strengthen the claims being made on behalf of intrusion detection.

The term IoT encompasses many areas and opportunities but several of them raise security concerns. There are always inputs from the devices in the online world thus making it have loopholes.par. According to Vegesna (2023) although IoT is a wonderful concept, some of its software has not been developed fully, more so they have not been secured fully and as such need to be examined for their weaknesses. There are many security concerns including but not limited to loss and theft of business sensitive information. Wu et al. (2023) presented a smart home authentication design to guarantee only authorized people are allowed into contact with devices, using both, outward and rigorous security analysis modalities. This underlines the fact that great emphasis should be placed on the effective protection of users' privacy in smart homes and smart city systems.

2.4. Research Summary

Table 1: Research Summary of all the papers researched

Paper Title	Authors	Dataset Used	Model Used	Result Summary
SDN-based multi-level framework for smart home services	Gilani, S.M.M., Usman, M., Daud, S., Kabir, A., Nawaz, Q., Judit, O.	dataset comprising various smart home devices and their operational data	Control Layer; Data Layer; Application Layer	Latency: Reduced response times for device interactions. Throughput: Increased data handling capacity. Security: Improved protection against

				network threats through dynamic resource allocation.
A context-aware internet of things-driven security scheme for smart homes	Khanpara, P., Lavingia, K., Trivedi, R., Tanwar, S., Verma, A., Sharma, R.	utilized datasets from real-time smart home applications	machine learning algorithms	significant improvements in detecting unauthorized access and intrusions, showcasing the efficacy of context-aware methodologies in enhancing smart home security.
Real-time Monitoring of Activity Recognition in Smart Homes: An Intelligent IoT Framework	Aziz, A., Mirzaliev, S., Maqsudjon, Y.	raw sensor data collected from various IoT devices in smart homes. The data undergoes preprocessing, including median filtering to minimize noise and enhance data quality.	framework is a Gated Recurrent Unit (GRU) model for activity recognition. The authors employ a federated distillation-based training strategy,	The framework achieved impressive performance metrics, recording an accuracy of 95% and an F1-score of 0.94.
Cyber-security of embedded IoTs in smart homes: challenges, requirements, countermeasures, and trends	Aldahmani, A., Ouni, B., Lestable, T., Debbah, M.	discusses various IoT security frameworks and implementations in smart homes.	layered security architecture and describes protocols like MQTT and HTTP for data transport, suggesting secure data handling across perception, network, and application layers to enhance IoT security within smart homes.	identify critical challenges in securing IoT devices in smart homes, such as data integrity, privacy issues, and device authentication.
An Internet of Things-Integrated Home Automation with Smart Security System	Sayeduzzaman, M., Hasan, T., Nasser, A.A., Negi, A.	discusses the design and implementation of an IoT-integrated home automation system.	implemented a model that integrates various IoT devices for home automation, focusing on user interface applications on Android devices to control and monitor security features.	The study presents an automated system designed to enhance home security and improve the convenience of daily tasks.

Design and Implementation of a Framework for Smart Home Automation Based on Cellular IoT	Esposito, M., Belli, A., Palma, L., Pierleoni, P.	implementation of a framework using cellular IoT technologies for smart home automation.	The study introduces a framework that integrates various IoT devices and cellular networks for enhanced connectivity and reliability in smart home environments.	The proposed framework aims to improve automation, energy efficiency, and user convenience. The results indicate that leveraging cellular IoT can enhance communication between devices, providing better control and monitoring capabilities for users
Constructing an integrated IoT-based smart home with an automated fire and smoke security alert system	Hasan, T., Abrar, M.A., Saimon, M.Z.R., Sayeduzzaman, M., Islam, M.S.	implementation of various sensors within an IoT framework for fire and smoke detection.	with an IoT framework to automate alerts in case of fire or smoke detection.	he study demonstrates that the developed system can effectively detect fire and smoke, providing timely alerts through a mobile application, thereby enhancing safety in smart homes.
Design of an intrusion detection model for IoT-enabled smart home	Rani, D., Gill, N.S., Gulia, P., Arena, F., Pau, G.	The paper utilizes publicly available datasets for training and testing the intrusion detection model, although specific datasets are not named.	propose a machine learning-based intrusion detection model that analyzes network traffic to identify unauthorized access attempts in IoT-enabled smart homes.	The results indicate that the model effectively detects various intrusion types with a high accuracy rate. The study highlights the necessity of robust security measures to protect IoT devices from cyber threats
IoT-enabled smart homes: Architecture, challenges, and issues	Malik, I., Bhardwaj, A., Bhardwaj, H., Sakalle, A.	mainly focuses on the architecture and challenges of IoT in smart homes.	. It emphasizes the use of machine learning and data analytics in enhancing the functionality and security of smart home environments.	paper identifies several challenges such as interoperability, security vulnerabilities, and user privacy issues in IoT-enabled smart homes.

2.5. Research Gap

Research on IoT-based smart homes has advanced highly, but there are still several gaps, particularly in the areas of scalability, security, and reliability. The integration of the different multi-level SDN frameworks suggested by Gilani et al. (2024) has shown a beneficial impact on packet drop; however, the stability of these systems in practical, high volumes of applications has not been thoroughly examined. Furthermore, while there are well-known methods such as blockchain solutions for safe data management (Popoola et al., 2023) and federated learning for age-related activity recognition (Aziz et al., 2023), the adaptability of these methods in this setting has not been thoroughly studied. To verify the effectiveness of the suggested strategy, ESCI-AKA (Alasmary and Tanveer, 2023), a security framework for smart homes, needs to be evaluated in various settings where smart houses can be installed.

According to Vardakis et al. (2024), one of the main shortcomings is the absence of strategies for changing the user's awareness and behaviour. Even while encryption techniques and sophisticated intrusion detection systems have advanced, the end user counterpart of the scheme has not received the attention it deserves and, more significantly, has not been incorporated into the systems. Finally, Bhardwaj et al. (2023) noted the quick growth of IoT devices and the need for firmware and software upgrades to reduce threats, although this has not yet been included in the large-scale solution. To close these gaps, future research should focus on the following suggestions: In addition to user-centred and monitoring designs, real-world utility should be at the forefront of the research agenda.

