International Journal of Machine Learning and Cybernetics (2021) 12:3179–3202
https://doi.org/10.1007/s13042-020-01241-0

ORIGINAL ARTICLE

Ensemble machine learning approach for classification of IoT devices in smart home

Ivan Cvitić[1] · Dragan Peraković[1] · Marko Periša[1] · Brij Gupta[2,3,4]

Received: 31 August 2020 / Accepted: 19 November 2020 / Published online: 3 January 2021

Abstract
The emergence of the Internet of Things (IoT) concept as a new direction of technological development raises new problems such as valid and timely identification of such devices, security vulnerabilities that can be exploited for malicious activities, and management of such devices. The communication of IoT devices generates traffic that has specific features and differences with respect to conventional devices. This research seeks to analyze the possibilities of applying such features for classifying devices, regardless of their functionality or purpose. This kind of classification is necessary for a dynamic and heterogeneous environment, such as a smart home where the number and types of devices grow daily. This research uses a total of 41 IoT devices. The logistic regression method enhanced by the concept of supervised machine learning (logitboost) was used for developing a classification model. Multiclass classification model was developed using 13 network traffic features generated by IoT devices. Research has shown that it is possible to classify devices into four previously defined classes with high performances and accuracy (99.79%) based on the traffic flow features of such devices. Model performance measures such as precision, F-measure, True Positive Ratio, False Positive Ratio and Kappa coefficient all show high results (0.997–0.999, 0.997–0.999, 0.997–0.999, 0–0.001 and 0.9973, respectively). Such a developed model can have its application as a foundation for monitoring and managing solutions of large and heterogeneous IoT environments such as Industrial IoT, smart home, and similar.

Keywords Boosting · Cybersecurity · Supervised learning · Internet of things · ML

1 Introduction

The application of the IoT concept in different economic sectors is becoming a key factor for business improvement. According to [1], 92% of companies believe the IoT concept will be important for their business by the end of 2020. Consequently, the companies consider that security, privacy, costs, and regulatory issues pose the greatest challenges of implementing and applying the IoT concept. Research [2] conducted in 1,430 companies (small, medium, and large) points to a number of advantages seen by the vast majority (95%) of adopters of the IoT concept. In doing so, more than half (53%) confirm significant benefits of implementing the IoT concept in business, while 79% of those surveyed believe that by applying the IoT concept, they achieve positive results in different areas of work that they would not otherwise be able to achieve.

According to Gartner, the largest representation and application of the IoT concept, according to the number of IoT devices used until 2017, was in the area of smart building environments. After 2017, the smart home concept is the environment that brings together the largest number of IoT devices [3]. More precise insight into the representation of IoT devices by individual areas of application is provided by the research of the company IHS Markit [4]. It can be seen that the smart home concept has the largest number of

✉ Ivan Cvitić
ivan.cvitic@fpz.unizg.hr

✉ Brij Gupta
bbgupta.nitkkr@gmail.com

1 Department of Information and Communication Traffic, University of Zagreb, Faculty of Transport and Traffic Sciences, Zagreb, Republic of Croatia

2 Department of Computer Engineering, National Institute of Technology Kurukshetra, Kurukshetra, Haryana, India

3 Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan

4 Macquarie University, Balaclava Rd, Macquarie Park, NSW 2109, Australia

installed IoT devices (822.6 million) compared to other areas of application. The annual growth rate (prediction by 2021) is 19.6%, which makes the smart home concept [5–7], along with the industrial IoT concept (CAGR 23.4%), the fastest growing area of application of the IoT concept. The classification of IoT devices is essential for several reasons. Successfully identifying IoT devices in a particular scenario and environment can be vital in identifying illegitimate devices, unauthorized devices, unwanted devices, devices that do not behave as expected, and have the potential to cause a security incident within the system. Besides, useful device classification and identification of new and hitherto unseen devices can enable more efficient traffic management as well as network capacity required in the environments in which IoT devices exist [8, 9].

The rest of this paper is organized as follows: the second chapter deals with the current research, their shortcoming, and the positioning of our research according to previous findings. In the third chapter, data collection approach is explained, which includes laboratory environment establishment, raw network traffic collection, data preprocessing, and device class definition as key activities for further classification model development. The fourth chapter explains the classification model development as well as the ensemble supervised machine learning method used for that purpose. In the fifth chapter, the results from the developed model were analyzed and discussed. In the final chapter, the authors give their conclusion and further research direction.

2 Related work

According to the forecasts presented in [10], by the end of 2020, approximately 31 billion IoT devices will be globally used, and until 2025 there will be 75 billion IoT devices. At the same time, 41%, i.e. 12.86 billion IoT devices will be installed within a smart home (SH) [11]. IoT device limitations in general, and thus SHIoT (smart home IoT) devices, are described in the research [12]. Limitations include hardware limitations, requirements for high autonomy and low production cost, which reduces the possibility of implementing advanced protection methods and increases the risk of many threats shown in [13]. The traffic generated by SHIoT devices or MTC (Machine Type Communication) traffic differs from the traffic generated by conventional devices, HTC (Human Type Communication) traffic, which was shown by research [14]. Specific features of MTC traffic have been used to solve several problems in the communication network. Research [15] looks at the impact of MTC traffic on QoS during integration with HTC traffic in the LTE (Long-Term Evolution) communications network. Identification and classification of IoT devices in smart cities [5, 6] and campuses and in smart environments using MTC traffic features have been presented by research [16] and [17]. Research [18] seeks to identify new requirements and challenges in the design and management of a mobile communication network imposed by the generation of MTC traffic.

SHIoT traffic can be observed through network activity features such as traffic volume (sum of the total traffic received and total traffic transferred), traffic flow duration (time between first and last packet in traffic flow), and device inactivity time (the period in which the device has no active traffic flow). The network behavioral modeling is an often-used approach to address communication network challenges such as detecting illegitimate events based on traffic generated by devices on the network. In general, current approaches seek to identify traffic characteristics at the network packet level and the traffic flow level [19]. The analyzed research shows more frequent consideration and use of traffic features at the level of traffic flow than at the level of network packages. Likewise, the mentioned studies use the presented features to identify individual devices or their classification based on the semantic characteristics of the observed devices. [20]. Authors in [21] developed a tool for automatic extraction of packet-level signatures of IoT devices form the network traffic. They have extracted packet-level features of 18 smart home devices, which was used as a basis for development of classification model with a recall of 97%. Although research represents high results of the developed classification model, it remains unclear how the model will behave on the previously unseen devices. It should be trained again for every new device that comes on the market. Such an approach is not suitable considering the nature of the IoT concept. In research [22], the authors present the LSIF (Locality-Sensitive IoT Fingerprinting) approach for identification of IoT devices. The presented approach does not require feature extraction from the traffic. Although this approach has its benefits, it is lacking in performance such as precision (93%) and recall (90%). Also, this approach is focused on the identification of individual devices, which raises already mentioned shortcomings. The research presented in [23] used artificial neuron network for the classification. They developed a model that can identify nine known devices with approximately 99% accuracy. In research [24] the primary goal is to develop a model for network anomaly detection caused by IoT devices. For that purpose, the authors first developed a classification model for profiling of normal behavior. They used J48 machine learning method for model development with precsion, recall and F-measure, 96.2%, 96.8% and 96.9%, respectively. The developed model is actual only for nine devices they used in research because profiling was done for an individual device. The lack of research can be noticed in the number of used devices and insufficient generalization of the problem, where for every new device a new model needs to be learned, trained, and

validated. In research [25], the authors use the decision trees and deep learning based methods for identification, classification, and anomaly detection of IoT devices. This research tries to use a more general approach to the classification of network traffic by using tree classes of traffic (actuation, sensing, video streaming). Such an approach is suitable for recognizing normal (expected) behavior of network traffic generated by IoT devices, and it is useful in resolving problems such as anomaly detection. Negative sides of this research are the number of used devices (7), the amount of network traffic (5 days), and the results of the developed classification model (93.5%).

