Subsystem of anomaly detection in the Smart House system based on machine learning

Maxim Prodeus[1,*,†], Andrii Nicheporuk[1,†], Andrzej Kwiecien[2,†], Dmytro Martiniyuk[1,†], Oleksii Lyhun[2,†]

[1] *Khmelnytskyi National University, Institutska str., 11, Khmelnytskyi, 29016, Ukraine*
[2] *Silesian University of Technology, Akademicka str., 2A, Gliwice, Poland*

Abstract

With the deepening implementation of Smart Home systems, the role of anomaly detection subsystems becomes increasingly important for ensuring the security and stability of these complex environments. This paper proposes a new anomaly detection subsystem for Smart Home systems, based on advanced machine learning technologies. The architecture of this subsystem is designed to process various data streams generated by Internet of Things (IoT) devices, utilizing packet preprocessing to optimize data before further analysis. The application of the Random Forest algorithm allows for the construction of a machine learning model for effective anomaly detection in the system. To evaluate the effectiveness of the proposed subsystem, the CICIDS2017 dataset is utilized, which is divided into training and validation sets. Comparative analysis is conducted with the J48 tree algorithm in detecting various types of cyberattacks, such as Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). The proposed subsystem aims to enhance the security and reliability of Smart Home systems by facilitating timely detection and response to potentially dangerous anomalies.

This work represents a significant contribution to the field of smart systems as it addresses the security issue within the smart home environment, where a large number of connected devices are typically characterized by limited resources and increased requirements for confidentiality and availability. The application of machine learning methods for anomaly detection enables the identification of unusual and potentially hazardous interactions between devices and the network, indicating attacks or security breaches.

Particular attention should be paid to the experiment results, which demonstrated the high effectiveness of the proposed system compared to traditional methods. Anomaly detection using the Random Forest algorithm proved to be effective in various attack scenarios, providing high accuracy and a low error rate. This suggests the potential use of such approaches for protecting smart systems in the future.

Keywords

Smart house, network security, anomaly detection, threat detection, cyber defense, random forest

1. Introduction

The Smart Home systems have seen rapid adoption in recent years, offering homeowners convenience, energy efficiency, and increased security. However, the proliferation of interconnected Internet of Things devices in these environments also creates vulnerabilities and potential security threats. Anomaly detection subsystems play a crucial role in identifying and

ICyberPhyS-2024: 1st International Workshop on Intelligent & CyberPhysical Systems, June 28, 2024, Khmelnytskyi, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

mprodeus99@ukr.net (M. Prodeus); andrey.nicheporuk@gmail.com (A. Nicheporuk); andrzej.kwiecien@polsl.pl (A. Kwiecien); martiniyuk.dim14@gmail.com (D. Martiniyuk); andrzej.kwiecien@polsl.pl (O. Lugyn)

0009-0002-2968-4648 (M. Prodeus); 0000-0002-7230-9475 (A. Nicheporuk); 0000-0003-1447-3303 (A. Kwiecien); 0009-0002-3524-872X (D. Martiniyuk); 0000-0003-1447-3303 (O. Lugyn)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3736/paper25.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

mitigating these threats, allowing for proactive responses to malicious activities or system anomalies [1]. This document presents a comprehensive approach to anomaly detection in smart homes using machine learning methods for network traffic analysis and detection of suspicious patterns [2]. The architecture of the proposed subsystem, along with the utilization of the Random Forest algorithm, is detailed. Additionally, experimental methodologies and evaluation metrics are described to assess the system's effectiveness in detecting various types of cyberattacks [3].

With the increasing number of devices connected to the network and the volume of generated data, there is a significant need for reliable and efficient anomaly detection methods [4, 5]. Traditional methods such as signature analyzers and rules often struggle to effectively cope with complex and evolving threats, necessitating enhanced approaches. The developed subsystem serves as a response to this challenge, enabling cybersecurity systems to adapt to changing conditions and detect anomalous behavior patterns that may indicate potential cyber threats. In the context of cybersecurity, anomalies can take various forms: from unjustified user activity and attacks on infrastructure to suspicious transactions and leaks of confidential information. Detecting these anomalies is a challenging task that is difficult to address without the use of automated and intelligent tools. The main advantages of using this method in cybersecurity are its ability to operate in real-time, identify new forms of threats, and adapt to changes in the cyber landscape [6]. The core principle of the subsystem's operation is based on the idea of using machine learning algorithms to recognize and learn the normal behavior pattern of the system or user. Instead of the traditional approach, where it is assumed that the form of a malicious attack is known, the subsystem adopts an intuitive approach: it learns the normal state and then detects anomalies that deviate from it.

A subsystem specifically designed for deployment in the Internet of Things (IoT) environment is also introduced. IoT is a concept that involves connecting various physical objects to the network for data exchange and process automation. This technology finds applications in various sectors, from household devices to industrial systems [7].

As the number of connected devices grows, the issue of cybersecurity in IoT becomes increasingly relevant. One of the most common threats is Distributed Denial of Service (DDoS) attacks, aimed at overloading network resources [8].

