Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Innovative IoT security protocol: High-accuracy device identification and resilience against credential compromise (HADIRACC)

Joseph Teguh Santoso[a*], Mars Caroline Wibowo[a] and Budi Raharjo[a]

[a]*Department of Computer and Business, STEKOM University, Semarang, Indonesia*

CHRONICLE

Article history:
Received: January 6, 2024
Received in revised format: February 20, 2024
Accepted: April 24, 2024
Available online: April 24, 2024

Keywords:
IoT Security
Device Authentication
Machine Learning
Proximity-based Authentication

ABSTRACT

The IoT ecosystem faces increasingly complex security challenges due to the rapid growth of global IoT devices. Security risks related to device identification and credential compromise are on the rise, especially with the proliferation of IoT devices in various aspects of life. This research highlights the need to address these vulnerabilities through the development of robust security protocols, aiming to create a more secure IoT ecosystem and enhance user trust in this technology. The objective of the research is the development of an innovative IoT security protocol; High-Accuracy Device Identification and Resilience Against Credential Compromise (HADIRACC). This paper contributes significantly to enhancing the security and reliability of the IoT ecosystem. The research methods employed encompass the development of security protocols, the development of a proximity-based solution, and the classification of IoT devices using data processing techniques and machine learning-based classification. This study involves the collection and preprocessing of datasets, training different classifiers using 70% of the dataset, and testing the classifiers using the remaining 30%. The proposed protocol can effectively enhance the security of IoT devices by addressing various scenario-based attacks. Furthermore, the results of the analysis of the five classifiers used in this study indicate that Random Forest has the highest F1 score accuracy, reaching 88.8%. This suggests that Random Forest, as a classifier, can make the most accurate predictions compared to other classifiers.

1. Introduction

In the contemporary era marked by the swift evolution of the Internet of Things (IoT), where millions of devices are globally interconnected to provide services and collect data, security challenges are becoming increasingly complex. IoT security is crucial given the widespread adoption of IoT devices in various aspects of life, including industries and smart homes. Alongside the significant benefits offered by the IoT ecosystem, security risks related to device identification and credential compromise are becoming more tangible and concerning (Omolara et al., 2022). According to the latest report, the number of connected IoT devices worldwide is estimated to reach billions, with projections of continuous growth across various industries (Chataut et al., 2023). Despite the growth, attacks on IoT devices have also experienced a significant increase. Since 2021, various attacks on IoT devices have been identified using various methods (Ghose et al., 2024; Hansdah et al., 2022; Yu et al., 2021) including successful credential compromises, which have become one of the main threats to the security of devices and associated data. This research aims to address existing security vulnerabilities by developing a protocol capable of providing device identification with high accuracy and maintaining resilience against potential credential compromises. Thus, the research is directed towards creating a more robust security layer within the IoT ecosystem, reducing the risk of attacks, and enhancing user trust in this technology. The success of the proposed security protocol can have a positive impact on various sectors that rely on IoT devices. Confronted with increasingly complex security challenges in the IoT world, this research emerges as a proactive step to safeguard the integrity, confidentiality, and availability of data handled by connected

* Corresponding author.
E-mail address joseph_teguh@stekom.ac.id (J. T. Santoso)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.4.014

devices. Through the development of innovative security protocols, it is hoped for this research to provide an effective solution to enhance the security of the IoT ecosystem and ensure the sustainability and broader adoption of this highly potential technology. In this context, the research aims to present an innovative IoT security protocol that provides high accuracy in device identification and resilience against credential compromises. The successful implementation of this protocol is expected to offer a leading solution in dealing with the security challenges faced by the current IoT ecosystem.

2. **Related Work**

Currently, numerous IoT devices are connected to the internet, encompassing various household appliances, automotive devices, medical devices, smart locks, monitoring sensors, cameras, and so forth, each serving different functions (Rawat, 2022). These various IoT devices store users' personal information, making the security of these devices vulnerable to data theft and firmware vulnerabilities (Liu et al., 2023). One viable approach to tackle this concern involves employing policy-driven access controls to thwart insecure devices from gaining control over the home network (Tomer & Sharma, 2022). However, that solution cannot differentiate between the configurations of owner and hacker devices. Some existing research has improved the authentication process of IoT devices within the network by employing blockchain-based authentication (Ghose et al., 2024). Other research by Thomer and Sharma (2022) uses a different approach to authenticating IoT, they employ machine learning methods to predict IoT devices. However, to the extent of their research, there is no comprehensive protocol or solution has been identified. A method entirely independent of user interaction based on machine learning, where the protocol is detached from network attributes, susceptible to accurately identifying various types of IoT devices, and provides a comprehensive and effective solution applicable in practical scenarios. Therefore, in this research, an enhancement of IoT device authentication is carried out by proposing a credential compromise scenario that has not been discussed in previous studies. One way to achieve this is through the implementation of an Innovative IoT Protocol, which boasts high accuracy in device identification and resilience against credential compromise. This proposition is put forth because it addresses the need for a robust solution that not just accurately identifies IoT devices but also provides a safeguard against potential security vulnerabilities associated with credential compromise. By adopting such an innovative protocol, there is a proactive approach to enhancing the security and efficiency of IoT networks, thereby ensuring a more reliable and resilient system in the face of emerging challenges. This is proposed because various existing machine learning models thus far, fail to accurately identify and authenticate IoT devices within the network.

This research addresses cases where vulnerable IoT devices can be hacked through network traffic and the activities of devices of different types. In the study by (Hansdah et al., 2022; Yu et al., 2021) they employed successful credential hacking methods, which have become one of the major threats to the security of devices and associated data. To tackle this, the study proposes the use of existing device type policies (Barua et al., 2022; Ghose et al., 2024) combined using machine learning for device categorization to limit network access based on device capabilities. The research introduces a new technique for classifying device types and device-based authentication using cross-layer data to classify them into four categories: Smart Entry, Cleaning, Appliance, and Home Sensors. This prevents hackers or criminals from stealing data on the device when it is powered on. The contributions of this research are as follows:

- Device-based authentication protocol capable of preventing criminals from replicating devices to steal data or engage in malicious activities through the compromised devices.
- Enhanced security analysis of the proposed technique.
- Experiments were performed using personally acquired datasets from a machine learning platform for device type classification. The aim was to showcase the effectiveness of different classifiers employing various algorithms. This study focused on utilizing only four algorithmic models as classifiers.