2.6. Research Outcome

The main contribution of this research is that this research proposes a novel conceptual framework to design an IDS for IoT-based smart homes using state-of-the-art machine learning innovative techniques, and structured data pre-processor. The study can overcome the usual issue of having unequal numbers in the classes by using SMOTE, thus will provide equal representation in the class for model training purposes which is very important. To reduce the possibility of black-box models and improve the reliability and adaptability of the intrusion detection system, six different machine learning models are employed, including Decision Trees, Logistic Regression, LightGBM, XGBoost, etc. Additionally, the incorporation of the models into smart home systems involving joblib enables the immediate deployment of the models into the system while achieving efficient real-time detection, an improvement over existing IDS packages. Separately, this work integrates EDA, dataset balancing, high-performance end-to-end ML, and near-real-world deployment-readiness into a scalable and flexible threat response system against emerging cyber threats targeting IoT-based smart home systems.

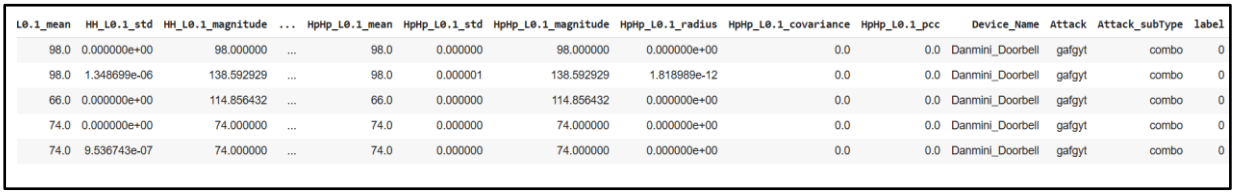
3. **Research Methodology**

3.1 Data Collection

The datasets utilized in this study are pivotal for understanding and enhancing IoT security in smart homes. The BoTNeTIoT-L01 dataset is the most recent dataset containing traffic from

nine IoT devices, sniffed using Wireshark in a local network through a central switch. It includes two prominent botnet attacks: Mirai and Gafgyt. The dataset contains twenty-three statistically engineered features extracted from the .pcap files. These features were computed over a 10-second time window with a decay factor of 0.1, facilitating detailed temporal analysis. The RT_IOT2022 dataset complements this by including various attack types, thus providing a broader evaluation spectrum for the proposed security framework. These datasets offer a robust foundation for training and testing machine learning models due to their comprehensiveness and relevance.

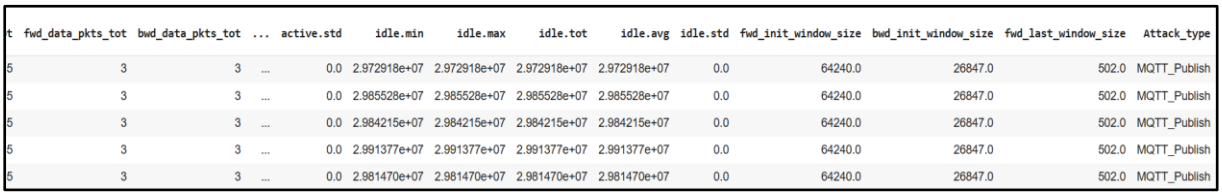
3.1.1 BoTNeTIoT-L01 dataset



L0.1_mean	HH_L0.1_std	HH_L0.1_magnitude	...	HpHp_L0.1_mean	HpHp_L0.1_std	HpHp_L0.1_magnitude	HpHp_L0.1_radius	HpHp_L0.1_covariance	HpHp_L0.1_pcc	Device_Name	Attack	Attack_subType	label
98.0	0.000000e+00	98.000000	...	98.0	0.000000	98.000000	0.000000e+00	0.0	0.0	Danmini_Doorbell	gafgyt	combo	0
98.0	1.348699e-06	138.592929	...	98.0	0.000001	138.592929	1.818989e-12	0.0	0.0	Danmini_Doorbell	gafgyt	combo	0
66.0	0.000000e+00	114.856432	...	66.0	0.000000	114.856432	0.000000e+00	0.0	0.0	Danmini_Doorbell	gafgyt	combo	0
74.0	0.000000e+00	74.000000	...	74.0	0.000000	74.000000	0.000000e+00	0.0	0.0	Danmini_Doorbell	gafgyt	combo	0
74.0	9.536743e-07	74.000000	...	74.0	0.000000	74.000000	0.000000e+00	0.0	0.0	Danmini_Doorbell	gafgyt	combo	0

Figure 9: BotNeTIoT sample examples
Source: https://www.kaggle.com/datasets/azalhowaide/iot-dataset-for-intrusion-detection-systems-ids/data

3.1.2 RT_IOT2022 dataset



t	fwd_data_pkts_tot	bwd_data_pkts_tot	...	active.std	idle.min	idle.max	idle.tot	idle.avg	idle.std	fwd_init_window_size	bwd_init_window_size	fwd_last_window_size	Attack_type
5	3	3	...	0.0	2.972918e+07	2.972918e+07	2.972918e+07	2.972918e+07	0.0	64240.0	26847.0	502.0	MQTT_Publish
5	3	3	...	0.0	2.985528e+07	2.985528e+07	2.985528e+07	2.985528e+07	0.0	64240.0	26847.0	502.0	MQTT_Publish
5	3	3	...	0.0	2.984215e+07	2.984215e+07	2.984215e+07	2.984215e+07	0.0	64240.0	26847.0	502.0	MQTT_Publish
5	3	3	...	0.0	2.991377e+07	2.991377e+07	2.991377e+07	2.991377e+07	0.0	64240.0	26847.0	502.0	MQTT_Publish
5	3	3	...	0.0	2.981470e+07	2.981470e+07	2.981470e+07	2.981470e+07	0.0	64240.0	26847.0	502.0	MQTT_Publish

Figure 10: RT_IOT2022 dataset sample examples
Source: https://www.kaggle.com/datasets/supplejade/rt-iot2022real-time-internet-of-things

3.2 Data Pre-processing

Pre-processing is a critical step to ensure the data is ready for machine learning models. The following steps were undertaken:

3.2.1 Handling Categorical Data

BoTNeTIoT-L01: Label encoding was employed to convert categorical variables to numerical values, enabling the models to process them effectively.
RT_IOT2022: Label encoding was similarly applied after identifying the categorical columns, ensuring consistency across datasets.