According to the above, the possibility of developing an efficient classification model of IoT devices based on the characteristics of the generated traffic flows is set as a hypothesis of the current research. The research aims to develop a classification model based on an ensemble supervised machine learning method that will be able to assign IoT devices to predefined classes based on the values of their traffic flows. Current research in this domain is trying to identify the individual device. Such an approach is not suitable in the fast-evolving, heterogeneous, and dynamic environment such as IoT, where the number of new devices is rising exponentially. Because of the mentioned, the approach in this research brings novelty and gives the opportunity to recognize a class of new and unseen IoT devices based on its network traffic behavior. Our approach tends to generalize the identified problem and develop a solution that would be adjusted to the nature of the IoT environment. Accordingly, IoT devices need not be observed individually in solving a problem such as certain types of management of IoT devices, detecting network anomalies generated by IoT devices, or identifying unauthorized IoT devices in the network. For that purpose, a classification model is needed that would be able to assign previously unseen devices to generic behavior profile. This research, compared to previous ones gives contribution in a larger set of observed devices, longer period and larger amount of collected data, innovative approach in the classification of IoT devices, and better performance results of the developed classification model.

3 Proposed approach

This research has been conducted in three phases with the activities shown in Fig. 1. In the first phase the research problem was identified and laboratory environment established. The dataset was formed from primary and secondary data sources. In the second research phase index C_u was extracted for each device and IoT device classes were defined. The collected data have been preprocessed which included feature engineering and data normalization (dealing with null and categorical values). In the final, third phase the dataset was balanced, and the classification model was developed. For the model development, the ensemble supervised machine learning method was used. The developed model performance was measured using standard validation measures for the classification models such as confusion matrix, accuracy, kappa coefficient, TPR (True Positive Ratio), FPR (False Positive Ratio), F-measure, ROC (Receiver Operating Characteristics) curve and other.

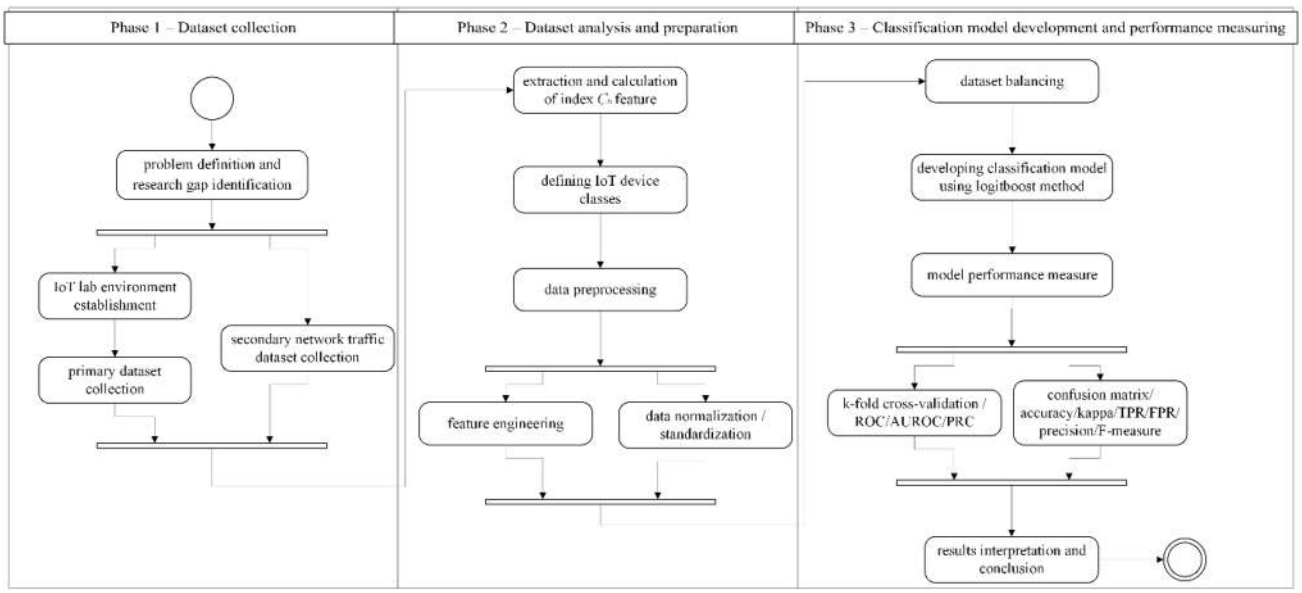

Fig. 1 Research phases and activities

One of the crucial research activities was primary data collection for which the laboratory environment with SHIoT devices was established. SHIoT devices are supplied by authorized distributors and representatives of each device manufacturer. They are connected to the communication network as recommended by the manufacturer, and in no way are the devices modified at the software and hardware level. Therefore, it is assumed that the devices that are used to collect legitimate traffic in this research work are as designed and are in no way previously compromised in terms of security.

The network topology, as well as the characteristics of the smart home environment, can be seen in Fig. 1. The devices are connected, directly or indirectly, by Wi-Fi communication technology to the Fortinet AP 221C wireless access point, except for Phillips Hue, which communicates with the rest of the local network via Ethernet (IEEE 802.3) communication standard. Some devices, such as the Blink smart camera, Netatmo smart thermostat, and Philips Hue smart lighting fixtures, use an IoT hub with which they communicate wirelessly, but with ZigBee technology. The reason is the energy efficiency of the device since they use the battery as the power source of the end device, which gives them advantages in terms of mobility and independence of the device from electricity as a power source. The IoT hub is connected to Wi-Fi (or Ethernet in the case of Phillips Hue devices) technology with a wireless access point. Based on the above, a wireless access point has been determined as an adequate collection point for traffic generated by SHIoT devices. Due to the known modes of operation and characteristics of computers, and thus wireless Wi-Fi networks, traffic in the communication network cannot be collected directly. Several methods are available for traffic collection, often using physical port mirroring on the switch. This method is efficient in several studies, such as [15, 26–28], which provides a basis for the application of the same method in conducting this research.

A software-hardware platform consisting of a Fortinet AP 221C wireless access point, a Cisco 2960 Catalyst 48 PoE switch (Power over Ethernet) and an HP Pavillion dm1 workstation (Microsoft Windows 10 10.0.17134 build 17,134, ×64 processor architecture, AMD E-350, 1600 MHz 2 cores, 4 GB RAM) has been set up to collect traffic by port mirroring with Wireshark software tool version 2.6.3 installed.

As shown in Fig. 2, port mirroring is configured for the physical communication ports (FA0 / 1 and FA0 / 3) of the switch to which the wireless access point and IoT hub for the Phillips Hue device are connected. These ports are configured as a source, which means that all traffic coming to or from these ports will be mirrored (mapped) to the destination communication port (FA0 / 2). A traffic collection workstation is connected to this port (Fig. 2).