2.1 Machine Learning-Based Authentication

This is one of the latest innovations in the realm of information security that offers a new approach to verifying user identities. This method utilizes machine learning algorithms to analyze unique behavioral patterns of users, enabling the system to recognize whether someone is the legitimate owner of an account or device. The primary advantage of machine learning-based authentication is its ability to continuously learn and adapt to changes in user behavior over time, proactively boosting security. This authentication method can involve the analysis of various factors, including typing patterns, mouse movement patterns, or even the time of access to the system. Kumar, Saha, et al. (2022) suggests this approach can minimize the risk of false attacks and improve effectiveness in identifying genuine users. However, like other technological innovations, some challenges need to be addressed in the implementation of machine learning-based authentication. Special attention is required for privacy and ethics in collecting user data needed to train machine learning models. Additionally, companies and organizations must ensure adequate security against potential attacks on the machine learning model itself. It is imperative to meticulously devise and execute the entire framework of machine learning-based authentication, aiming for a harmonious equilibrium among security, privacy, and efficacy.

In the study by Bao et al. (2020) there is a system capable of automatically enhancing IoT security through a hybrid deep learning approach and limiting the communication of susceptible devices to reduce potential network harm. However, authenticating new devices to the network depends on the physical address for identification, which can be forged. Another study by (Salman et al., 2022) identifies IoT devices based on a framework, however, the present study does not address the compromise of previously authenticated devices or the re-authentication procedure when a device is reintegrated into the network. (Salman et al., 2022) focuses on the identification and detection of malicious traffic using machine learning classifiers to provide relevant security policies to devices. (Salman et al., 2022) also indicates limitations of formal authentication protocol capable of incorporating the introduced machine learning methodologies to safeguard the network against real-world attacks. The study by Ravikumar et al. (2022) explores various security challenges related to IoT using deep learning techniques, and intelligently monitoring IoT device security, they focused on using deep learning. Previous researchers (Babu & Veena, 2021; Saba et al., 2021), present classification methods with IoT device types using multiple classifiers on datasets trained with various measurable data types. They also present combined classifiers with accuracies exceeding 99%. Nevertheless, while offering exceptionally precise models for categorizing diverse IoT device types. Therefore, this research adopts a more efficient approach and proposes a formal authentication protocol that is safer and more sustainable.

2.2 Proximity

A proximity-based solution is proposed by (He et al., 2022) to enhance energy efficiency and find a balance between energy consumption and the security strength of existing proximity-based IoT device authentication protocols. Another study by Zhang et al. (2023) concentrates exclusively on a proximity-based approach involving specific physical activities performed by users. These activities include actions like bringing a smartphone closer to and moving at a distance from IoT devices and turning the smartphone for authentication. Although this study makes a noteworthy contribution to the domain of IoT device authentication, it demands a substantial level of user involvement. Furthermore, Zhang et al. (2023) primarily address the initial authentication phase and does not delve into the procedures to be followed in the event of a compromise involving an already authenticated but vulnerable device. On the other hand, (Ghose et al., 2024) suggests an alternative proximity-based solution for IoT device authentication, relying exclusively on wireless communication interfaces. The approaches seek to distinguish between authorized and unauthorized authentication requests by utilizing ambient radio signals to assess the proximity of an IoT device. Nevertheless, this solution has constraints, as it overlooks the possibility that nearby devices may be compromised due to security vulnerabilities. This scenario could enable malicious actors to execute attacks like actuation, network poisoning, and intercepting network traffic. Study by Liu et al. (2023) used techniques to identify compromised devices or not. They present a structured authentication protocol capable of utilizing the suggested solutions to authenticate IoT devices before their connection to the network. Lastly, in another study by Sobot et al. (2022) they proposed device identification based on fingerprinting the chipset of wireless devices. However, their solution is unable to identify legitimately compromised devices. Therefore, this research proposes the *High-Accuracy Device Identification and Resilience Against Credential Compromise* (HADIRACC) protocol, aiming to refine and address gaps identified in several previous studies.

3. **Proposed Model**

In the proposed system model, there are three components: the system model, cracker model, and security model. The system model consists of legitimate devices or the main device (D), ports as the central hub, and a verification server. Meanwhile, the threat model includes potential threats that could be used by adversaries to record, manipulate, and control IoT devices for malicious purposes. Additionally, the security framework of the HADIRACC protocol operates to validate devices according to the categorization of device types, ultimately concluding with the techniques and machine learning algorithms utilized in this research.

3.1 System Model

The system model for HADIRACC can be seen in Fig. 1. The system models comprise three components, akin to a network that includes IoT devices (Fig. 1), namely the device (D), Port (P), and Server Verification (S/V). Here, D represents legitimate devices that establish trust with the network using existing techniques (Ghose et al., 2024; Hansdah et al., 2022; Yu et al., 2021).

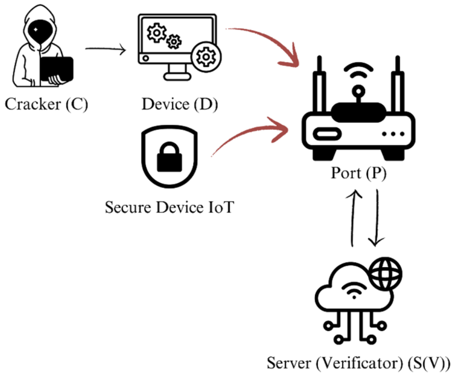

Fig. 1. Proposed and Utilized System Model in the Study (Source: Author's elaboration, 2023)

In this case, there are no limitations related to security requirements and device capabilities. Additionally, a dependable communication channel between D and S/V is assumed, which can be established using modern cryptographic techniques, ensuring authentication implementation encryption.

3.2 Cracker Model (C)

In the context of the Cracker model (C), it is presumed that it possesses the capability to compromise any of the valid devices through various means, such as exploiting firmware vulnerabilities or compromising a database of previously shared secrets (Barua et al., 2022). The Cracker (C) can leverage the acquired knowledge to target susceptible devices within the network, and manipulating devices to mimic the behavior of other device types. In this scenario, it is assumed that the Cracker (C) does not possess prior information about the traffic patterns of any legitimate compromised devices. This presumption is deemed logical, as adversaries, when scrutinizing compromised secrets, lack the ability to reach legitimate devices to capture and analyze traffic patterns. In simpler terms, the adversary remains unaware of the traffic patterns on the compromised devices.