3.2.2 Class Balancing

To address class imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. This technique generates synthetic samples for the minority class, ensuring a balanced distribution and improving the model's performance on imbalanced data.

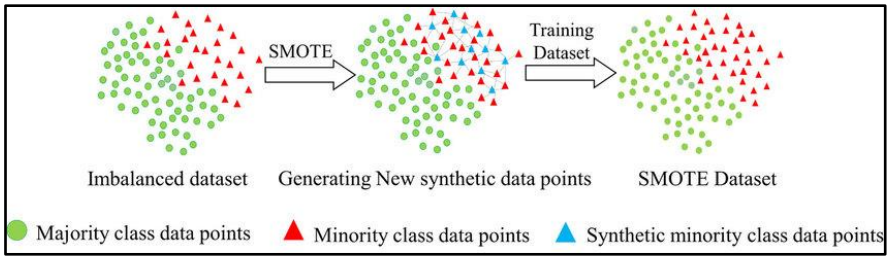


Figure 11: SMOTE application on imbalanced dataset so that the data is balanced (Source: AI Mind)

3.2.3 Feature Engineering

The BoTNeTIoT-L01 dataset included twenty-three statistically engineered features, such as mean, variance, count, magnitude, radius, covariance, and correlation coefficient. These features were computed over a 10-second time window with a decay factor of 0.1. This comprehensive feature set facilitated a nuanced understanding of the data. For the RT_IOT2022 dataset, features were meticulously selected and cleaned to ensure consistency and reliability, providing a solid foundation for model training.

4. Design Specification

A diverse range of machine learning models were selected to identify the most effective approach for intrusion detection in smart home IoT environments:

4.1 Decision Tree Classifier

Decision Trees are non-linear models that partition data based on feature values recursively such that splits are optimized on maximizing metrics like information gain or Gini impurity. Gini impurity is the probability of choosing a wrong class for classification, and entropy is another way of quantifying the uncertainty. Decision trees have been found to be interpretable and effective for classification and regression tasks but are prone to overfitting.

4.2 Random Forest Classifier

This is an ensemble learning method that creates multiple decision trees during the training process and combines the output of these trees for enhanced accuracy and reduction in overfitting. It applies techniques such as bootstrap aggregating, where subsets of data are randomly sampled to train individual trees.

Each tree makes predictions and then the overall prediction is decided by majority voting. Information gain, or the difference between the entropy of the dataset and the weighted sum of the entropy of its subsets, can be used to optimize the splits within a tree. This method is very powerful for complex high-dimensional data.

$$IG(T, A) = Entropy(T) - \sum_{v \in A} \frac{|T_v|}{|T|} Entropy(T_v)$$

Here, $IG(T, A)$ is the information gain for the dataset A while T_v are the subset of T after the splits.

4.3 Logistic Regression

Logistic Regression is a statistical and machine learning method for binary classification problems, where the outcome or target variable is categorical in nature and has two outcomes, which are often denoted as 0 (negative class) and 1 (positive class).Instead of making a continuous value prediction as in linear regression, logistic regression predicts the probability that a given input belongs to the positive class by using the logistic or sigmoid function. Then the probabilities of the outputs are translated into class labels by a decision threshold, commonly set at 0.5.

4.4 LightGBM Classifier

LightGBM is a highly optimized, powerful, and efficient gradient boosting framework developed by Microsoft for classification, regression, and ranking tasks. Its optimization for speed and memory usage makes it best suited for large datasets that can have millions of rows and features. Unlike the traditional gradient boosting methods, LightGBM uses an algorithm based on histograms to discretize continuous features into bins. It uses a leaf-wise tree growth strategy, where it grows the leaf with the maximum loss reduction, thus giving deeper and more accurate trees than level-wise growth. LightGBM also natively handles missing values and categorical features, so no pre-processing like one-hot encoding is required. With its speed to train, scalability in efficient usage, and very high predictive accuracy, LightGBM is widely used in fraud detection, recommendation systems, and ranking tasks.

4.5 XGBoost Classifier

XGBoost is the advanced gradient boosting framework, wherein it builds trees sequentially with an attempt to correct the residual error of the previous trees. Its objective function has combined a loss term and the regularization term, so this balances model complexity with its performance. XGBoost optimizes for both computational efficiency and the prediction accuracy, making it a very popular method in working with tabular data.

4.6 Multilayer Perceptron (MLP) Classifier

A multilayer perceptron (MLP) is a type of artificial neural network (ANN) that includes at least three layers: an input layer, one or more hidden layers, and an output layer. Every connection between nodes, or neurons, of each layer is associated with weights and biases. The MLP is a feedforward network meaning information flows one way and never loops back, flowing from the input layer towards the output layer. Every neuron in the hidden and the output layer takes a weighted sum of its inputs adds a bias, and feeds that to an activation function which gives non-linearity so it can learn all those hard patterns in your data. MLPs are trained using a supervised learning algorithm, typically backpropagation, to minimize a loss function and optimize the weights. They are widely used for tasks like classification, regression, and function approximation.