Mitigating DDoS in IoT is extremely important because a large number of connected devices can be used to create a botnet, which in turn can easily initiate DDoS attacks. This can lead to service disruptions in critical systems such as medical devices, power systems, or automotive transportation networks. Securing IoT involves using measures to detect and suppress DDoS attacks. Utilizing anomaly detection systems, traffic restriction from suspicious sources, and implementing authentication mechanisms are key components of DDoS protection in IoT [9].

Coordinated security measures in IoT not only help maintain the functionality of systems but also prevent potential consequences of attacks on critical infrastructure [10]. Thus, effective DDoS mitigation in IoT is a necessary component to ensure stability and security in the modern connected world.

Despite the significant potential of the subsystem, there are certain challenges. One of them is the need for a large amount of data for effective algorithm training. Collecting and processing this data can be resource-intensive [11]. Additionally, algorithms' high sensitivity to changes in input data may lead to false positives. In the future, the development of the subsystem will be associated with improving algorithms for automatic anomaly detection in real-time, developing methods to reduce false positives, and expanding its application in other areas [12]. Enhancing measures to ensure confidentiality and ethics in the use of the subsystem, especially in processing personal data, is also important.

In a world of constantly evolving threats and rapidly changing technologies, machine learning for anomaly detection proves to be an indispensable tool for cybersecurity and identifying unknown threats. Its capabilities in detection and adaptation, application in various sectors, and potential to address complex issues make this subsystem a key element in modern security and protection strategies. Engaging intelligent systems based on machine learning is a crucial step in ensuring the resilience and efficiency of information systems in response to constantly growing cybersecurity threats [13].

The objective of this work is to develop and present a comprehensive approach to anomaly detection in smart home environments and the Internet of Things (IoT) using machine learning methods. Specifically, the focus is on designing a subsystem capable of identifying and mitigating potential cybersecurity threats, including Distributed Denial of Service (DDoS) attacks, within these interconnected systems. The aim is to enhance the security and resilience of smart homes and IoT devices by implementing intelligent anomaly detection mechanisms.

2. Related works

The study and application of machine learning methods for anomaly detection in modern cybersecurity and other fields of scientific research have garnered significant interest and are aimed at addressing pertinent challenges in security, medicine, finance, and beyond [14]. Below is a review and analysis of some key works investigating the application of machine learning for anomaly detection and their contributions to the development of respective fields [15].

In "Detection of DNS DDoS Attacks with Random Forest Algorithm on Spark" by Liguo Chen, Yuedong Zhang, Qi Zhao, Guanggang Geng, ZhiWei Yan, the authors conducted a review of anomaly detection methods focused on cybersecurity [16]. They explore various approaches and algorithms, such as the Random Forest method, statistical approaches, and deep learning methods.

The work represents an important scientific research in the field of detecting DDoS attacks. The aim of the work is not only to indicate whether a DNS server is under attack, but also to distinguish normal requests from abnormal ones. Some additional features are also added to distinguish unusual domain names from regular domain names.

The authors selected features based on which anomaly activity recognition in the data stream would be performed. After that, in the experiment process, three main steps were executed. Firstly, data preprocessing was carried out to normalize the data and give the subsystem the ability to perceive values. Secondly, the subsystem was trained to differentiate between normal and anomalous data. And thirdly, they characterized the threat assessment system, which is based on the selection made by Random Forest.

This work presents a new method for reducing DDoS traffic on TLD servers, where traffic filtering based on machine learning algorithms is applied to the core recursive DNS servers on the Internet. Their classification model is built on Spark and operates with a 0.0% FPR and 4.36% FNR, meaning that practical requirements for accuracy and performance are met. In future work, the authors plan to apply a streaming approach, which is more suitable for real-time rule creation by the firewall.

In "Transport in the IP-based Internet of Things: status report" by J. Antonio Garcia-Macias, the focus is on IoT networks using IP at the network level [17]. A simplified IP-based IoT stack is depicted in comparison to the traditional Internet protocol stack. Being network-agnostic in its scheme, IP does not make assumptions about underlying physical layers and data link layers. The development of IP-based IoT networks has been focused on IEEE 802.15.4 at the physical and data link layers, as well as the use of Bluetooth Low Energy (BLE) system.

Other technologies such as ITU-T G.9959, DECT ULE, MS/TP, NFC, IEEE 1901.2, and IEEE 802.11ah have been studied by the author for their utilization in IPv6-based IoT networks. Many application protocols have been developed considering IoT scenarios. For instance, the Constrained Application Protocol (CoAP) has been proposed as an alternative to the Hypertext Transfer Protocol (HTTP); Message Queuing Telemetry Transport (MQTT) and Advanced Message Queuing Protocol (AMQP) are used for message queue-based applications; and Lightweight Machine-to-Machine (LwM2M) has been proposed as a solution for device management and service provisioning.

In the work "A Survey of Anomaly Detection in Internet of Things," the utilization of anomaly detection methods in Internet of Things (IoT) networks is investigated [18]. Authors Moustafa, N. and Slay, J. examine the characteristics of IoT that pose challenges for anomaly detection and present various methods, such as statistical methods, machine learning-based methods, and deep learning, that can be applied for effective anomaly detection in the context of IoT.

The work represents a review of anomaly detection methods in the Internet of Things (IoT). Nurul Moustafa and Jillian Slay systematically explore a wide range of approaches and methods used for anomaly detection in complex IoT systems.