3.3 Security Requirements

The security requirements of HADIRACC involve authenticating devices based on the classification of device types. Port (P) is tasked with validating credentials and assessing identified and observed device types through traffic pattern analysis. P and S/V can be considered a unified entity functioning as a secure gateway, handling tasks such as initial trust establishment, policy-driven network access, and ongoing device authentication and re-authentication. This is achieved using existing methods (Barua et al., 2022; Ghose et al., 2024; Hansdah et al., 2022; Yu et al., 2021). Following the user's initiation of initial trust formation in this context, it is presumed that unique credentials are subsequently assigned to each device. These credentials may be employed for subsequent authentication or to establish forthcoming security attributes like integrity verification or confidentiality. Additionally, the network has the ability to enforce access levels based on recognized vulnerabilities associated with specific device types (Barua et al., 2022). To execute this strategy on a granular level, we devised classification methods capable of distinguishing among diverse IoT device types. Vulnerabilities inherent in these device types can be extracted from databases containing known vulnerabilities, as exemplified in the research carried out by (Jeon & Kim, 2021) and utilized for policy formulation.

Ultimately, the rationale behind the third approach is to introduce supplementary modalities in both re-authentication and continuous authentication stages. Beyond credentials, the device's conduct should conform to established patterns. Collaboratively, P and SV function as an integrated and secure gateway. Furthermore, the network can implement access levels depending on vulnerabilities associated with different device types. This study developed classification techniques proficient in distinguishing various types of IoT devices, enabling vulnerability extraction from databases for customized policy implementation. Moreover, during continuous and re-authentication, additional modalities are provided. Besides credential usage, the device's behavior must align with previously documented patterns. P and S/V store traffic pattern information along with credentials, using it as a parameter in the authentication process. Consequently, malevolent devices are required not just to breach credentials but also to imitate recognized traffic patterns from compromised devices for authentication within the network.

3.4 Machine Learning Model

This research utilizes a supervised machine-learning strategy to classify the different kinds of connected IoT devices in the network. The approach incorporates four distinct machine learning algorithm models (Babu & Veena, 2021; Saba et al., 2021). The main goal of this research is to improve the accuracy of IoT device identification and optimize network performance through the implementation of sophisticated machine-learning techniques. Supervised learning in machine learning involves algorithms trained (Fig. 2). Data is processed, and divided into training and test sets, and the algorithm searches for patterns to associate labels with input data. In the prediction phase, the supervised algorithm uses the learned patterns to determine the labels of unseen test data.

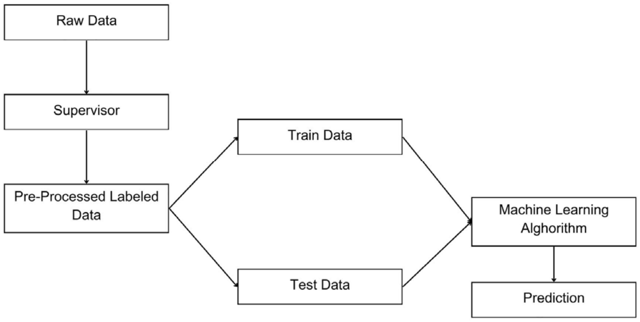

Fig. 2. Supervised Learning Model (Source: Author's elaboration, 2023)

3.4.1 Random Forest (RF)

This is a classification and regression algorithm based on the concept of ensemble learning, involving the construction of multiple decision trees during the training process, and combining the results to improve the model's performance and stability. Each tree is built randomly by selecting a random subset of features and training data. During prediction, each tree casts a vote, and the majority vote determines the result. The main advantage of Random Forest is its ability to handle overfitting issues often encountered with single decision trees. This algorithm is effective for high-dimensional data and can be used for classification and regression in various types of problems (Zhou & Wang, 2022). Random Forest is generally considered a robust and versatile model in machine learning.

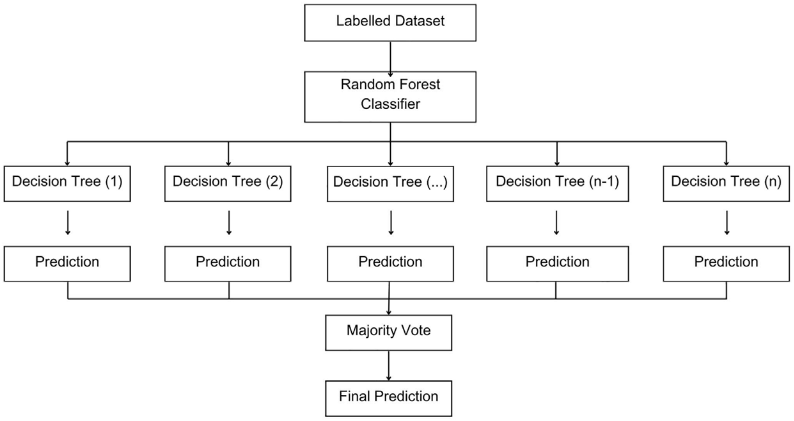

Fig. 3. Random Forest (Source: Author's elaboration, 2023)

3.4.2 K-Nearest Neighbors

This machine-learning classification algorithm functions on the principle that objects with similarities tend to be in proximity (Saba et al., 2021). In the context of classification, every data point is treated as a point within a high-dimensional space. Therefore, when classifying new data, the algorithm identifies the *K* nearest data points in proximity to it. The predominant class among these *K* nearest data points dictates the classification for the new data. The selection of the *K* value significantly impacts the model's accuracy, with a smaller *K* potentially increasing sensitivity to noise and a larger *K* reducing the model's flexibility.

3.4.3 Support Vector Machine

An algorithm within the realm of machine learning, SVM is employed for classification and regression purposes, seeking the optimal hyperplane that divides two classes within a multidimensional space. The primary objective of SVM is to identify a line or surface separator possessing the maximum margin, characterized as the maximum distance between the hyperplane and the nearest points belonging to each class. SVM is effective in handling high-dimensional data and can manage situations where classes are not linearly separable by using kernel functions to transform data into higher dimensions (Babu & Veena, 2021). SVM is known for its reliability in handling complex classification problems and its good performance in practice.