5. Implementation

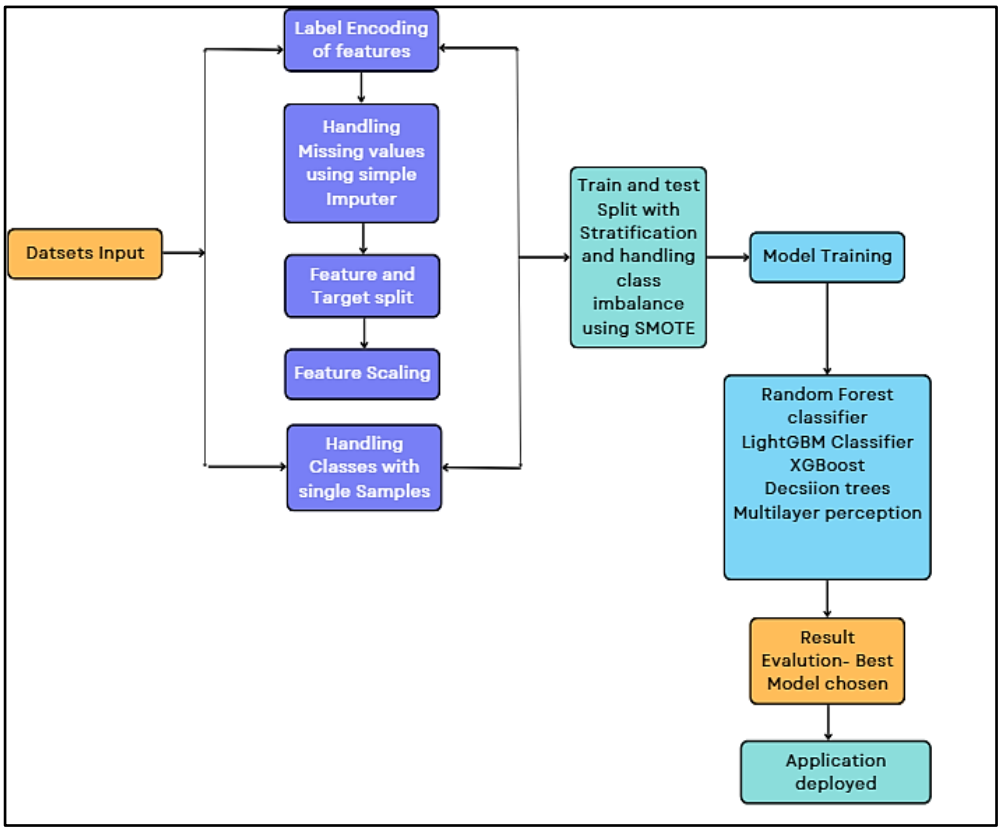


Figure 21: Implementation flow

5.1 Algorithm Flow

1. Data Loading and EDA: The objective is to perform EDA (Exploratory data analysis) to find out the anomalies, missing values and to get access to the data distribution. Uses graphs like bar plots to illustrate the target variable label's distribution for creating the visualization. The BoTNeTIoT-L01 and RT_IOT2022 were loaded using the panda's library. Here it show how it is done:
2. Data Pre-processing: Pre-processing is a crucial step to prepare the data for machine learning models. This includes handling categorical data, dealing with missing values, scaling features, and balancing classes. This is how it was done:
    i. Handling Categorical Data: Both datasets contained categorical variables that needed to be converted to numerical values using label encoding. To prepare the dataset for machine learning models, one must handle categorical data and convert it into numeric form. Using Python's Label Encoder, all categorical columns are identified with the select_dtypes method, and their unique string labels are encoded into integers, which allows models to process non-numeric data. It is appropriate to use Label Encoder here because the data is assumed to be ordinally or nominally encoded, but not one-hot encoded.
    ii. Handling Missing Values: Missing values in numerical columns will lead to some errors while training the model. Thus, the SimpleImputer is used with a strategy set to "mean." Now, missing values for those columns are replaced with mean values of their column to avoid loss of any data and ensure consistency in making predictions with minimum possible bias.

3. Separate Features and Target: The dataset is divided into features (X) and the target (y), where the target column, Attack_type, represents the labels the model is predicting. Features are everything else in the dataset after dropping the target column. This separation is important for training machine learning models effectively.
4. Feature Scaling and Standardization: By standardizing the feature, making features in comparable ranges, using MinMax Scaler, such that data now falls on range between 0 to 1; then I normalized scaled feature using the Standard Scaler and ensured feature would have the mean zero with standard deviation of value of one. That is considered important, because normally in cases that are too dependent on standardization; therefore, any machine-learning models are generally improved while implemented such as support-vector-machines, k nearest-neighbour types.
5. Handling rare classes in Target Variable: Classes with only one sample in the target variable will lead to imbalances and bad generalization of the model. Rare classes are identified by analyzing class distribution using value_counts(). The instances of these rare classes are re-labeled to the most frequent class calculated using mode. Thus, no label is underrepresented to the point of becoming statistically insignificant at training time.
6. Data Splitting: Finally, the dataset is split into training and testing sets using an 80-20 ratio. This allows us to train the model on one portion of the data while evaluating its performance on a separate, unseen portion, helping to prevent overfitting. The training data is used for model training, and the test data is reserved for performance evaluation.
7. Smote for Class Balancing: Before training the model, the class distribution in the training and test sets is analyzed by using a Counter to identify underrepresented classes, which are very typical in imbalanced datasets. To combat this, SMOTE (Synthetic Minority Oversampling Technique) is applied to the training data. Synthetic samples of the minority classes are then generated by interpolating between data points in the existing classes. This ensures that all classes have an equal representation, so the model will be learned suitably without class bias by the majority class. The good distribution of classes in the training set is checked once SMOTE is applied to ensure that the data is ready for training.
8. Model Training and Evaluation: Six different machine learning models were trained and evaluated on preprocessed and balanced datasets. Each model was assessed using metrics like accuracy, precision, recall, F1 score, and ROC AUC score. The Models that were trained are: Decision Tree Classifier, Random Forest Classifier, Logistic Regression, LightGBM Classifier, XGBoost Classifier, Multilayer Perceptron Classifier.
9. Visualization and Analysis: Visualizing the results is crucial for understanding the performance and effectiveness of the models. Confusion matrices and ROC curves were plotted for each model to provide clear insights into their accuracy and predictive capabilities.
10. Model Deployment and Integration: The trained models were saved using joblib to facilitate easy deployment and integration into smart home systems. This ensures that the models can be utilized for real-time intrusion detection without the need for retraining.