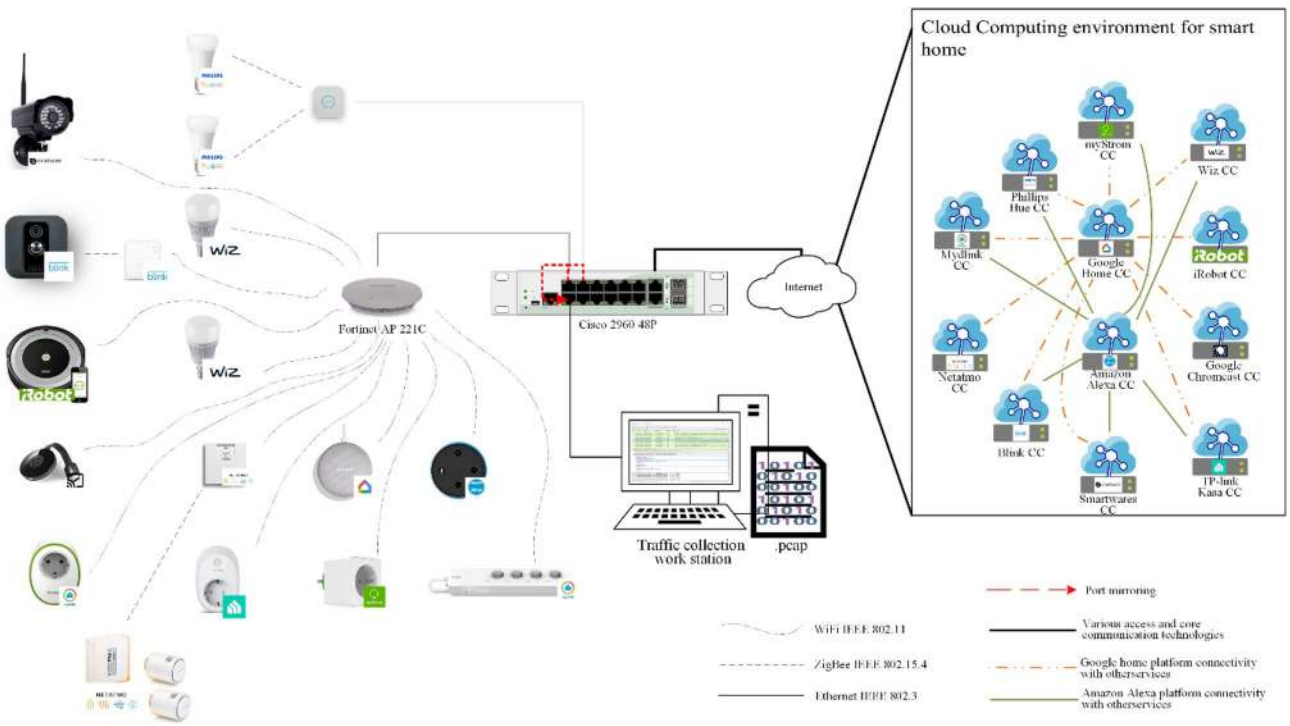

Fig. 2 Laboratory environment of a smart home formed for data collection [29]

3.1 Analysis of the used SHIoT devices

The laboratory environment of a smart home was formed to collect primary data. It contains SHIoT devices commercially available on the market, considering that, according to statistical indicators, such devices feature continuous growth of the application. Figure 3 shows the distribution of SHIoT devices, i.e., the representation of each group in the total number of devices and the number of devices that will be used to collect the primary and secondary data. The complete list of SHIoT devices included in this research is shown in Table 1.

The smart home laboratory environment was formed within the Laboratory for security and forensic analysis of the information and communication system of the Department for information and communication traffic at the Faculty of Transport and Traffic Sciences. In addition to SHIoT devices intended for the collection of primary data, for the subject research, secondary data already collected through various SHIoT devices within the existing research were used [17, 30, 31].

Table 1 lists the MAC (Media Access Control) addresses as the unique identifiers of the SHIoT device in the network, the device name, the P / S code indicating whether the observed device was used to collect primary or secondary data, and to which functional group the observed SHIoT the device belongs.

A total of 41 devices in a smart home environment were used for the research, part of which was already shown in [29]. According to statistics, there are differences in the estimate of the average number of SHIoT devices per household that has implemented a specific form of a smart home. These estimates range from 6.53 to 14 SHoT devices per household. In the Republic of Croatia, the representation of smart homes is still low, and telecom operators are taking on the role of smart home service providers through the offer of SHIoT devices for end-users. For example, the Internet service provider Iskon Internet offers customers the opportunity to purchase a smart home package consisting of four SHIoT devices [32]. In comparison, the telecom operator A1 offers customers the opportunity to implement a total of five SHIoT devices in a smart home environment [33].

Despite the above, this research sought to achieve the highest possible diversity of SHIoT devices due to the need to define device classes based on the characteristics of the generated traffic. Therefore, the number of devices used is higher than the current statistical estimate of the average value of SHIoT devices per smart home in the Republic of Croatia and the world. The predictions shown in [34] refer to the period until 2023, but given the upward trend in the growth of the number of devices, it is to be assumed that the number of devices will reach 40 per smart home in the foreseeable future.

3.2 Descriptive statistical analysis of collected data

The primary dataset formed for this research consists of a total of 103 files in.pcap format that contain a complete record of network traffic. The secondary dataset consists of 41 files of the same format as the primary set, which makes a total of 144 network traffic files generated by various SHIoT devices and represents the legitimate network traffic. Each of the 144 files contains traffic generated in a 24-h time interval.

Table 2 shows the statistical description of the dataset through statistical measures of standard deviation, minimum, maximum, and mean values at the level of 24-h intervals of collected traffic for primary and secondary data and the consolidated dataset. Statistical description is represented through three logical parts: primary data, secondary data and total.