The work begins with defining the concept of IoT and its importance in the modern world. The authors emphasize the diversity and scope of interpretations of IoT, which encompass various types of devices, communication protocols, and applications ranging from smart homes to industrial systems. In their further investigation, the authors delve into the challenges and issues associated with ensuring security in IoT, particularly anomaly detection. They analyze the characteristics of IoT data such as large volume, diversity, and dynamism, which complicate the task of anomaly detection.

3. Architecture of the anomaly detection subsystem in the Smart House system based on machine learning. Preprocessing of IoT traffic packets.

In modern Smart Home systems, where a large number of various sensors and devices provide a continuous flow of data, anomaly detection becomes a critically important task to ensure the security and efficiency of system operation [19]. Therefore, an integrated anomaly detection subsystem based on machine learning has been proposed for anomaly detection in the Smart Home system (see Fig. 1). Its main function is to analyze network traffic and data received from sensors in the smart home and apply machine learning algorithms to detect abnormal behavioral patterns. Control influences are implemented through the microcontroller system of the Smart Home. Although threats to IoT may also include metamorphic viruses [20], one subsystem alone cannot fully cope with all aspects of the system. Therefore, interaction with other subsystems working in different protocols and planes [21] is not excluded. Only in such a case can protection against harmful influences be maximized.

Along with the anomaly detection subsystem, the proposed architecture includes a notification subsystem and a data recording and logging subsystem. The data recording and logging subsystem provide the ability to recover event histories for further analysis and anomaly detection. These data can be useful for comparison when detecting subsequent anomalies and for preserving formed behavioral patterns. To alert the user about potential cyber attacks, a notification subsystem is included, which generates messages in the form of sending an email and SMS notification. A local data storage ensures the storage of copies of important data to ensure availability and access speed. In the proposed architecture, it is used for data analysis and recovery after detecting anomalies or cases of loss of communication with the cloud service.

The interaction between components of the smart home will be facilitated by the TCP/IP protocol, which is well-suited for IoT operations. The critical aspect in protocol selection was the universality and flexible protocol set supporting various types of IoT devices and networks, offering advantages such as compatibility, scalability, and security. Initially, the CoAP protocol was considered, but due to its incompatibility with the chosen dataset, the TCP/IP protocol was chosen [22]. Interoperability allows IoT devices to communicate with each other and with the cloud, regardless of their hardware, software, or network architecture. This will enable future implementation of enhanced security during user connection through cloud access [23]. Moreover, TCP/IP can handle a large number of IoT devices and data traffic using methods such as subnetting, routing, and addressing. Additionally, TCP/IP can provide security features for IoT devices such as encryption, authentication, and firewall. It also supports protocols like TLS (Transport Layer Security) and IPSec (Internet Protocol Security), which can enhance data transmission security and network access. Finally, TCP/IP can support various application protocols such as HTTP, MQTT, and CoAP, catering to different IoT scenarios and requirements.

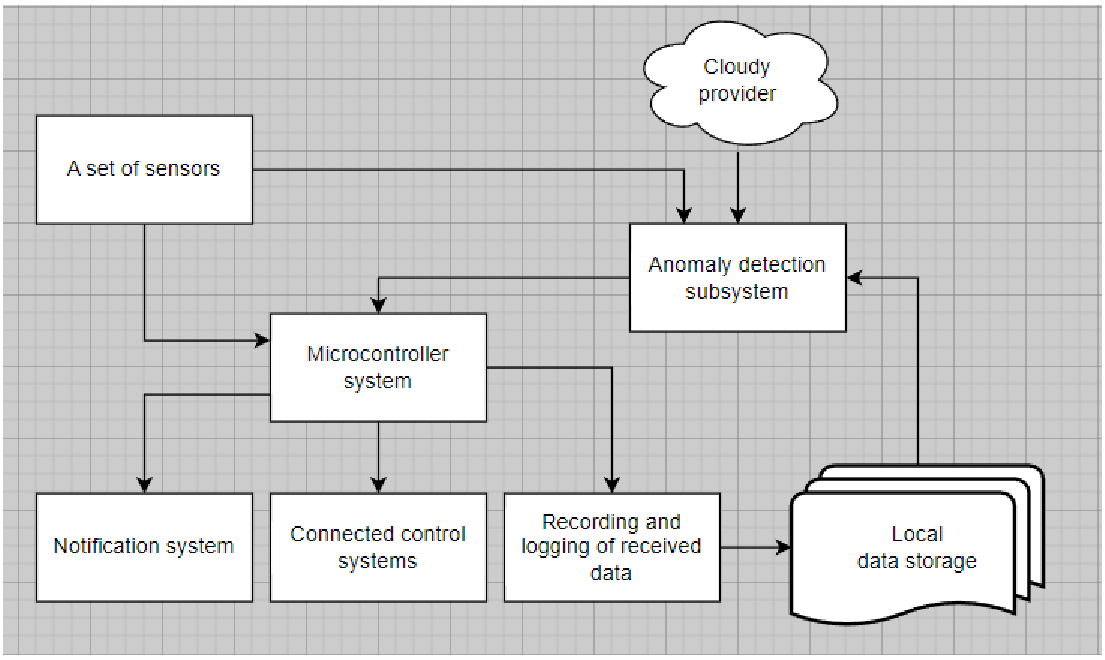
Cloudy provider
A set of sensors
Anomaly detection subsystem
Microcontroller system
Notification system
Connected control systems
Recording and logging of received data
Local data storage

Figure 1: Anomaly detection subsystem in a smart home.