6. Evaluation

Each model underwent rigorous training using preprocessed and balanced datasets. The models were evaluated on the test set using various metrics, including accuracy, precision, recall, F1 score, and ROC AUC score. Confusion matrices and ROC curves were plotted for each model, providing a visual comparison of their performance. These evaluations ensured that the models were robust, reliable, and capable of detecting IoT-related intrusions effectively.

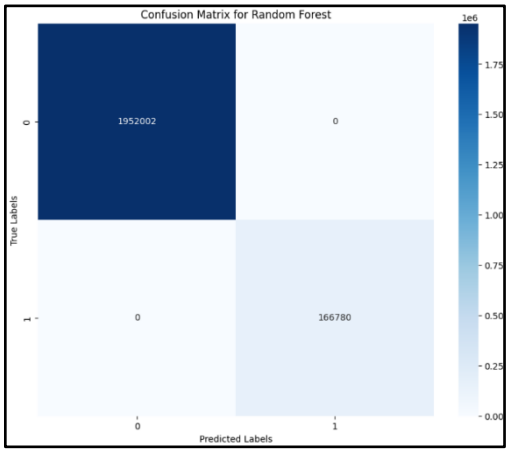
6.1 Model Saving and Streamlit Deployment

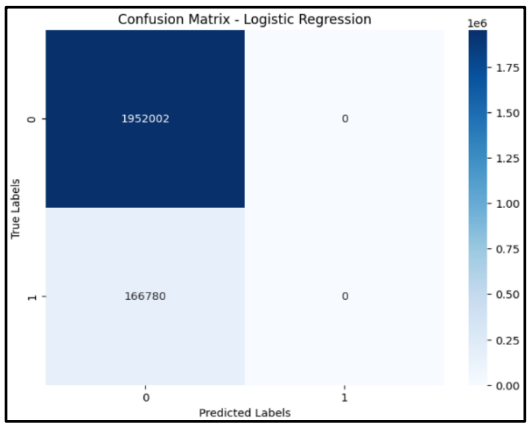
To facilitate future use and deployment, the trained models were saved using joblib. This step ensured that the models could be easily loaded and integrated into real-time intrusion detection systems in smart homes. The saved models provide a ready-to-use solution for enhancing IoT security, offering a practical and efficient approach to addressing smart home vulnerabilities.

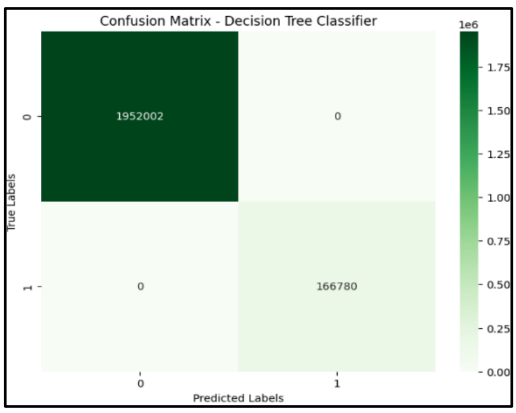
6.2 Results and Analysis

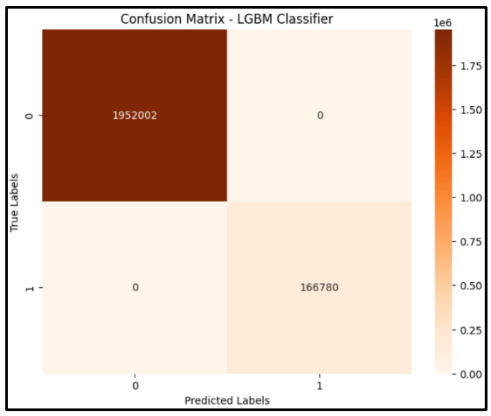
6.2.1 Case 1: Using BoTNeTIoT-L01 is a dataset

The Decision Tree Classifier demonstrates exceptional performance with very high values across all key metrics. The accuracy of 1.00 indicates that the model correctly predicts almost all instances. Both precision and recall values (1.00) show that the model is highly effective at identifying positive cases while minimizing false positives and false negatives. The F1 score of 1.00, which is the harmonic mean of precision and recall, further supports that the model strikes a strong balance between both metrics. These results suggest that the Decision Tree Classifier is highly reliable for this classification task.