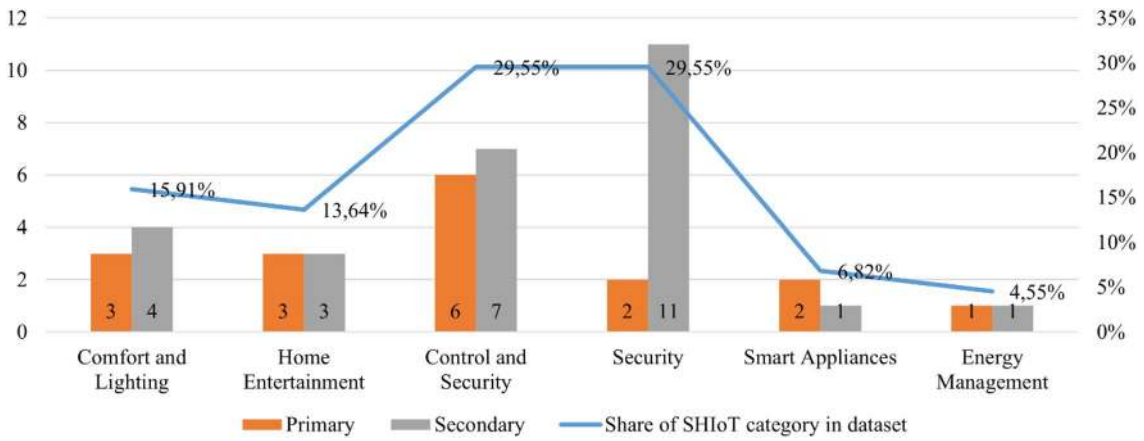

Fig. 3 Distribution of SHIoT device groups

Table 1 SHIoT devices for data collection purposes

No.#	Device	Label	Device name	Data collection source	Functional category of SHIoT device
1	ph_hue_2	u10	Phillips Hue Starter kit 2xE26	P	CL
2	ph_hue_4	u33	Phillip Hue Starter kit 4xE26	S	CL
3	wiz_F3	u2	WiZ Colors ESP_0531F3	P	CL
4	wiz_B0	u4	WiZ Colors ESP_0506B0	P	CL
5	lifx	u20	Light Bulbs LiFX Smart Bulb	S	CL
6	wit_aura	u14	Withings Aura Sleep Tracking Mat	S	CL
7	google_chr	u36	Google Chromecast	P	HE
8	triby	u39	Invoxia Triby Speaker	S	HE
9	pix	u34	PIX-STAR Photo-frame	S	HE
10	amz_dot	u17	Amazon Alexa Dot	P	HE
11	amz_echo	u30	Amazon Alexa Echo	S	HE
12	google_mini	u41	Google Home mini	P	HE
13	hs110	u12	TPlink Smart Plug HS110	P	CC
14	hs105	u3	TPlink Smart Plug HS105	S	CC
15	my_strom	u24	MyStrom switch	P	CC
16	w245	u15	D-link DSP-W245 plug	P	CC
17	w115	u6	D-link DSP-W115 plug	P	CC
18	ihome	u7	iHome Power Plug	S	CC
19	belk_sw	u13	Belkin Wemo switch	S	CC
20	sams_st	u5	Samsung Smart Things	S	CC
21	bc_blood	u26	Blipcare Blood Pressure meter	S	CC
22	aw_aq	u40	Awair air quality monitor	S	CC
23	i896	u23	iRoobot Roomba 896	P	SA
24	i895	u22	iRoobot Roomba 895	P	SA
25	wit_body	u35	Withings Body	S	SA
26	smartw_cam	u21	Smartwares C923IP Camera	P	S
27	blink_cam	u18	Blink XT2 Camera	P	S
28	cana_cam	u11	Canary View Camera	S	S
29	net_cam	u32	Netatmo Welcome Camera	S	S
30	tp_cam	u1	TPlink Day Night Cloud NC220 camera	S	S
31	sams_cam	u19	Samsung SmartCam	S	S
32	nest_cam	u38	Nest Dropcam	S	S
33	belk_cam	u25	Belkin NetCam Camera	S	S
34	inst_cam	u28	Insteon HD WiFi Camera	S	S
35	wit_baby	u8	Withings Smart Baby Monitor	S	S
36	belk_mot	u31	Belkin Wemo Motion Sensor	S	S
37	nest_smoke	u9	NEST Protect Smoke Alarm	S	S
38	aug_door	u29	August Doorbell Cam	S	S
39	ring_vd	u37	Ring Video Doorbell	S	CL
40	net_therm	u16	Netatmo smart thermostat	P	EM
41	net_weath	u27	Netatmo Smart Weather Station	S	EM

P primary; *CL* comfort and lighting; *HE* home entertainment; *CC* control and connectivity; *S* secondary; *S* security; *SA* smart appliances; *EM* energy management

For every logical part we gave standard deviation, minimum, maximum and mean value for the parameters such as Number of collected packets, File size, Amount of collected data, Average data transfer rate, Average packet transfer rate, and Average packet size. These measures show the characteristics of the collected data. For example, it can be concluded that secondary dataset is bigger than the primary one or that the average packet size is smaller in the secondary dataset than in the primary. All previously mentioned can be explained with a high level of device heterogeneity

Table 2 Statistical description of the collected legitimate network traffic data

Statistical measure	Number of collected packets	File size (Byte)	The amount of data collected (Byte)	Average data transfer rate (B/s)	Average packet transfer rate (packets/s)	The average packet size (Byte)
Primary data						
Standard deviation	2,613,702.80	2,503,780,307	2,462,270,632	28,498.13	30.25	175.48
Minimum value	1,019,339	288,056,862	271,747,414	3,145.21	11.80	252.50
Maximum value	14,815,959	13,562,522,315	13,325,466,947	154,232.80	171.48	899.40
Mean value	4,428,879.60	3,416,448,737	3,345,586,639	38,721.71	51.25	677.78
Secondary data						
Standard deviation	1,646,515.40	804,291,290.10	765,130,504.3	12,047.91	186,6204	20.69
Minimum value	527,035	89,615,024	71,959,664	832.87	136.54	6.10
Maximum value	7,720,905	3,483,660,828	3,322,027,939	60,188.48	910.56	730.47
Mean value	2,365,097.30	986,887,232.4	908,751,070	11,565.99	330.12	45.21
Total						
Standard deviation	2,557,564.50	2,429,423,670	2,396,969,893	27,864.11	165.57	329.45
Minimum value	527,035	89,615,024	71,959,664	832.87	11.80	6.10
Maximum value	14,815,959	13,562,522,315	13,325,466,947	154,232.80	910.56	899.40
Mean value	3,835,384.10	2,714,894,474	2,641,775,718	30,881.53	132.69	492.91

Table 3 Characteristics of the initial traffic dataset

	Number of files	Number of collected packets	File size (GB)	The amount of data collected (GB)	Collection period (hours)
Primary (sum)	103	456,174,601	351.89	344.59	2,472.01
Secondary (sum)	41	99,334,088	41.44	38.16	986.45
Total	144	555,508,689	393.33	382.75	3,458.47

and diversity in both datasets which are characteristics of the devices in IoT concept. The characteristics of the initially collected data are shown in Table 3. They are expressed through the number of collected files containing 24-h intervals of generated traffic, number of collected packets, file size, amount of collected data, and the total period of data collection.

The network traffic acquisition tool (Wireshark) uses specific metadata that it records within files with the collected traffic, which makes a difference between the size of the file and the amount of collected data (traffic) contained in the file.

3.3 Extraction of identified traffic features

To develop the SHIoT device classification model, the process of filtering traffic from an individual.pcap file according to the MAC address of the device was performed. The reason for this way of filtering is the assignment of an IP (Internet Protocol) address to devices via a DHCP (Dynamic Host Configuration Protocol) server, which is why it can change over time and does not represent a reliable feature according to which it is possible to accurately filter traffic to a particular device over time.

The research observes the traffic characteristics for individual SHIoT devices covered by the research (41 devices) at the traffic flow level. The traffic flow is defined by a sequence of packets with equal values of source IP address, destination IP address, source communication port, destination communication port and the protocol used, TCP (Transmission Control Protocol) or UDP (User Datagram Protocol) [35]. The reason for choosing the traffic flow as the level of observation and analysis of traffic characteristics is that it represents the aggregated (statistical) data of the packet header for communication between the source and the destination. The analysis of packet-level traffic features encompasses more information such as packet content, and also requires more computing resources to store and process them. An example of the relationship between the number of traffic flows and the number of packages in 24 h is visible for the Google Chromecast device (covered by this study), where 11,877 separate traffic flows were generated while the number of packets is 2,459,538. Nowadays, the number of devices

and applications uses cryptographic methods for communication. The contents of the packet cannot be observed and analyzed in an economically, temporally, and legally acceptable way. Therefore, the observation and analysis of traffic characteristics at the traffic flow level represent an acceptable and frequently used approach in numerous studies.