The functioning of the anomaly detection subsystem in a smart home can be represented as a sequence of stages (see Fig. 2), generally including model training and utilization [24].

Initially, data is collected from various sources such as server logs, network packets, and sensor data. These data are then normalized and standardized for uniform representation, ensuring their consistency and interoperability. Data processing involves considerations such as IP addresses, packet sizes, port checks, DNS protocols, and sensor data. Min-max normalization is applied to unify the data, as all input data come in different formats. IP addresses, DNS protocols, and ports are categorical, while packet sizes and sensor data are numerical [25].

During the preprocessing stage, features are extracted, and missing or incorrect values are handled by comparing them with the average normal values. An important parameter for anomaly detection is the identification of features indicating data irregularities. For IP address analysis, the source and destination IP (Source IP/Destination IP) are checked. Packet size is determined by parameters such as the total length of forward packets (Fwd), total length of backward packets (Bwd), minimum packet length, maximum packet length, packet length mean, and packet length standard deviation. DNS protocol parameters include query repetition, name length, and domain names with invalid characters. Ports are checked for source and destination packet ports.

For sensors, critical anomaly detection parameters include the absence of measurements for a long period, sudden changes in parameters, and large discrepancies with average values. The Random Forest algorithm is used as the machine learning algorithm. For model training, in addition to collected data (in the form of features), the target indicator (label) was included. During the cyber attack detection stage, the trained model is used to classify new data as "normal" or "anomalous".

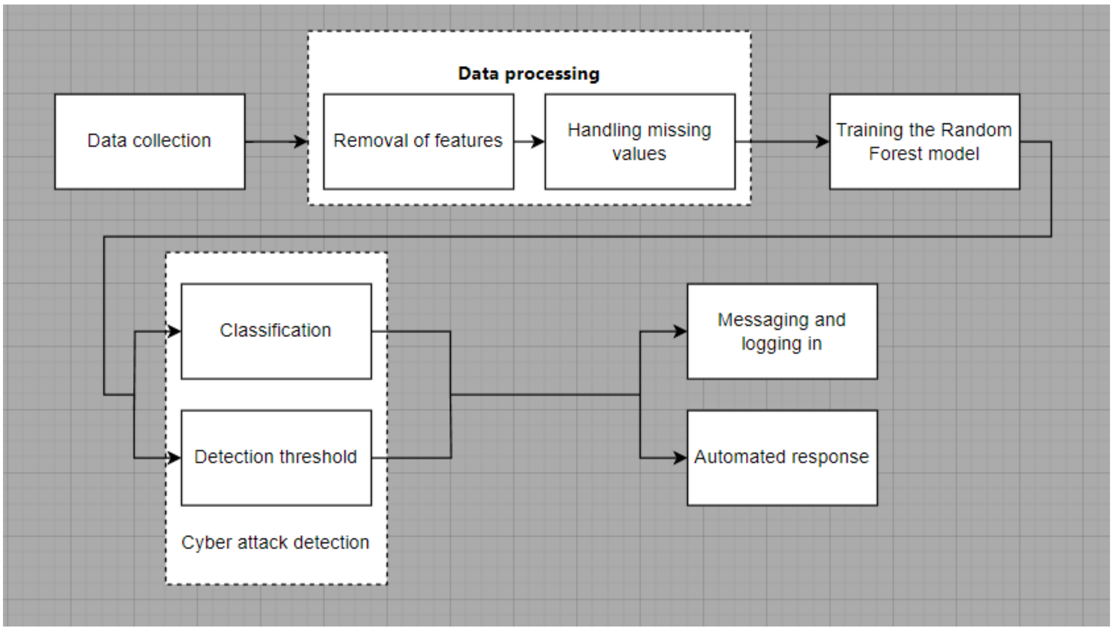

Figure 2: Architecture of anomaly detection subsystem in a smart house.

3.1 Data processing module

Data preprocessing is a critical stage in the system of data collection and cyber attack detection using Random Forest. This stage involves optimizing input data before using it in the model to prevent excessive energy consumption [26]. Initially, it is important to identify key features for the model, considering their influence. As mentioned earlier, fundamental features that significantly impact anomaly detection or cyber attacks have been selected.

During preprocessing, decisions are made on how to handle missing data, where the system compares them with regular data flow and their average values. It is also essential to normalize the data, transforming them into a uniform range of values to avoid the influence of voluminous parameters. Data transformations are also used to improve their distribution and highlight important characteristics.

For communication of messages in IoT networks, the TCP/IP protocol is used. TCP/IP provides mechanisms for communication between different devices in the IoT network. It enables devices to exchange data, including sensor information, device management, and other data. TCP/IP provides some basic security features such as message integrity, confidentiality, and endpoint identity protection using SSL/TLS (Secure Sockets Layer/Transport Layer Security).