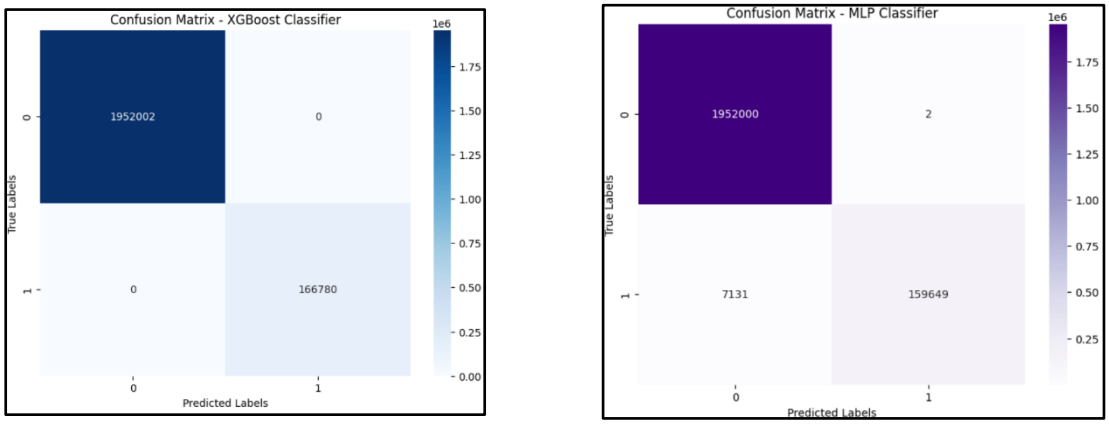


Figure 23: Case 1 Performance evaluation of different models

Random Forest Classifier is highly performing as well, with almost perfect metrics. Overall accuracy is 1.0 which is clear evidence that the model is classifying benign and malicious instances correctly. The precision is around 1.0, and recall is 1.0. The F1-scores 1.0, showcasing the model's excellent performance in distinguishing between the two classes. The LGBM Classifier also performs outstandingly. All the metrics are closer to perfection. An accuracy value of 1.0 shows that the model classifies nearly all instances correctly. Precision of 1.0 shows that the model does a great job minimizing false positives, and its recall of 1.0 shows that it detects true positives effectively. The F1 score of 1.0 confirms the perfect balance between precision and recall. These metrics point out that LGBM Classifier is a good and stable classifier for classification tasks. The accuracy of logistic regression is about 0.921. The precision is 0.84. Recall and F1 Score was about 0.92 and 0.88 respectively. The XGBoost Classifier is good but the metrics are lower compared to Decision Tree and LGBM classifiers. The accuracy stands at 1.0, meaning that the model gets most instances correct. However, precision of 1.0 has relatively more false positives compared to the other models. The recall of 1.0 is a huge strength in identifying true positives while the F1 score indicates a fair balance between the precision and recall, at about 1.0. Although effective, the XGBoost Classifier might need fine-tuning or optimization specifically for this task. The MLP Classifier has on average a strong performance with 0.9966 accuracy. Precision is about 0.996 This further translates to F1-scores of 0.996 and recall of 0.996. The slightly lower recall for the malicious class seems to indicate that a few malicious instances were misclassified.

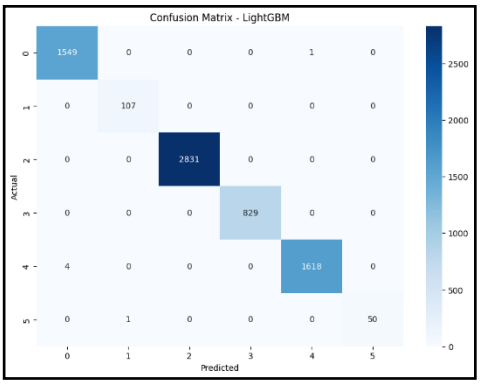
Table 2: Model Comparison of BoTNeTIoT-L01 Dataset

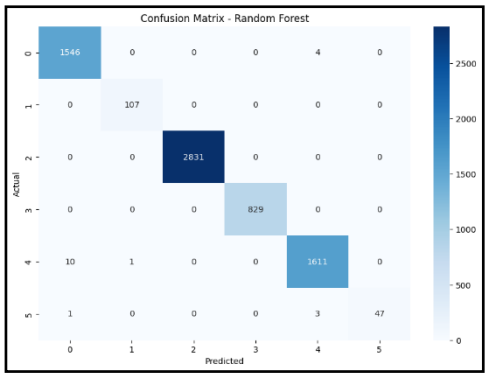
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0
LightGBM	1.0	1.0	1.0	1.0

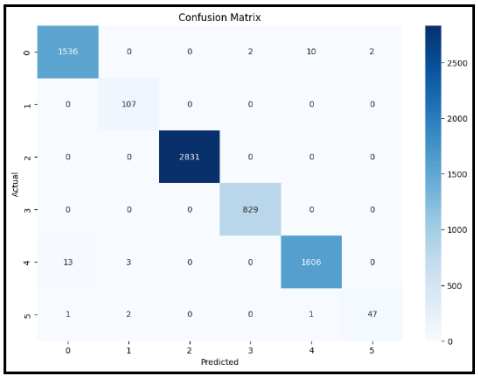
MLP	0.9966	0.9966	0.9966	0.996
Logistic Regression	0.92	0.848	0.921	0.88
XGBoost	1.0	1.0	1.0	1.0

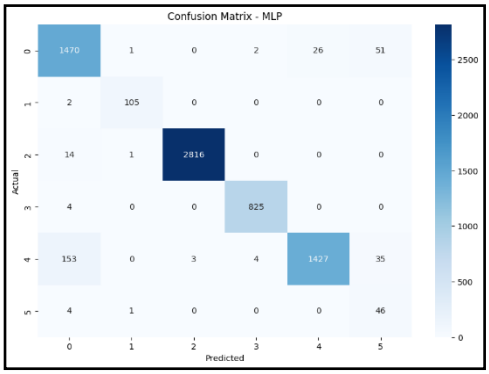
6.2.2 Case 2: Using IoT Intrusion Detection dataset

The Decision Tree classifier achieved an accuracy of 0.9951 on the test dataset, demonstrating excellent performance across most classes. It displayed high precision, recall, and F1-scores, indicating its capability to correctly classify instances while minimizing false positives and false negatives.









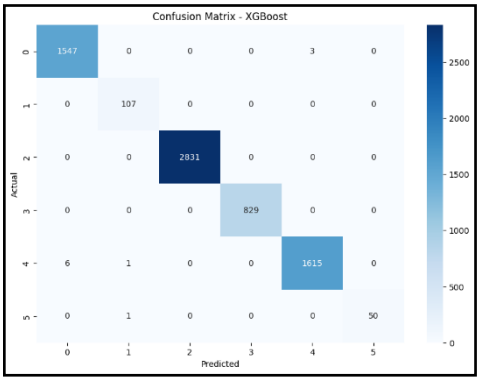


Figure 24: Performance Analysis of different models for the dataset 2

The Random Forest classifier achieved an accuracy of 0.997, indicating strong performance in multi-class classification. The model demonstrated near-perfect precision, recall, and F1-scores across most classes. While class 5.0 showed a slight dip in recall, its precision and F1-score remained robust. The LightGBM classifier achieved an accuracy of 0.9991, demonstrating excellent performance. The model showed near-perfect precision, recall, and F1-scores across most classes, with slight variations in class 5.0, where recall dropped marginally to 0.98. This is still a very high score, indicating that the model handles even rare cases well, though with some slight limitations. The XGBoost classifier achieved an excellent accuracy of 0.998, showcasing strong overall performance. Most of the precision, recall, and F1-scores are still very close to 1.00. Class 1.0 and class 5.0 show a slight drop in precision and recall, but they still maintain high values, indicating that the model handles imbalanced classes quite well.