3.4.4 Gaussian Naïve Bayes

This classifier is based on Bayes' probability theorem. The method is considered “naïve” because it assumes independence between each pair of features, although there may be dependencies. Naive Bayes is suitable for data with a multitude of features and can be used for text classification, spam detection, and other applications (Babu & Veena, 2021), as it evaluates the likelihood of the target class according to input features and designates the class with the highest probability as the prediction.

3.5 HADIRACC Protocol

Before delving into the HADIRACC protocol, the device types are first classified to generate fingerprints using the features present in the packets. Once the fingerprint data is obtained, the next step is to classify the device types based on traffic patterns. Subsequently, the device types are used for additional verification during the authentication process. The devices used are classified into four types (Smart Entry, Smart Cleaning, Smart Home Appliance, and Smart Sensor). Moreover, the effectiveness and precision of the classification procedure are heightened by incorporating iterative classification techniques based on thresholds.

3.6 Fingerprint Characteristic Data

This study suggests employing an array of n packets $\{fD(1), fD(2), . . . ,fD(n)\}$ for each device (D). In this investigation, data extracted from each packet includes 19 characteristics defined as features $f(i, j)$. Only essential features are considered in this instance. Streamlining the features improves effectiveness in practical situations by minimizing the time needed for training and classifying models, as well as the necessary memory. Consequently, we compute importance scores for these features based on predictions from a basic random forest classification. Seven features with important scores exceeding 0.05, as detailed in Table 1 and grouped into four main categories, are selected. Equation 1 illustrates the seven fingerprint characteristics for the packet pD, subsequently categorized into four distinct parts.

$$f_D = \begin{bmatrix} f_D(1,1) & f_D(2,1) & \cdots & f_D(n,1) \\ f_D(1,2) & f_D(2,2) & \cdots & f_D(n,2) \\ \vdots & \vdots & \ddots & \vdots \\ f_D(1,7) & f_D(2,7) & \cdots & f_D(n,7) \end{bmatrix} \tag{1}$$

In the initialization stage, Port (P) sends the fingerprint (f_D) to S/V using a trusted channel after establishing a secure connection with S/V. Subsequently, in the initialization stage, the classification device of S/V will choose the initial classification corresponding to Port (P), so S/V will receive the fingerprint type $f_D(b, d)$ and accuracy $A(b, d)$. If $accuracy \geq S/V$, then S/V will send $f_D(i, x)$ to P, and vice versa. For the evaluation of the target classification, S/V will take the three highest accuracies $f_D(b, d)$, $f_D(c, d)$, and $f_D(e, d)$. Furthermore, S/V will evaluate f_D using 3 classifications, namely *cb, cc*, and *ce* according to the highest accuracy. If any accuracy is obtained from the classification $A(z, z) \geq {}_D$, then S/V will send the corresponding type to *P*; otherwise, the process will be repeated.

Table 1
Features with important scores selected for use in predictions

Feature	OSI Model Layer	Importance Score
`tcp.port`	Transport Layer	0.066480
`tcp.stream`	Transport Layer	0.094845
`frame.time_delta`	Physical Layer	0.096504
`ip.len`	Network Layer	0.099793
`ip.ttl`	Network Layer	0.102245
`tcp.window_size`	Transport Layer	0.125575
`frame.time_relative`	Physical Layer	0.163713

Device classification is divided into 4 types: smart home, smart cleaning, smart home appliance, and smart sensor. The rationale for choosing these categories lies in the fact that cameras and sensors acquire information, while home assistants can execute actions. Consequently, this choice facilitates effective establishment policies, preventing devices focused on information gathering from executing actions. Moreover, we distinguish between information-collecting devices, specifically cameras and sensors, recognizing their varying degrees of privacy invasion during data collection. The effectiveness and accuracy of the classification procedure are improved by employing iterative classification techniques based on thresholds. Initially, the dataset is split into 70% (train) and 30% (test). Subsequently, five unique models are individually trained by associating each with the training data.

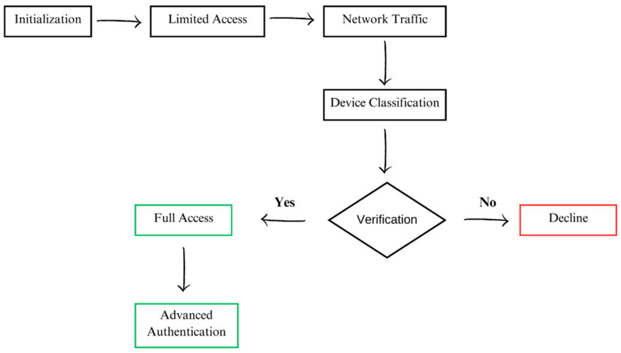

Fig. 4. HADIRACC Protocol (Source: Author's elaboration, 2023)

The classification procedure commences with initialization, during which the secure connection established transmits fingerprints to the verification center through the trusted channel. Next, the initial classifier is evaluated until the accuracy of the verifier type is obtained. Finally, the second classification process is performed to select the highest accuracy. Figure 4 illustrates the HADIRACC protocol. The HADIRACC protocol begins with Initialization, where the user must undergo initial authentication to transmit data to the port (P), followed by Limited Access. The device subsequently captures traffic to initiate a network connection and send data to the cloud service. The next step is the classification of the device type through the previously authenticated trusted channel to retrieve fingerprint data based on identity. Subsequently, the verification of the device type is performed; if successful, full access is granted, if not successful, access is denied and the process returns to the initial step. Once full access is granted, additional authentication can be carried out for each IoT device in the network, ensuring the security of the device network can be controlled.