The CICFlowMeter software tool was used to extract traffic flow features. CICFlowMeter is a tool developed at the Canadian Institute of Cyber Security, University of New Brunswick [36]. The tool was developed in the Java programming language, which provides flexibility in selecting traffic flow features that can be calculated as well as adding new features. By using this tool, a total of 83 traffic flow features were extracted ($z1,\ldots,z83$). The extracted traffic flow characteristics are the result of the analysis and identification of relevant traffic characteristics for MTC traffic resulting from the research [20]. The reason is to collect as many features as possible in the initial set in order to determine in the later stages of the research (classification of SHIoT devices and anomaly detection) which independent features have the most significant influence on the change of the selected dependent feature.

Figure 4 shows the distribution of traffic flows (feature vectors), i.e., the share of traffic flows extracted from the collected traffic of SHIoT devices covered by the research.

The total number of collected traffic flows is 2,045,052. The presented feature vectors were used in the later phases, which include defining of SHIoT device classes and developing a SHIoT device classification model.

3.4 Defining classes of IoT devices

Identification of devices in the IoT environment is an important step and the basis for activities related to the security of the environment in which such devices exist, such as the detection of unauthorized activities, unauthorized devices within the network, malicious program code. The authors in the research [16] use the cluster method for the purpose of classifying 21 IoT devices whereby the devices are classified separately based on 11 features. Based on the identification of the device, research [28] seeks to detect unauthorized devices connected to the observed network. For this purpose, a total of 11 IoT devices was used, which are classified according to the semantic characteristics of the devices, i.e., their purpose (child monitoring devices, motion sensors, refrigerators, security cameras, smoke sensors, sockets, thermostats, televisions, clocks). A similar method of classification, based on the semantic characteristics of the device, is shown in research [37] in which the authors use a secondary dataset collected in [16]. The research included a total of 15 devices that are classified into four categories concerning the purpose of each device (concentrators, electronic devices, cameras, and sockets). Based on the analysis conducted by the research, the authors point out that the diversity of devices included in the data collection phase is more critical for the classification of SHIoT devices than the size of the dataset (the period of collection and amount of collected traffic).

From previous research, it is noticeable that the classification approaches so far are based mainly on semantic features, which means that the device classes are defined according to the application of such devices or their functionalities. The lack of such an approach for defining classes can be observed from the aspect of the dynamism of the smart home environment. According to the statistical indicators presented in [34], the number of SHIoT devices is continuously increasing, which is accompanied by an increase in the number of companies developing new solutions and new SHIoT devices. Therefore, SHIoT device classes need to be defined in a way that will apply to the upcoming SHIoT devices that will differ in functionality and application from the currently available devices.

3.4.1 Determining the traffic flow feature for the definition of device classes

The predictability of IoT device behavior is a phenomenon that is the result of the communication activities of IoT devices observed in research [15, 27, 38]. Since SHIoT devices possess a limited number of functionalities, specific devices will behave approximately equally in time according to the values

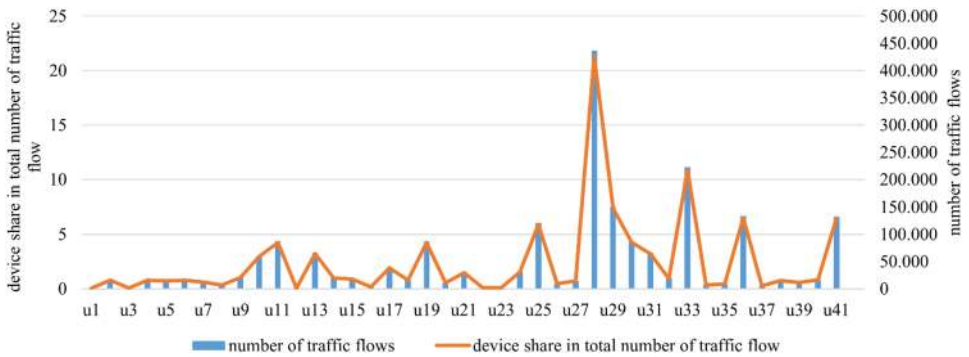

Fig. 4 Distribution of the number of traffic flows according to the SHIoT device

of the observed traffic characteristics. Unlike IoT devices, the conventional devices (smartphones, desktops, laptops, and servers) support the installation of a large number of applications where the communication activity of such devices depends on the end-users and the way the device is used. Accordingly, the index of the level of predictability of the behavior of IoT devices expressed by the coefficient of variation of the received and sent data (C_u index) is a measure based on which it is possible to determine the behavior of SHIoT devices in a certain period. The closer the index (C_u) is to 0, the smaller the deviation of the observed device in relation to the amount of received and sent data, and it is considered that the level of predictability of the behavior of such a device is higher than the device whose C_u index is farther than 0. All notation used in paper are shown in.

The C_u index was calculated for the mean values of consecutive traffic flows of an individual SHIoT device in 30 days according to expression (1).

$$C_u = CVar_u = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N} x_i} \tag{1}$$

where:

$C_u = CVar_u$ traffic predictability level index for SHIoT device *u;*

N total number of mean values of the ratio of received and sent traffic for consecutive traffic flows in period *T;*

x_i the amount of the mean value of the ratio of received and sent traffic volume for consecutive traffic flows.

In order to avoid the mean values to weigh 0, which is a problem of applying the method of the coefficient of variation, as normalized values of dispersion, traffic flows in which the ratio of received and sent data is equal to 0 are removed from the data set.

3.4.2 Defining IoT device classes based on coefficients of variation

To define the device classes based on the C_u index value, we used the method of coefficients of variation classification used in research [29, 39–42]. It assumes a normal distribution of data. Since the distribution of the obtained values (C_u index) is asymmetric (slanted to the left), the data are transformed. The data transformation method was selected using the Ladder of powers method (Tukey method), which clearly shows the appropriate data transformation function to achieve a normal distribution [43].

From the results obtained by the applied method, the suitability of the application of the logarithmic function for data transformation is observed, since in this case, it results in a normal distribution. The distribution of data is closest to normal the closer *chi2* is to 0, i.e., the closer *P* (*chi2*) is to 1. The normal distribution of the obtained data was confirmed by both the Shapiro–Wilk and Shapiro-Francia normality test, seen in Table 4, wherein both cases, $p > 0.05$ and the null hypothesis (that the values of the log (C_u) variable follow the normal distribution) cannot be rejected. Parameters *W* and *V* represent coefficients that indicate the deviation from the normal distribution of data where the value of $W \approx 1$ indicates the normal distribution of data, while *z* is a *z*-statistic that indicates how many standard deviations are observed data away from the mean value [44].

To apply the coefficients of variation classification method, the logarithmic values of C_u index were normalized by the min–max method according to expression (2):

$$C_{u(norm)} = \frac{\log\left(C_u\right) - \log(C_{u_{\min}})}{\log(C_{u_{\max}}) - \log\left(C_{u_{\min}}\right)} \tag{2}$$

where:

$C_{u(norm)}$ normalized value of a logarithmically transformed value C_u in the interval [0,1];

$\log\left(C_u\right)$ logarithmic value of C_u for device *u*;

$\log(C_{u_{\min}})$– minimum logarithmic value of C_u of all devices;

$\log(C_{u_{\max}})$ maximum logarithmic value of C_u of all devices.

After establishing the normal distribution of data and their normalization, the method of defining classes based on coefficients of variation was applied as a result of the mean values of the coefficients of variation and their standard deviation.