The MLP classifier achieved an accuracy of 0.95, which is a strong performance overall, though slightly lower compared to the previous models. The model displayed solid results across most classes, particularly for classes 1.0, 2.0, and 3.0. Class 5.0 exhibited lower precision (0.35), which resulted in a significantly lower F1-score (0.50), but recall was still relatively high (0.90). Class 4.0 also showed a notable decrease in recall (0.88), leading to an F1-score of 0.93.

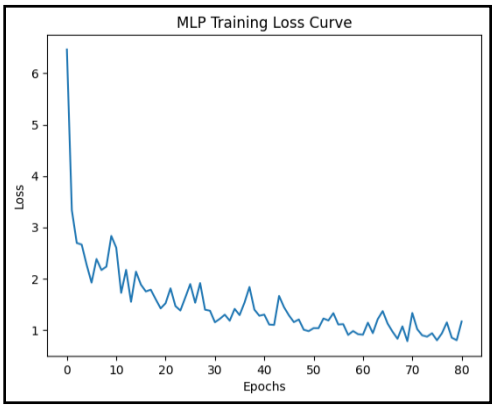


Figure 25: Loss Curve of MLP

Table 3: Model Comparison of IoT Intrusion Detection dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.9951	0.96	0.92	0.94
Random Forest	0.997	1.00	0.92	0.96
LightGBM	0.991	1.00	0.98	0.99

MLP	0.95	0.35	0.90	0.50
XGBoost	0.99	1.00	0.98	0.99

6.2.3 Real Time Analysis

Screenshots of a Streamlit application that deploys an IoT intrusion detection model are shown. The user interface accepts key features of network activity, including packet size, protocol, and flow duration. After such inputs, the model will analyze the data and provide a predicted label. For this example, the model gives a predicted label of 0, meaning that the input is benign traffic. This outcome thus shows the potential of the model in real-time distinction between normal and malicious IoT activity.