However, it is important to note that the use of the TCP/IP protocol alone is insufficient to ensure communication security in IoT. Therefore, additional security measures such as authentication and access control need to be implemented. In this regard, the protected endpoint interacting with IoT devices should be determined by business logic, not by transport protocols and endpoint availability, as depicted in Figure 3. This is achieved through message protection at different network layers, even in the case of low-power radio devices, while maintaining system performance [27].

Devices with limited resources require a specialized protocol for secure communication that minimizes performance impact while flexibly supporting various trust models. The gateway used to support cross-device communication with the cloud can perform essential functions, but it cannot be fully trusted with access to application-level data.

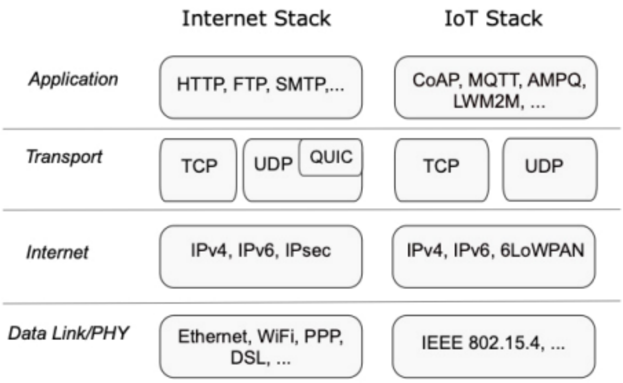

Figure 3: Interaction of the TCP/IP protocol with the IoT architecture.

The Random Forest classifier was chosen because it does not assume linear interactions or even linear functions. Random Forest is also a bagging method, which is easily scalable since each decision tree can be trained on each working node of the cluster.

In the case of DDoS Water Torture attacks, analytical functions may interact linearly, as the names of the attacked subdomains are generated randomly and queried by a large number of IP addresses. In this case, QR, SS, and SIS will interact linearly. However, during DDoS AMP attacks, capturing the name record will result in a small SS and a large QR. Therefore, tree-based models are better than linear models.

In large datasets, there may be a need for selective analysis to reduce the volume of data without losing representation. Sampling can be random, based on certain criteria, or another method to ensure the optimal amount of data for model training [28].

Some additional features may be created based on existing ones to improve data representation or detect additional relationships. For example, combining or extracting certain characteristics can help the model better understand complex interactions in the data.

These data preprocessing steps are important to ensure the quality and effectiveness of the Random Forest model in detecting cyber attacks and ensuring optimal utilization of training on input data.

3.2 Model training

The machine learning model is a key component in the proposed anomaly detection subsystem in the Smart Home system. The model is trained based on pre-processed and normalized training data. Using the trained Random Forest model, the system classifies the input data, considering each input as either "normal" or "anomalous". The main idea is that attacks often manifest as anomalous deviations from typical system behavior.

The training of the Random Forest algorithm followed the following steps:

1. Loading the dataset.
2. Applying preprocessing techniques. Discretization.
3. Partitioning the dataset into four datasets.
4. Splitting the dataset into training and testing sets.
5. Selecting the best feature set using a feature subset selection measure.
6. Passing the dataset to the Random Forest for training.
7. Passing the testing dataset to the Random Forest for classification.
8. Calculating accuracy, detection rate, and false alarm rate.

Additionally, when evaluating data packets, the Random Forest will detect anomalies based on the ensemble of branches and decisions made by the system. Let the first tree (see Fig. 4) decide whether the received data packets match normal values. Initially, the packet is compared with the total length of all packets sent forward from the source to the destination during network monitoring (Total Length of Fwd Packets) and vice versa, packets sent in the reverse direction (Total Length of Bwd Packets), which is 100 bytes. Then it checks against a maximum allowable value (Max Packet Length), for example, 50 bytes, and a minimum allowable value (Min Packet Length), 5 bytes. Also, consider the parameter of packet length mean (Packet Length Mean), 20 bytes. And take into account the standard deviation of packet length (Packet Length Std) of 10 bytes from the norm. With these parameters, the first tree in the Random Forest system can be constructed.

4 Experimental studies

For conducting experiments, all test data was split into two sets: a training set (70%) and a validation set (30%). The training set is provided to the Random Forest classifier for training, while the validation set is used to assess the classifier's performance.

All experiments were conducted using the Weka tool. For analysis, was utilized the CICIDS2017 dataset. The CICIDS2017 dataset consists of 42 attributes, with the last attribute comprising the class label [29].

Was tested various numbers of Random Forest trees. The following performance metrics were used to evaluate the classifier: 10-fold cross-validation was employed for classification [30]. Accuracy is calculated first, equation 1. Accuracy – Defined as the ratio of correctly classified samples to the total number of samples

$$Accuracy = \frac{Samples\ correctly\ classified\ in\ test\ data}{Number\ of\ samples\ in\ test\ data}, \tag{1}$$

Then the detection rate is the ratio of the total number of attacks detected by the system to the total number of attacks present in the data set, equation 2.

$$DR = \frac{TP}{\mathrm{TP} + \mathrm{TN}}, \tag{2}$$

where DR(Detection rate) is the proportion of detection, TP - correctly detected positive results, TN - correctly detected negative results.

After this false Alarm Rate – The false alarm rate is defined as equation 3.