The mean value of the coefficient of variation was calculated according to expression (3):

$$A_{C_{u(norm)}} = \frac{1}{N}\sum_{u=1}^{n}\frac{C_{1(norm)} + C_{2(norm)} + \cdots + C_{n(norm)}}{N} \tag{3}$$

where:

$A_{C_{u(norm)}}$ arithmetic mean of the coefficients of variation of all devices;

Table 4 Results of Shapiro–Wilk and Shapiro-Francia normality tests

Variable	Observations	W	V	z	p > z
$\log(C_u)$ – Shapiro–Wilk	41	0.98831	0.471	− 1.588	0.94382
$\log(C_u)$ – Shapiro-Francia	41	0.98887	0.495	− 1.367	0.91420

N number of devices;

$C_{u(norm)}$ coefficient of variation of device u.

The standard deviation of the coefficients of variation was calculated according to expression (4):

$$\sigma_{C_{u(norm)}} = \sqrt{\frac{1}{N-1}\sum_{u=1}^{n}(C_{u(norm)} - \overline{C})^2} \quad (4)$$

where:

$\sigma_{C_{u(norm)}}$ standard deviation of the coefficients of variation of all devices;

N number of devices;

$C_{u(norm)}$ coefficient of variation of device u;

$\overline{C}$ arithmetic mean of the coefficients of variation of all devices.

Based on the previously performed data processing, a total of four classes of IoT devices were defined according to the method used in the research [41]. The first class includes devices where the condition is met $C_{u(norm)} \leq A_{C_{u(norm)}} - \sigma_{C_{u(norm)}}$. The second class includes devices that meet the condition $A_{C_{u(norm)}} - \sigma_{C_{u(norm)}} < C_{u(norm)} \leq \frac{A_{C_u} + \sigma_{C_u}}{2}$. The third class includes devices that meet the condition $\frac{A_{C_u} + \sigma_{C_u}}{2} < C_{u(norm)} \leq A_{C_u} + \sigma_{C_u}$, while the last class includes devices that satisfy the condition $C_{u(norm)} > A_{C_u} + \sigma_{C_u}$.

Values of C_u index, logarithmically transformed values, and min–max normalized values for each analyzed device are shown in Table 5. According to the data shown in Table 5, a total of four device classes was defined based on the values of the C_u index. The first class (C1) includes all devices whose logarithmically transformed, and normalized value of C_u index is $C_{u(norm)} \leq 0.253722$. The second class (C2) includes devices which met the condition $0.253722 < C_{u(norm)} > 0.354866$. The third class (C3) includes devices that met condition $0.354866 < C_{u(norm)} \leq 0.709732$ while the last class (C4) includes devices that met condition $C_{u(norm)} > 0.709732$.

Class C1 denotes IoT devices with a very high level of behavioral predictability since the coefficient of variation of the ratio of received and sent data is closest to 0. This means that such devices behave approximately equally over time from the aspect of the observed feature. If a Class C1 IoT device is used by a user, another device, or the environment, there will be no significant effect on the change in the C_u index value.

Class C2 combines devices with a high level of predictable behavior. If a device in the specified class is used by a user, another device, or the environment, it can result in minor changes to the ratio of received and sent data. Devices integrated into class C3 represent devices with a medium level of predictable behavior. The impact of user interaction, other devices, or the environment on the relationship between received and sent data can be significant. This behavior can result in additional functionality of the device that, at certain times, results in a larger amount of data in the incoming or outgoing direction.

The latest class (C4), combines SHIoT devices with a low level of predictable behavior. The use of such devices and their interaction with the user, other devices, or the environment significantly affects the relationship between the received and sent data. The reason is a significantly higher amount of data in the incoming direction (download) as a result of user requests. An example is seen with a device such as Google Chromecast, where video content is played at the user's request, which requires it to be downloaded via the Youtube service. This class also includes the Google Home mini, a smart speaker that can provide a variety of audio contents at the user request, which also causes a more considerable variation in the ratio of received and sent traffic.

Figure 5 shows an example of the behavioral relationships of four SHIoT devices (TPlink Day Night Cloud NC220 camera, NEST Protect Smoke Alarm, iRoobot Roomba 896, and Google Home mini) belonging to different classes for 1,000 consecutive traffic flows. There is a difference in the variation of the ratio of received and sent traffic ($C_{google\_mini} = 4.18$) in relation to the devices TPlink Day Night Cloud NC220 camera ($C_{tp\_link} = 0.042$), NEST Protect Smoke Alarm ($C_{nest\_smoke} = 0.19$), and iRoobot Roomba 896 ($C_{i896} = 0.37$).

For the development of a classification model based on the method of logistic regression improved by the concept of supervised machine learning, a dataset was formed containing the values of extracted characteristics of SHIoT devices traffic flows and belonging to the class of each device for each traffic flow in the dataset. The process of forming a dataset that contains aggregated data on the values of the characteristics of individual traffic flow and the affiliation of the traffic flow to the defined classes is shown by the UML (Unified Modeling Language) flow diagram in Fig. 6.

Each traffic flow is generated by a SHIoT device belonging to a particular class according to the classification shown in Table 5. Accordingly, each traffic flow is associated with a corresponding class, as shown in Table 6.

The extraction of traffic flow characteristics generated by an individual SHIoT device and the definition of SHIoT device classes are the basis for the formation of a data set of SHIoT device traffic flows to which class labels are associated.

4 Development of SHIoT device classification model

In order to develop a multiclass classification model of SHIoT devices, the logitboost method was used. The method used belongs to the ensemble machine learning methods

Table 5 Defined device classes according to C_u index value

No.#	Device	Index C_u	log(C_u) transformation	min–max normalization ($C_{u(norm)}$)	Class defition	Class label
1	tp_cam	0.042916917	− 1.36737	0	$C_{u(norm)} \leq A_{C_u} - \sigma_{C_u}$	C1
2	wiz_F3	0.075820416	− 1.12021	0.124242056		
3	hs105	0.076231674	− 1.11786	0.125423008		
4	wiz_B0	0.08086321	− 1.09225	0.138299504		
5	sams_st	0.123562483	− 0.90811	0.230861447		
6	w115	0.142241675	− 0.84697	0.261595627		
7	ihome	0.148887517	− 0.82714	0.271564558	$A_{C_u} - \sigma_{C_u} < C_{u(norm)} \leq \frac{A_{C_u} + \sigma_{C_u}}{2}$	C2
8	wit_baby	0.176239975	− 0.7539	0.308384178		
9	nest_smoke	0.192606687	− 0.71533	0.327771139		
10	ph_hue_2	0.200187894	− 0.69856	0.336199355		
11	cana_cam	0.209863653	− 0.67806	0.346504073		
12	hs110	0.24742122	− 0.60656	0.382445795	$\frac{A_{C_u} + \sigma_{C_u}}{2} < C_{u(norm)} \leq A_{C_u} + \sigma_{C_u}$	C3
13	belk_sw	0.254614637	− 0.59412	0.388702406		
14	wit_aura	0.261184872	− 0.58305	0.394264423		
15	w245	0.27041724	− 0.56797	0.401848085		
16	net_therm	0.290797956	− 0.53641	0.417711253		
17	amz_dot	0.318918293	− 0.49632	0.437862868		
18	blink_cam	0.344500361	− 0.46281	0.454707915		
19	sams_cam	0.34686605	− 0.45984	0.456201948		
20	lifx	0.346886878	− -0.45981	0.456215056		
21	smartw_cam	0.357559305	− 0.44665	0.462830477		
22	i895	0.358681004	− 0.44529	0.463514273		
23	i896	0.379012744	− 0.42135	0.475551248		
24	my_strom	0.432393144	− 0.36412	0.5043173		
25	inst_cam	0.479119397	− 0.31956	0.526719365		
26	bc_blood	0.479127026	− 0.31955	0.526722841		
27	net_weath	0.543491131	− 0.26481	0.554240633		
28	belk_cam	0.565787022	− 0.24735	0.563017747		
29	aug_door	0.610206124	− 0.21452	0.579517618		
30	amz_echo	0.632948837	− 0.19863	0.587506285		
31	belk_mot	0.724907331	− 0.13972	0.617121319		
32	net_cam	0.764635407	− 0.11655	0.628769456		
33	ph_hue_4	0.791347539	− 0.10163	0.636265899		
34	pix	0.958787396	− 0.01828	0.678167108		
35	wit_body	1.140461786	0.057081	0.716048538	$C_{u(norm)} > A_{C_u} + \sigma_{C_u}$	C4
36	google_chr	1.267801595	0.103051	0.739157175		
37	ring_vd	1.370122066	0.136759	0.756101612		
38	nest_cam	1.985562839	0.297884	0.837096166		
39	triby	2.468462951	0.392427	0.884621355		
40	aw_aq	2.553917945	0.407207	0.89205118		
41	google_mini	4.187473486	0.621952	1		