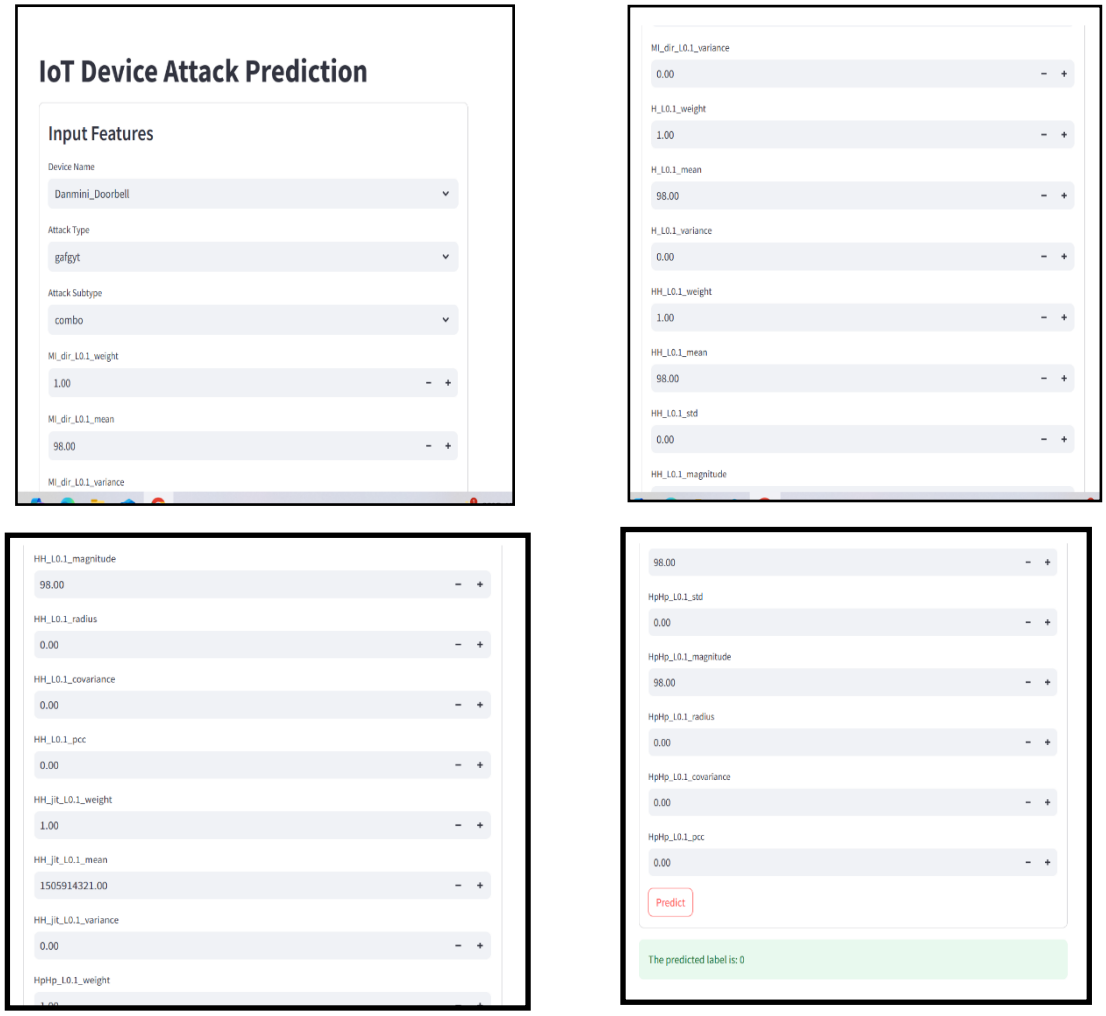


Figure 26: Real Time Analysis for the application

Link: https://iot-project.streamlit.app/

https://github.com/sanx123/iot-project

7. Conclusion and Future Work

In this research, a novel machine learning IDS has been devised for IoT smart homes to meet important research gaps including class imbalance problem and real-world implementation. Hence, to achieve an unbiased model training and performance evaluation, the research used SMOTE for class balancing and trained six models: Decision Tree, Random Forest, Logistic Regression, LightGBM, XGBoosting, and Multilayer Perceptron. The incorporation of these models in real-time systems using joblib makes it possible to implement their real-time intrusion detecting capability. This approach can supplement the lack of theoretical and practical machine learning studies in the IoT security field, by providing a viable approach to protect smart home environments.

Future work can propose improvements to this work where the work can incorporate deep learning models that include CNN and the transformers approach to feature extraction and better proceed with detecting. Besides, the given framework is acknowledged to be extensible towards the integration of continual learning and thus, the IDS can develop with emerging threats. The evaluation with different datasets of IoT systems will provide generalization and overall validity to the proposed models. Lastly, incorporating the user feedback systems and the real-time adaptive functionalities into the system may add value and improve the performance of reaction. These developments will reinforce the status of IDS for adopting eliminating cyber threats to smart homes equipped with IoT technology.

References

Gilani, S.M.M., Usman, M., Daud, S., Kabir, A., Nawaz, Q. and Judit, O., 2024. SDN-based multi-level framework for smart home services. *Multimedia Tools and Applications*, *83*(1), pp.327-347.

Khanpara, P., Lavingia, K., Trivedi, R., Tanwar, S., Verma, A. and Sharma, R., 2023. A context-aware internet of things-driven security scheme for smart homes. *Security and Privacy*, *6*(1), p.e269.

Aziz, A., Mirzaliev, S. and Maqsudjon, Y., 2023. Real-time Monitoring of Activity Recognition in Smart Homes: An Intelligent IoT Framework. *Journal of Intelligent Systems & Internet of Things*, *10*(1).

Aldahmani, A., Ouni, B., Lestable, T. and Debbah, M., 2023. Cyber-security of embedded IoTs in smart homes: challenges, requirements, countermeasures, and trends. *IEEE Open Journal of Vehicular Technology*, *4*, pp.281-292.

Sayeduzzaman, M., Hasan, T., Nasser, A.A. and Negi, A., 2024. An Internet of Things-Integrated Home Automation with Smart Security System. *Automated Secure Computing for Next-Generation Systems*, pp.243-273.

Esposito, M., Belli, A., Palma, L. and Pierleoni, P., 2023. Design and Implementation of a Framework for Smart Home Automation Based on Cellular IoT, MQTT, and Serverless Functions. *Sensors*, *23*(9), p.4459.

Hasan, T., Abrar, M.A., Saimon, M.Z.R., Sayeduzzaman, M. and Islam, M.S., 2023. Constructing an integrated IoT-based smart home with an automated fire and smoke security alert system. *Malaysian Journal of Science and Advanced Technology*, pp.1-10.

Rani, D., Gill, N.S., Gulia, P., Arena, F. and Pau, G., 2023. Design of an intrusion detection model for IoT-enabled smart home. *IEEE Access*, *11*, pp.52509-52526.

Popoola, O., Rodrigues, M., Marchang, J., Shenfield, A., Ikpehia, A. and Popoola, J., 2023. A critical literature review of security and privacy in smart home healthcare schemes adopting IoT & blockchain: problems, challenges and solutions. *Blockchain: Research and Applications*, p.100178.

Vardakis, G., Hatzivasilis, G., Koutsaki, E. and Papadakis, N., 2024. Review of Smart-Home Security Using the Internet of Things. *Electronics*, *13*(16), p.3343.

Alasmary, H. and Tanveer, M., 2023. ESCI-AKA: Enabling secure communication in an iot-enabled smart home environment using authenticated key agreement framework. *Mathematics*, *11*(16), p.3450.

Malik, I., Bhardwaj, A., Bhardwaj, H. and Sakalle, A., 2023. IoT-enabled smart homes: Architecture, challenges, and issues. *Revolutionizing Industrial Automation Through the Convergence of Artificial Intelligence and the Internet of Things*, pp.160-176.

Bhardwaj, A., Kaushik, K., Bharany, S. and Kim, S., 2023. Forensic analysis and security assessment of IoT camera firmware for smart homes. *Egyptian Informatics Journal*, *24*(4), p.100409.

Uppuluri, S. and Lakshmeeswari, G., 2023. Secure user authentication and key agreement scheme for IoT device access control based smart home communications. *Wireless Networks*, *29*(3), pp.1333-1354.

Rahim, A., Zhong, Y., Ahmad, T., Ahmad, S., Pławiak, P. and Hammad, M., 2023. Enhancing smart home security: anomaly detection and face recognition in smart home IoT devices using logit-boosted CNN models. *Sensors*, *23*(15), p.6979.

Wang, M., Yang, N. and Weng, N., 2023. Securing a smart home with a transformer-based iot intrusion detection system. *Electronics*, *12*(9), p.2100.

Vegesna, V.V., 2023. Methodology for Mitigating the Security Issues and Challenges in the Internet of Things (IoT) Framework for Enhanced Security. *Asian Journal of Basic Science & Research*, *5*(1), pp.85-102.

Wu, T.Y., Meng, Q., Chen, Y.C., Kumari, S. and Chen, C.M., 2023. Toward a secure smart-home IoT access control scheme based on home registration approach. *Mathematics*, *11*(9), p.2123.