4. Security Analysis of HADIRACC

First, the device to be used must be authenticated to establish the initial connection between the device and the central port (P). Through the utilization of classification techniques, the HADIRACC protocol can identify both known and foreign devices. HADIRACC addresses attackers (C-crackers) who may target devices through the network in various ways, similar to the methods employed by attackers in the study (Kumar, Abhishek, et al., 2022; Troscia et al., 2022). HADIRACC detects threats and attacks using classifications stored in the database, where the classifier has been trained using authenticated fingerprints from old devices. When a machine learning model is tasked with making predictions based on a dataset resembling its training data, it delivers remarkably precise classifications. In the event of a device compromise, alterations in the content of data packets occur, and the identification of these changes can be accomplished by comparing them with the classifications in the database, as the fingerprints will deviate from those stored in the database. Finally, after SV sends the results to P, P will decide on access and only provide restricted access to the network devices. Additionally, previously authenticated credential devices will be removed from the authentication list, and if re-authenticated at some point, the device needs to be reintroduced to the network. If the device stays compromised throughout the re-authentication procedure, SV will struggle to furnish precise classification for the device due to disparities in the device's fingerprint. For example, if a Smart Lock door IoT is compromised, the SV classification will not categorize the device because the fingerprint of the compromised device differs from the fingerprint of the Smart Lock door IoT used to train the classification.

4.1 Analysis of Classification Techniques

Unlike studies (Ghose et al., 2024; Hansdah et al., 2022; Kumar, Abhishek, et al., 2022; Troscia et al., 2022; Yu et al., 2021) that only identify new devices based on the physical address where attackers can easily spoof the PA of authenticated devices and authenticate themselves with port (P), in the approach of this research beyond prior authentication, the device will be classified every time it needs to be re-authenticated with the server. The use of physical address in this research is to check the presence of the device in the database. Thus, if the Physical Address is manipulated, the S/V will face challenges in achieving the highest classification accuracy, as the device's fingerprint won't correspond to the information stored in the database. This strategy effectively addresses compromised devices by considering traffic patterns. Consequently, the S/V (Verifier) can automatically relay this confirmation to the port (P) that while the PA matches, the fingerprint does not, designating the device as unauthorized and preventing its connection. This analytical approach depends on the concept that a machine learning model, when trained and assessed using the same dataset or a similar one, should generate predictions with the utmost accuracy. This rationale underscores the importance of capturing the traffic of new devices, training the model exclusively on the fingerprints of authenticated devices post-authentication, and storing that refined model in the database.

5. Implementation

The protocol addresses three scenarios for device authentication within a network. First, for a new device, its fingerprint and network details are collected, and based on device type classification, it is granted full network access. The fingerprint is stored for future classifications. The second situation entails a previously verified device transitioning in and out, utilizing its fingerprint and MAC address for the re-authentication process. The verifier classifies the device type and notifies the hub for access decisions. Lastly, devices constantly inside the network require continuous authentication, similar to re-authentication. Authentication correlates MAC addresses with stored ones and utilizes classification techniques. The protocol ensures security by detecting compromised devices through classification models, preventing unauthorized actions. However, compromise detection occurs during the next re-authentication, allowing a window for malicious activities. Security administrators can control re-authentication frequency based on device and network security requirements.

5.1 Dataset, Device and Data Processing

The communication between the access points and the devices listed in Table 2 was derived from information gathered through a dedicated platform provided by a specific brand specializing in smart home and security products. More than 70 collected data were IoT-based devices, but they were further selected and filtered to obtain approximately 50 devices for use in the study (Table 2). The subset of collected data was made open source, and several devices and several relevant data were extracted into a .csv file. Subsequently, fingerprint matrices were created for classifying devices based on their respective types into different class groups for the training and testing processes of the model. The process of identifying devices involved the utilization of device fingerprints extracted from the devices listed in Table 2. This information will be employed to construct a scalable machine-learning model, facilitating seamless training of new categories of IoT devices from the dataset. The combined classifier and directed classifier will be trained using device packet captures, enhancing the adaptability and efficiency of the entire device identification system.

Table 2
Dataset - List of IoT devices according to their classifications (Source: (*EZVIZ - Creating Easy Smart Homes*, n.d.)

Device Name	Category	Class
Battery Power Video Doorbell	Smart Entry	1
Battery Power Video Doorbell Kit	Smart Entry	1
Front Door Protection	Smart Entry	1
Front Door Protection 2K resolution	Smart Entry	1
Smart Knock Door	Smart Entry	1
Wi-Fi Video Doorbell	Smart Entry	1
Wi-Fi Video Doorbell Plus	Smart Entry	1
Smart Wi-Fi Chime	Smart Entry	1
Video Doorbell Companion	Smart Entry	1
Home Security System	Smart Entry	1
Home Security System Plus	Smart Entry	1
Smart Home Video Door phone	Smart Entry	1
Wire-free Peephole Doorbell	Smart Entry	1
Wire-free Peephole Doorbell 2K	Smart Entry	1
Smart Fingerprint Lock Non-WIFI	Smart Entry	1
Smart Fingerprint ZigBee Version	Smart Entry	1
Smart-Lock	Smart Entry	1
Smart Fingerprint Keyless	Smart Entry	1
Vacuum & Mop Combo	Smart Cleaning	2
Elevated & Simplify Cleaning	Smart Cleaning	2
Elevated & and Simplify Cleaning every day	Smart Cleaning	2
Easy Cleaning Plus	Smart Cleaning	2
Effortless Cleaning Vacuum	Smart Cleaning	2
Easy auto-cleaning Vacuum	Smart Cleaning	2
Light and Easy auto Cleaning Vacuum	Smart Cleaning	2
Self-Cleaning Vacuum	Smart Cleaning	2
Wet & Dry-Cleaning Vacuum	Smart Cleaning	2
Smart Security Wall-Light Camera	Smart Home Appliance	3
Smart CCTV with Lamp	Smart Home Appliance	3
Smart Plug 10B	Smart Home Appliance	3
Smart Plug 10A	Smart Home Appliance	3
Dimmable Wi-Fi LED Bulb Color	Smart Home Appliance	3
Dimmable Wi-Fi LED Bulb White	Smart Home Appliance	3
UV-C Air Purifier	Smart Home Appliance	3
Portable Power Station-Plenty	Smart Home Appliance	3
Portable Power Station-Robust	Smart Home Appliance	3
Portable Power Station-Grab and Go	Smart Home Appliance	3
Portable Solar Panel	Smart Home Appliance	3
Power Station	Smart Home Appliance	3
2 in-One Outdoor Security	Smart Home Appliance	3
Smart Control Aluminate	Smart Home Appliance	3
Smart Control Wi-Fi Relay	Smart Home Appliance	3
Smart Control Wi-Fi Relay-Multiple Safety	Smart Home Appliance	3
Smart Radiator Thermostat	Smart Home Appliance	3
Smart Plug Power Consumption Tracker	Smart Home Appliance	3
Fast Stable Wi-Fi	Smart Home Appliance	3
Temperature and Humidity Sensor	Home Sensors	4
T3C Smart Button	Home Sensors	4
T2C Open Close Sensor	Home Sensors	4
T1C PIR Motion Sensor	Home Sensors	4
Home Gateway	Home Sensors	4
EZVIZ 4-Piece Sensor	Home Sensors	4
Smart Siren	Home Sensors	4
Water Leak Sensor	Home Sensors	4
Home Gateway Apple	Home Sensors	4