and is based on the statistical method of logistic regression. Ensembles combine several models, as shown in Fig. 7, with each model solving the original problem to obtain a better composite global model with better performance than using a single model [45].

Boosting belongs to a set of ensemble methods that can convert multiple "weak" classifiers (models that predict the target class depending on the values of the observed feature vectors) into "strong" classifiers. In general, a "weak" classifier is a model whose class prediction accuracy is slightly better than random guessing, while a strong classifier is

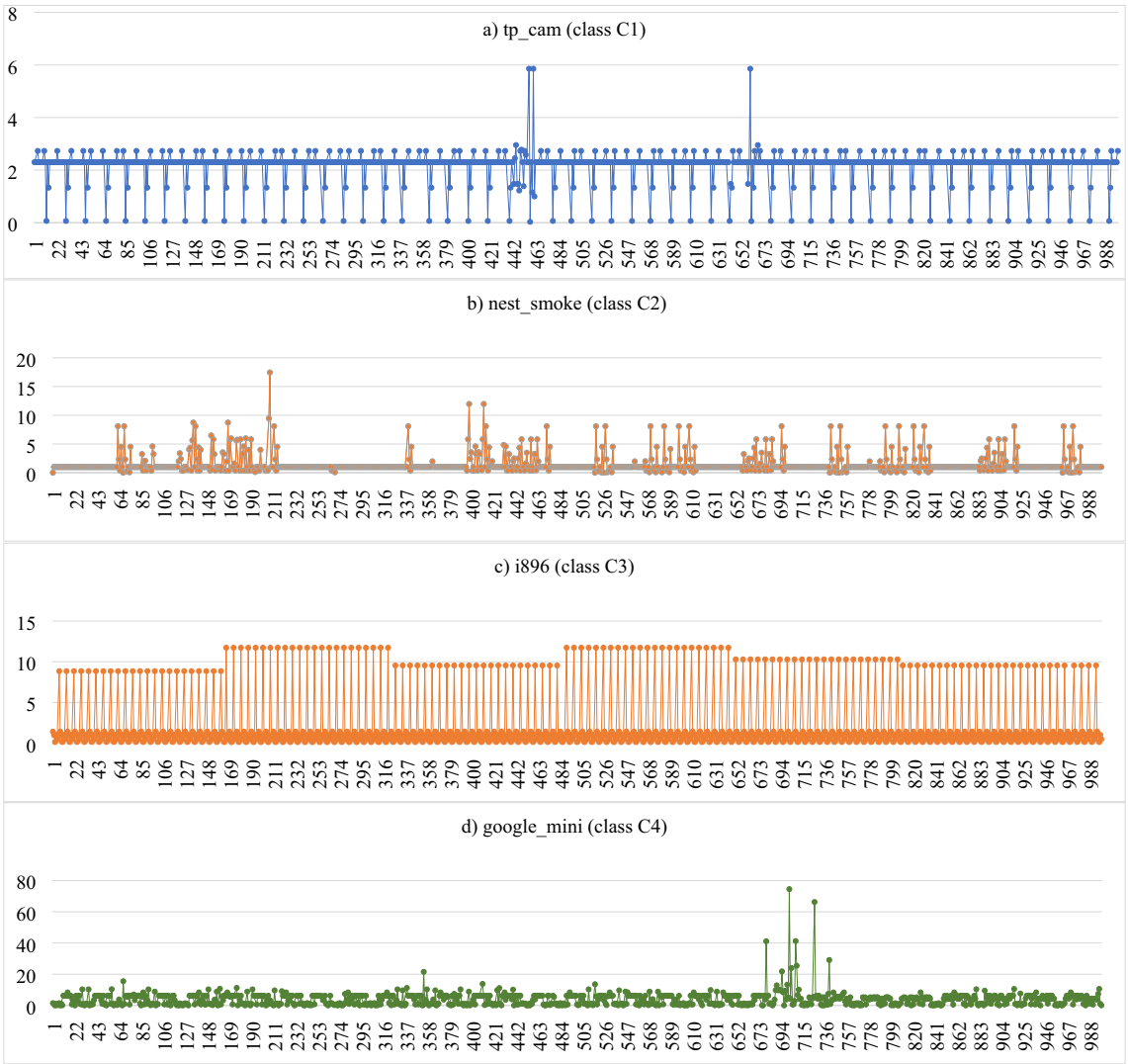

Fig. 5 Display of the difference in the behavior of four SHIoT devices in time according to the ratio of received and sent traffic for 1,000 consecutive traffic flows

characterized by near-ideal performance. Boosting methods have proven to be a suitable classification technique that provides excellent results in solving problems from different domains [46]. Given the classification problem that is being addressed and the proven effectiveness of the boosting group of machine learning methods, the logitboost method was used in this research.

4.1 Feature selection for development of SHIoT device classification model

Selecting the traffic characteristics generated by SHIoT devices is a crucial step in the process of developing a SHIoT device classification model. The importance of feature selection has been proven in numerous studies using statistical and machine learning methods, especially in the area of classification and regression. The aim is to identify a subset of the original feature set that is relevant to the classification problem being addressed and to remove those features that are irrelevant or redundant, thus reducing the dimensionality of the feature space as well as the entire dataset. The choice of features has a positive effect on the accuracy of the classification model, the speed of classification, and can reduce the occurrence of overfitting, which often leads to poor results in the validation process [47].

Features related to traffic flow identification (z1,.., z7) were preventively removed from the initial feature set to reduce its bias, a phenomenon that causes "wrong assumptions" during the model learning phase and results in a failure to identify the relevant relationships between independent and dependent features. Therefore, the initial set of independent features was reduced from 83 to 76.