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{3}$$

where FAR (False alarm rate) is the frequency of false alarms, FP is the number of false positive detections, TN is the number of correct negative detections.

Matthews Correlation Coefficient (MCC) is the ratio between observed and predicted binary classifications, equation 4.

$$\mathrm{MCC} = \frac{\mathrm{TN} \times \mathrm{TP} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{FP} + \mathrm{TP})(\mathrm{FN} + \mathrm{TP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \tag{4}$$

where TN – Correct negative; FN – False negative; FP – False positive result; TP – Correct positive.

For experimental analysis, was utilized the CICIDS2017 dataset in ARFF format. The following preprocessing steps were conducted:

1. Missing Values Imputation. Was applied the Weka filter to replace all missing values in the CICIDS2017 dataset. The filter utilizes the mean and mode from the training data to replace missing values.
2. Discretization. A discretization filter with 10 bins was used for numerical attributes.

To integrate the cyber-attack detection system into a smart home, this system can be considered as part of a comprehensive infrastructure solution that interacts with other elements of the smart home. The primary task is to position the system at a level that ensures the necessary security and efficiency of its operation.

Regarding the architecture of the cyber-attack detection system, and can employ distributed components for threat analysis and detection. This may include modules for data analysis, machine learning, as well as an interface for interaction with other systems in the smart home.

Thanks to the use of Weka and the corresponding architecture, it is possible to effectively model attacks and analyze their impact on the smart home, providing a reliable system of integration and protection against cyber threats.

For thorough analysis and evaluation of the effectiveness of cyber-attack detection methods, was conducted comparisons of their performance on different types of attacks [31]. In this study, four main classes of cyber threats were considered: DOS (Denial of Service), Prob (Probe), R2L (Unauthorized Access to Remote System), and U2R (Unauthorized Access by Privileged Users). The tables below display the results for each type of attack, including accuracy, sensitivity (DR), false alarm rate (FAR), and Matthews correlation coefficient (MCC).

In analysis and comparison, was used the Weka program, which provides a wide range of tools for machine learning and data analysis.

To understand the performance of the random forest approach, it was compared with the J48 tree method. The performance of the proposed approach is presented in Table 1. It can be seen from Tables 1, 2 and 3 that the proposed model achieved high DR and low FAR for attack classification. For DOS attacks, the proposed model achieved an accuracy of 94.98%, which is 7% higher than the J48 algorithm. The FAR recorded for the J48 is higher than the proposed model. A good classifier for detecting attacks should have a high DR and a low FAR (Figure 5).

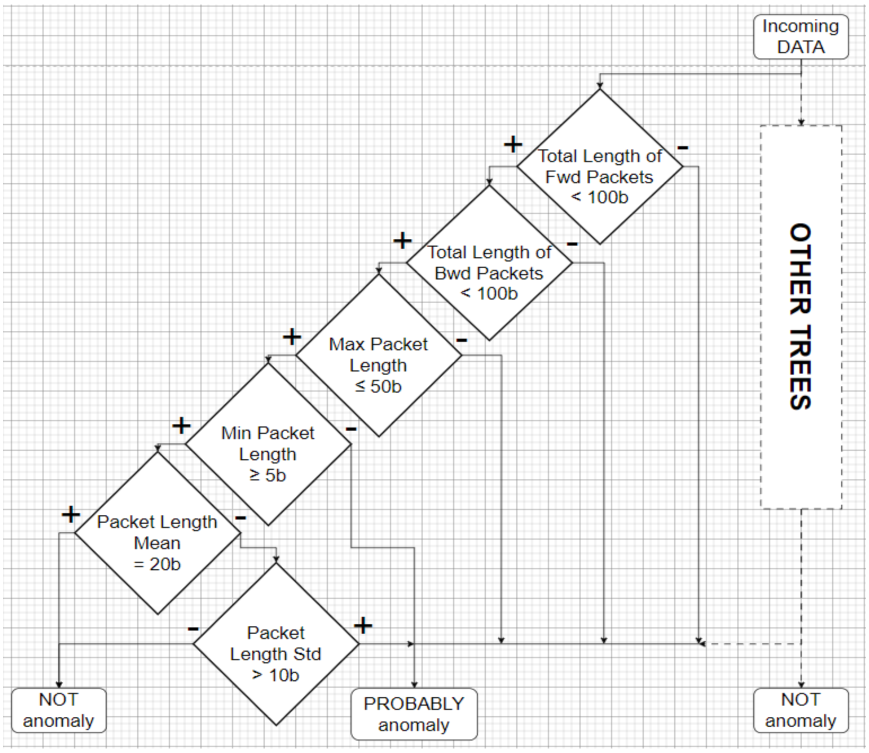

Figure 4: Anomaly search tree in Random Forest based on selected features.

After applying the symmetric uncertainty (SU) feature selection function, the measurement accuracy and DR were improved and the FAR was reduced.

Table 1
Performance index for a random forest (number of trees = 100).

№	Attack type	Accuracy	DR	FAR	MCC
1	Dos	94.98	95.83	0.00519	0.94
2	Prob	95.34	95.81	0.00501	0.93
3	R2L	95.27	95.81	0.00501	0.94
4	U2R	94.93	95.83	0.00548	0.92

Table 2
Performance indicator for the J48 tree.