Based on the research approach (Babu & Veena, 2021; Saba et al., 2021), this study employs four classifiers (RF, SVM, KNN, and GNB). This research builds upon previous studies classifying IoT devices and enhances the accuracy and efficiency of the classification process. Unlike previous studies that classified devices as either IoT or non-IoT, this research goes a step further by classifying IoT devices, such as smart CCTV, and smart door locks. The data used for classification undergoes several data processing techniques to ensure optimal classification results. The initial stage involves data cleaning and splitting, where irrelevant packets such as outbound traffic from the access point are removed, and empty columns that do not contribute are deleted. The database in this study is divided into two parts, with the first 70% allocated for training and the remaining 30% for testing. The next process includes feature standardization using the scikit.learn (Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.3.2 Documentation., n.d.) standard method to improve classifier accuracy. To address missing values in features, numeric imputation with median values is employed. Feature engineering is conducted by applying Random Forest to extract feature importance scores, and features with scores below the threshold of 0.05 are removed. Finally, features considered irrelevant, such as TCP. urgent pointer, and others, are eliminated from the dataset to enhance runtime efficiency, memory usage, and classifier accuracy.

5.2 Algorithm Selection and Data Pre-processing

Based on the approach of previous researchers (Babu & Veena, 2021; Saba et al., 2021), four classifiers (random forest, KNN, SVM, and Gaussian) are utilized in this study to enhance more efficient classification. Firstly, the dataset with irrelevant data points is converted to a csv file by removing packets with unnecessary source Ethernet addresses in the model used, encompassing all outbound traffic from the access point, as it avoids the need for classification and could potentially introduce bias to the classifier predictions due to the abundance of such packets in the data. Additionally, empty columns representing null values for features are eliminated as they hold no significance in the classification process. Subsequently, the complete dataset is categorized into distinct classes for both training and testing purposes. The data in CSV format is imported into a data table, and the labels are obtained from the dataset by dividing the data into A (features) and B (labels). Afterward, both A and B are randomly divided into training and testing datasets, with a distribution ratio of 70% and 30%, respectively.

5.3 Training and Testing

Following the collection and preprocessing of the dataset, 70% of the dataset is allocated for the individual training of each of the four classifiers. The training procedure involves utilizing the fit method from sklearn (Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.3.2 Documentation., n.d.), which aligns the model with the dataset and subsequently generates predictions. Throughout the training phase, the time taken for the complete training of each model is documented and detailed in Table 3.

Table 3
Training and testing time for each machine learning model

Classifier Model	Result for Training time	Result for Testing time
Random Forest	14.11 second	0.41 milliseconds
K-Nearest Neighbors	0.22 second	0.77 milliseconds
Support Vector Machine	98 minutes	2.81 milliseconds
Gaussian Naïve Bayes	40 milliseconds	0.18 milliseconds

In the testing phase, the pre-trained models were utilized to predict labels for the test data, and a comparison between the predicted labels and an analysis of the outcome labels was performed to calculate the classification accuracy provided by each model. The assessment employed the function for measuring accuracy available in the library (Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.3.2 Documentation, n.d.).

5.4 Classification Result

research focuses on the classification of data using four different machine learning models, to evaluate and compare the performance of each model in generating accurate predictions. Through this approach, we strive to gain a deeper insight into the strengths and weaknesses of each model, as well as to understand the context in which each model can deliver the best results. Obtaining a better understanding of the performance of these machine learning models is expected to contribute significantly to the development of more effective solutions in classifying data across various application fields.

3.1 F1 Score

In the assessment of classification, three primary metrics employed to gauge the model's performance include F1 Score, Recall (Sensitivity), and Precision. F1 Score, as a combined metric of Precision and Recall, provides a holistic overview of the model's ability to classify data, particularly useful when there is an imbalance between classes or when performance on a specific class cannot be sacrificed. Conversely, precision measures how well the model can identify positive instances without providing significant false positive results. To calculate F1, the formula used is as follows:

$$F1\ Score = 2\times\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4}$$

Table 4
F1 Score results of the algorithms in the classes used in the study

Class	Recall	Precision	F1 Score
Class 1	0.94	0.84	0.888
Class 2	0.88	0.88	0.886
Class 3	0.90	0.81	0.856
Class 4	0.87	0.77	0.822

The F1 score in this study was calculated using the scikit-learn library for metric evaluation. Class 1-4 represents Smart Entry, Smart Cleaning, Smart Home Appliances, and Home Sensors.

5.7 Accuracy Score

In this study, the model's performance was evaluated using accuracy as the primary indicator. Accuracy is the proportion of accurate predictions to the overall number of predictions conducted. The formula used to calculate overall accuracy is:

$$\text{Overall accuracy} = \frac{\text{Number of correct predictions}}{\text{True number of predictions}} \times 100 \qquad (5)$$

After the model was trained, we made predictions on the test data and contrasted them with the genuine labels to compute the accuracy score. The total accuracy score (overall accuracy score) is 89%, indicating that the model performs well in making predictions overall. However, it should be noted that the results for some classes may be lower, suggesting potential improvement in those specific classifications. While the overall accuracy score is satisfactory, focusing on improving performance in specific classes will enhance the model's applicability in a broader context. After the calculations, it was found that the F1 score for the Random Forest Classifier is 88.8%, and it is determined to be the most accurate model with the highest accuracy approaching 89%. The next classifiers in order are the K-Nearest Neighbors Classifier with an accuracy of 88.6%. Following in sequence, the subsequent classifier is the SVM, achieving an accuracy of 85.6%, and the final classifier in the lineup is the GNB, boasting an accuracy of 88.2%. These results are illustrated in Fig. 5.

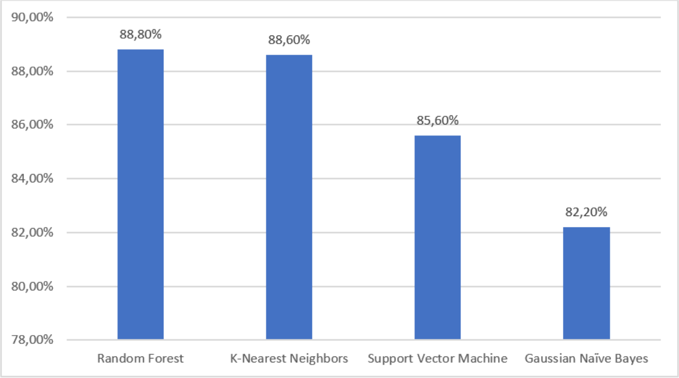

Fig. 5. Results of Accuracy Scores for each Classifier Model used in the study

6. Discussion

The proposed HADIRACC protocol focuses on high-accuracy device identification and resilience against credential compromise. This protocol leverages collaborative functions between Port (P) and Security Validator (S/V) to serve as a secure integrated gateway. Through the utilization of classification methods capable of discerning diverse types of IoT devices, it becomes possible to extract vulnerabilities linked to various device types from a database, facilitating the implementation of tailored policies. Additionally, the protocol provides additional modalities during continuous authentication to ensure device behavior aligns with documented patterns, enhancing security measures. Compared to previous research (Barua et al., 2022; Ghose et al., 2024; Hansdah et al., 2022; Yu et al., 2021), the HADIRACC protocol stands out by using supervised machine learning algorithms to categorize IoT devices accurately. This approach aims to improve device identification accuracy and optimize network performance through sophisticated machine learning techniques. The protocol's emphasis on continuous authentication and the incorporation of known device vulnerabilities for policy enforcement sets it apart from conventional security measures. Furthermore, the HADIRACC protocol addresses the threat of credential compromise by necessitating malicious entities to not only breach credentials but also replicate recognized traffic patterns from compromised devices to authenticate on the network. This layered approach adds an extra level of security, making it difficult for attackers to infiltrate the network. Moreover, the research methodology involves preprocessing datasets, converting irrelevant data points into CSV files, and using supervised learning algorithms to train and test machine learning models. Training and testing times for each classification are recorded, demonstrating the efficiency and effectiveness of this approach, which is quite superior compared to (Barua et al., 2022; Yu et al., 2021). In summary, the HADIRACC protocol presents a comprehensive and innovative solution to enhance IoT device security through advanced machine learning techniques, continuous authentication, and resilience against credential compromise. By combining known device vulnerabilities and leveraging collaborative functions between P and S/V, this protocol offers a robust security framework for IoT networks. Compared to previous research, the HADIRACC protocol introduces a more sophisticated approach to IoT security by combining machine learning algorithms for device identification and authentication. Traditional security protocols may lack the adaptability and accuracy provided by machine learning models, making them more vulnerable to attacks. The emphasis of the HADIRACC protocol on continuous authentication and leveraging known device vulnerabilities for policy enforcement sets it apart from conventional security measures, offering a more proactive and dynamic security solution for IoT environments.

7. Conclusion and Recommendations

This research has yielded an innovative security protocol known as High-Accuracy Device Identification and Resilience Against Credential Compromise (HADIRACC). The protocol aims to create a more robust security layer in the IoT ecosystem, reducing the risk of attacks and enhancing user trust in IoT technology. Additionally, the research proposes a proximity-based solution for authenticating IoT devices with a focus on energy efficiency and a balance between energy consumption and the security strength of proximity-based IoT device authentication protocols. Regarding the classification of IoT devices, this research has been successful in classifying IoT devices with a high level of accuracy. Previous studies have presented classification methods for IoT device types using multiple classifiers on datasets trained with various measurable data types. This research employs four machine learning classifiers with algorithm calculations and testing using scikit-learn. The classifiers used in this study are Random Forest, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes to identify the accuracy of IoT devices and classify types of IoT devices. Among all the models, the classifier that can make the most accurate predictions with the highest average F1 score in this research is Random Forest, with a score of 88.8%. In this study, the use of time in model training and the average time for classifier classification indicates that the Gaussian Naive Bayes classifier is the fastest for training and testing. However, it does not provide sufficiently accurate results. For this reason, this research leans towards using the Random Forest classifier, even though it takes slightly longer than Gaussian Naive Bayes for training and testing, the results it provides are the most accurate.

7.1 Future Recommendation

To address the limitations and remaining issues in the proposed protocol in this study, future research should consider several improvement suggestions. Firstly, the research can be expanded by involving model training for various types of IoT devices using larger datasets. This more comprehensive data collection will provide a stronger foundation for machine learning and enhance the model's ability to recognize differences between devices. Secondly, an essential measure involves incorporating supplementary features from packet captures corresponding to each device type. This additional information can offer deeper insights into the unique characteristics of each device, resulting in more accurate and reliable classification. Additionally, the research could explore other variations of classifiers besides those already used, aiming to increase model diversity. This approach may involve combinations of different classifiers to gain a more holistic view of the dataset, optimizing model performance, and minimizing potential bias or imbalance in classification. Lastly, future research should integrate model refinement techniques, such as fine-tuning or regression techniques, to ensure that the classification model continues to adapt to changes in the IoT environment. Thus, it can be expected that the proposed protocol will become stronger, more reliable, and capable of addressing challenges that may arise in the evolution of IoT technology.