№	Attack type	Accuracy	DR	FAR	MCC
1	Dos	94.91	95.3	0.00828	0.933
2	Prob	95.19	95.4	0.0093	0.933
3	R2L	95.13	95.3	0.010	0.948
4	U2R	94.98	95.3	0.0075	0.948

Table 3
After applying the FSS-symmetric uncertainty.

№	Attack type	Accuracy	DR	FAR	MCC
1	Dos	94.88	95.82	0.00467	0.95
2	Prob	95.43	95.73	0.00467	0.95
3	R2L	95.39	95.86	0.00501	0.94
4	U2R	94.97	95.82	0.00467	0.93

For the trial attack after applying the feature selection feature, the DR is recorded as 95.83%. For R2L and U2R, the MCC was recorded as 0.94 and 0.92, respectively, indicating the effectiveness of this approach for classifying attacks in IDS.

The average accuracy obtained by the proposed approach without feature selection is 95.13%, while it is only 95.05% for J48.

The Matthews correlation coefficient recorded by this model is high compared to the J48 classifier. Experimental results show that this approach can achieve high accuracy, high DR with low FAR.

Interestingly, for metamorphic viruses, which are capable of disguising themselves and changing their features, this method would be less effective.

For this, a more complex algorithm is needed, which can maintain efficiency at 85%, which is quite high for this type of threat [32].

Alternatively, employing feature obfuscation analysis may raise efficiency to 94% [33].

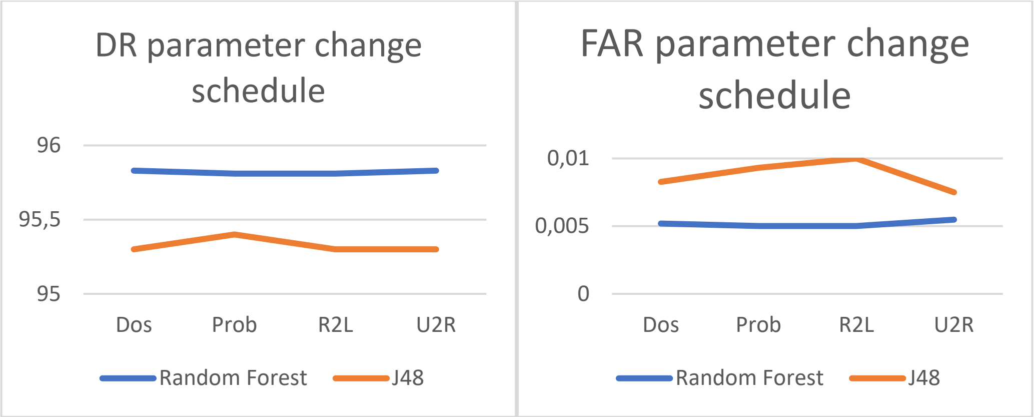

Figure 5: Comparison of DR and FAR threat detection parameters for Random Forest and J48.

Additionally, some research suggests using API Call Tracing for threat detection [34, 35].

The effectiveness of this method, according to studies, is 96.56%, but the high computational overhead would increase energy efficiency costs, which would be unacceptable and contradict the priorities of this system.

5 Conclusions

In summary, the development of anomaly detection subsystems based on machine learning algorithms represents significant progress in enhancing the security of Smart Home systems. The utilization of the Random Forest algorithm demonstrates promising results in effectively detecting anomalies in IoT device-generated network traffic. Experimental evaluations conducted on the CICIDS2017 dataset underscore the effectiveness of the proposed subsystem in various attack scenarios. By providing early detection and response capabilities, the subsystem contributes to safeguarding smart homes against potential cyber threats, ensuring the integrity and reliability of these interconnected environments.

This subsystem represents an important and timely direction in the field of cybersecurity and beyond. With the continuous advancement of technologies and the increasing number of network-connected devices, cybersecurity issues are becoming increasingly relevant, and this subsystem is an integral part of strategies to combat these threats.

The application of such subsystems manifests in various domains such as cybersecurity, finance, medicine, and others, indicating its versatility and significant potential. One of the main advantages is its ability to detect anomalies in real-time and adapt to new forms of threats. The subsystem enables the detection of patterns that may go unnoticed by traditional methods and provides the ability to preempt attacks or other malicious activities.

A review of related works highlights the diversity of methods and approaches in the field of cybersecurity, from statistical methods to deep learning based on various algorithms. However, considering the constant evolution and refinement of methods, such subsystems remain a key tool for ensuring cybersecurity and detecting anomalies and cyber-attacks in the modern world.

6 References

[1] G. K. Mehrotra, K. M. Chilukuri, H. Huang, Anomaly Detection Principles and Algorithms (Terrorism, Security, and Computation) 1st ed. (2017).

[2] C. Zhou, P. Zhang, J. Li, H. Luo, Y. Wang, Deep Learning for Anomaly Detection: A Review. Mathematical Problems in Engineering. (2021).

[3] C.S Smith, M. Koning, Decision Trees and Random Forests: A Visual Introduction For Beginners: A Simple Guide to Machine Learning with Decision Trees. (2017).

[4] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. (2022).