References

Babu, M. R., & Veena, K. N. (2021). A Survey on Attack Detection Methods For IOT Using Machine Learning And Deep Learning. *2021 3rd International Conference on Signal Processing and Communication (ICPSC),* 625–630. https://doi.org/10.1109/ICSPC51351.2021.9451740

Bao, J., Hamdaoui, B., & Wong, W.-K. (2020). IoT Device Type Identification Using Hybrid Deep Learning Approach for Increased IoT Security. *2020 International Wireless Communications and Mobile Computing (IWCMC),* 565–570. https://doi.org/10.1109/IWCMC48107.2020.9148110

Barua, A., Al Alamin, M. A., Hossain, M. S., & Hossain, E. (2022). Security and Privacy Threats for Bluetooth Low Energy in IoT and Wearable Devices: A Comprehensive Survey. *IEEE Open Journal of the Communications Society, 3*, 251–281. https://doi.org/10.1109/OJCOMS.2022.3149732

Chataut, R., Phoummalayvane, A., & Akl, R. (2023). Unleashing the Power of IoT: A Comprehensive Review of IoT Applications and Future Prospects in Healthcare, Agriculture, Smart Homes, Smart Cities, and Industry 4.0. *Sensors, 23*(16), 7194. https://doi.org/10.3390/s23167194

EZVIZ - Creating Easy Smart Homes. (n.d.). Retrieved December 31, 2023, from available: https://www.ezviz.com/id

Ghose, N., Gupta, K., Lazos, L., Li, M., Xu, Z., & Li, J. (2024). ZITA: Zero-Interaction Two-Factor Authentication using Contact Traces and In-band Proximity Verification. *IEEE Transactions on Mobile Computing,* 1–16. https://doi.org/10.1109/TMC.2023.3321514

Hansdah, R. C., Jamwal, J., & Gudivada, R. B. (2022). Dragonshield : An Authentication Enhancement for Mitigating Side-Channel Attacks and High Computation Overhead in WPA3-SAE Handshake Protocol. *Proceedings of the 23rd International Conference on Distributed Computing and Networking,* 188–197. https://doi.org/10.1145/3491003.3491021

He, Y., Zeng, K., Mark, B. L., & Khasawneh, K. N. (2022). Secure and Energy-Efficient Proximity-Based Pairing for IoT Devices. *2022 IEEE Globecom Workshops (GC Wkshps)*, 1359–1364. https://doi.org/10.1109/GCWkshps56602.2022.10008568

Jeon, S., & Kim, H. K. (2021). AutoVAS: An automated vulnerability analysis system with a deep learning approach. *Computers & Security, 106,* 102308. https://doi.org/10.1016/j.cose.2021.102308

Kumar, A., Abhishek, K., Ghalib, M. R., Shankar, A., & Cheng, X. (2022). Intrusion detection and prevention system for an IoT environment. *Digital Communications and Networks, 8*(4), 540–551. https://doi.org/10.1016/j.dcan.2022.05.027

Kumar, A., Saha, R., Conti, M., Kumar, G., Buchanan, W. J., & Kim, T. H. (2022). A comprehensive survey of authentication methods in Internet-of-Things and its conjunctions. *Journal of Network and Computer Applications, 204*, 103414. https://doi.org/10.1016/j.jnca.2022.103414

Liu, P., Ji, S., Fu, L., Lu, K., Zhang, X., Qin, J., Wang, W., & Chen, W. (2023). How IoT Re-using Threatens Your Sensitive Data: Exploring the User-Data Disposal in Used IoT Devices. *2023 IEEE Symposium on Security and Privacy (SP)*, 3365–3381. https://doi.org/10.1109/SP46215.2023.10179294

Omolara, A. E., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., Alshoura, W. H., & Arshad, H. (2022). The internet of things security: A survey encompassing unexplored areas and new insights. *Computers & Security, 112*, 102494. https://doi.org/10.1016/j.cose.2021.102494

Ravikumar, K. C., Chiranjeevi, P., Manikanda Devarajan, N., Kaur, C., & Taloba, A. I. (2022). Challenges in internet of things towards the security using deep learning techniques. *Measurement: Sensors, 24*, 100473. https://doi.org/10.1016/j.measen.2022.100473

Rawat, A. (2022). Recent Trends in IoT : A review. *Journal of Management and Service Science (JMSS), 2*(2), 1–12. https://doi.org/10.54060/jmss.v2i2.21

Saba, T., Haseeb, K., Shah, A. A., Rehman, A., Tariq, U., & Mehmood, Z. (2021). A Machine-Learning-Based Approach for Autonomous IoT Security. *IT Professional, 23*(3), 69–75. https://doi.org/10.1109/MITP.2020.3031358

Salman, O., Elhajj, I. H., Chehab, A., & Kayssi, A. (2022). A machine learning based framework for IoT device identification and abnormal traffic detection. *Transactions on Emerging Telecommunications Technologies, 33*(3). https://doi.org/10.1002/ett.3743

Scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation. (n.d.). Retrieved December 29, 2023, from available: https://scikit-learn.org/stable/

Sobot, S., Ninkovic, V., Vukobratovic, D., Pavlovic, M., & Radovanovic, M. (2022). Machine Learning Methods for Device Identification Using Wireless Fingerprinting. *2022 International Balkan Conference on Communications and Networking (BalkanCom)*, 183–188. https://doi.org/10.1109/BalkanCom55633.2022.9900723

Tomer, V., & Sharma, S. (2022). Detecting IoT Attacks Using an Ensemble Machine Learning Model. *Future Internet, 14*(4), 102. https://doi.org/10.3390/fi14040102

Troscia, M., Sgambelluri, A., Paolucci, F., Castoldi, P., Pagano, P., & Cugini, F. (2022). Scalable OneM2M IoT Open-Source Platform Evaluated in an SDN Optical Network Controller Scenario. *Sensors, 22*(2), 431. https://doi.org/10.3390/s22020431

Yu, J., Lian, H., Zhao, Z., Tang, Y., & Wang, X. (2021). Provably secure verifier-based password authenticated key exchange based on lattices (pp. 121–156). https://doi.org/10.1016/bs.adcom.2020.09.003

Zhang, Y., Han, D., Li, A., Li, J., Li, T., & Zhang, Y. (2023). SmartMagnet: Proximity-Based Access Control for IoT Devices With Smartphones and Magnets. *IEEE Transactions on Mobile Computing, 22*(7), 4266–4278. https://doi.org/10.1109/TMC.2022.3149746

Zhou, L., & Wang, H. (2022). A Combined Feature Screening Approach of Random Forest and Filterbased Methods for Ultra-high Dimensional Data. *Current Bioinformatics, 17*(4), 344–357. https://doi.org/10.2174/1574893617666220221120618