[5] I. Obeidat, M. AlZubi, Developing a faster pattern matching algorithms for intrusion detection system. International Journal of Computing, 18(3) (2019) 278-284. doi:10.47839/ijc.18.3.1520

[6] J. Howard. (2018). Introduction to Machine Learning for Coders.

[7] S.L. Cheruvu, A. Kumar, et al. (2019). Demystifying Internet of Things Security: Successful IoT Device/Edge and Platform Security Deployment.

[8] R. Wason. (2014). Internet of Distributed denial-of-service (DDoS) attack: Learn DDos Attack Methods.

[9] S. Ciuta. (2023). Securing the Internet of Things (IoT): Cybersecurity of Connected Devices.

[10] I. Goodfellow, Y. Bengio, A. Courville. (2017). Deep Learning.

[11] S. Shalev-Shwartz, S. Ben-David. (2015). Understanding Machine Learning: From Theory to Algorithms.

[12] S. Raschka, V.Mirjalili. Python Machine Learning, 2022.

[13] M. P. Deisenroth. Mathematics for Machine Learning. (2020).

[14] Z. Zhi-Hua. (2021). Machine Learning.

[15] A. Ng. Machine Learning on Coursera. URL: https://www.coursera.org/learn/machine-learning.

[16] L. Chen, Y. Zhang, Q. Zhao, G. Geng, Z. Yan. Detection of DNS DDoS Attacks with Random Forest Algorithm on Spark. 134 (2018) 310-315.

[17] J. Antonio Garcia-Macias. (2023). Transport in the IP-based Internet of Things: status report

[18] N. Moustafa, J. Slay, A Survey of Anomaly Detection in Internet of Things, 2019.

[19] Kiranyaz, S., Avci, O., Abdeljaber, O., & Incecik, E. (2017). Machine Learning for Anomaly Detection and Diagnosis in Aeronautics.

[20] O. Pomorova, O. Savenko, S. Lysenko, A. Kryshchuk, A. Nicheporuk, A technique for detection of bots which are using polymorphic code, Communications in Computer and Information Science. 431 (2014) 265-276.

[21] O. Savenko, S. Lysenko, A. Nicheporuk et al., Metamorphic Viruses' Detection Technique Based on the Equivalent Functional Block Search, CEUR Workshop Proceedings. 1844 (2017) 555-569.

[22] R. Achary, C. J. Shelke, K. Marx, Aishwarya Rajesh. (2023). Security Implementation on IoT using CoAP and Elliptical Curve Cryptography. doi: "doi.org/10.1016/j.procs.2023.12.105".

[23] I. R. Chiadighikaobi, N. Katuk, B. Osman, DMUAS-IoT: A Decentralised Multi-Factor User Authentication Scheme for IoT Systems. International Journal of Computing 21(4) (2022). 424-434, doi: "doi.org/10.47839/ijc.21.4.2777".

[24] S. Garcia, M. Grill, , J.Stiborek, A. Zunino, Cambiaso, E. Machine Learning in Cybersecurity: A Comprehensive Survey. 19 (2019).

[25] J. Antonio Garcia-Macias. (2023). Transport in the IP-based Internet of Things: status report. doi: doi.org/10.1016/j.procs.2023.09.006.

[26] O.Yakubu, B. C. Narendra, C.O. Adjei, A Novel IoT Based Smart Energy Meter with Backup Battery. International Journal of Computing, 20(3), (2021). 357-364, doi: doi.org/10.47839/ijc.20.3.2281.

[27] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A Survey. (2014).

[28] C. Zhou, P. Zhang, J. Li, H. Luo, Y. Wang, Deep Learning for Anomaly Detection: A Review. (2018).

[29] H. Lindstedt. Methods for network intrusion detection. Evaluating rule-based methods and machine learning models on the CIC-IDS2017 dataset. (2022).

[30] N. A. Farnaaz, J. Akhil. Random Forest Modeling for Network Intrusion Detection System. (2016). doi: doi.org/10.1016/j.procs.2016.06.047.

[31] A. Feijoo-Añazco, D. Garcia-Carrillo, Jesús Sanchez-Gomez, Rafael Marin-Perez. Innovative security and compression for constrained IoT networks. (2023). doi: "doi.org/10.1016/j.iot.2023.100899".

[32] O. Pomorova, O. Savenko, S. Lysenko, A. Nicheporuk, Metamorphic Viruses Detection Technique based on the the Modified Emulators, CEUR Workshop Proceedings. 1614 (2016) 375-383.

[33] A. Kashtalian, S. Lysenko, O. Savenko, A. Nicheporuk, T. Sochor, V. Avsiyevych, Multi-computer malware detection systems with metamorphic functionality. Radioelectronic and Computer Systems. 2024(1), 152-175. doi: 10.32620/reks.2024.1.13

[34] O. Savenko, A. Nicheporuk, S. Lysenko, et al., Dynamic signature-based malware detection technique based on API call tracing CEUR Workshop Proceedings, 2393 (2019) 633-643.

[35] V. Khoroshko, V. Kudinov, M. Kapustian Evaluation of quality indicators of functioning cyber protection management systems of information systems, Computer Systems and Information Technologies. 2 (2022) 47–56. doi: 10.31891/csit-2022-2-6.