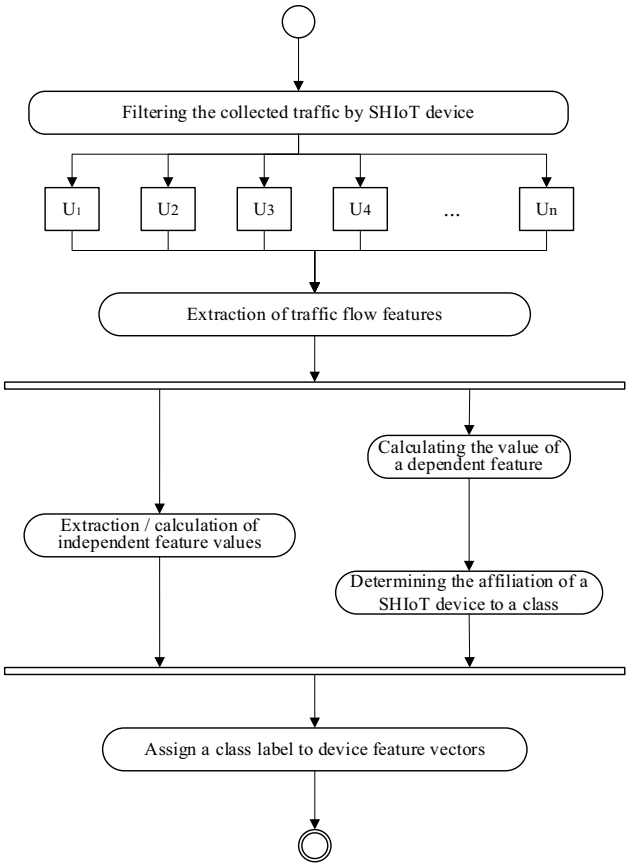

Fig. 6 UML activity diagram of the data set creation process

For the purpose of selecting features, the information gain (IG) method was used. The selected method is based on entropy and belongs to a set of feature ranking methods. This group of methods is characterized by simplicity and good results in practical applications, which is why it is often used in the process of selecting features in different domains such as text categorization, genome analysis, anomaly detection in communication networks, and bioinformatics [48–51].

According to [52], the IG method belongs to the measures based on correlation and serves to calculate the degree of correlation between the selected independent feature and the dependent feature (device class) and to evaluate the suitability of the feature for the classification purpose (goodness of feature). According to [53], an independent feature is appropriate if it is relevant to the observed dependent feature, but it is also not redundant with other relevant independent features. IG expresses a measure to reduce the uncertainty of identifying a dependent feature in the case where the value of the independent feature is unknown. The uncertainty calculation is based on information theory, and Shannon entropy to select those independent features that have the most significant impact on the dependent feature. The entropy of the dependent feature X is defined by expression (5) [53].

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i)) \tag{5}$$

where:

Table 6 Example and aggregation of traffic flows and class labels

No.#	Device label	z8	z9	z10	z11	z12	z13	z14	z15	z16	z17	z18	…	z83	class
1	u6	110,176,901	5	4	372	648	186	0	74	102	186	154	…	54,900,000	C1
2	u6	110,117,149	5	4	372	648	186	0	74	102	186	154	…	54,800,000	C1
3	u4	113,285,202	30	23	2012	3831	267	0	67	78	1460	0	…	5,740,188	C1
4	u4	8,334,253	3	2	96	80	32	32	32	0	48	32	…	5,269,207	C1
5	u4	79,698,183	25	21	1,805	4905	267	0	72	83	1460	0	…	5,767,895	C1
6	u11	9,383	7	1	2,156	308	308	308	308	0	308	308	…	0	C2
7	u11	1,649	3	1	924	308	308	308	308	0	308	308	…	0	C2
8	u10	4,785,250	17	1	5,104	296	305	296	300	4	296	296	…	0	C2
9	u10	4,795,180	17	1	5,104	296	305	296	300	4	296	296	…	0	C2
10	u10	90,304,076	4	5	0	74	0	0	0	0	39	0	…	30,000,000	C2
11	u19	2,461	1	3	33	143	33	33	33	0	61	33	…	0	C3
12	u19	2,337	1	3	33	143	33	33	33	0	61	33	…	0	C3
13	u30	108,109,786	4	1	596	149	149	149	149	0	149	149	…	27,000,000	C3
14	u30	108,113,739	4	1	636	159	159	159	159	0	159	159	…	27,000,000	C3
15	u30	119,577,279	9	11	164	246	41	0	18	22	41	0	…	29,100,000	C3
16	u41	141,088	4	7	454	2,881	357	28	114	162	1350	16	…	0	C4
17	u42	68,231	1	3	32	140	32	32	32	0	60	32	…	0	C4
18	u43	15,158,091	9	9	6,181	3,268	1,350	23	687	660	1,350	20	…	14,900	C4
19	u44	15,125,583	8	8	4,966	3,058	1,350	23	621	657	1,350	20	…	14,900	C4
20	u45	75,000,643	3	4	26	26	26	0	9	15	26	0	…	29,800	C4

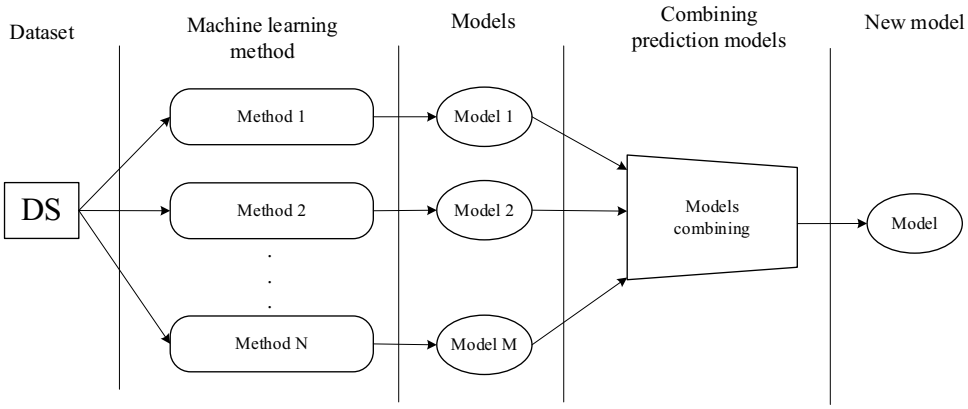

Fig. 7 Generalized presentation of the working principle of an ensemble machine learning method

H(X) entropy of dependent feature *X;*

$P(x_i)$ probability of occurrence of value x_i for feature *X.*

The entropy of the dependent feature *X*, after observing the value of the independent feature *Y*, is defined by expression (6).

$$H(X|Y) = -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j) log_2\left(P(x_i|y_j)\right) \quad (6)$$

where:

$P(y_i)$ probability of occurrence of value y_j for feature *Y;*

$P(x_i|y_j)$ conditional probability of feature *X* concerning values of feature *Y.*

The information gain reflects the amount by which the uncertainty of an individual value identification of the dependent feature *X* (device class) decreases with respect to the values of the observed independent feature *Y* according to the expression (7).

$$IG = H(X) - H(X|Y) \quad (7)$$

Since the dependent feature X can only take four values (four possible classes), the maximum value of IG is 2 (*log2X*). Therefore, the value obtained for an individual independent feature represents the amount of information of the independent feature, i.e., the amount by which the observed independent feature reduces the entropy (uncertainty) of the dependent feature. Table 7 shows the characteristics of the traffic flow with the expressed value of IG. From the presented table it can be seen that, for example, feature *z12* almost completely reduces the entropy of the dependent feature (IG = 1.832) while certain features (e.g. *z67, z39, z37*) do not contribute to the decrease of the entropy of the dependent feature (IG = 0).

Accordingly, the set of 76 has been further reduced to 58 independent features. In doing so, those features that satisfy the condition IG > 0 are considered. The obtained subset of features cannot be considered final since, in the development phase of the SHIoT device classification model, it is necessary to examine the model performance further if features with lower IG values are removed. The goal is to use the minimum set of features that gives the best performance of the classification model in order to reduce the time required to predict the class, reduce complexity, and reduce the occurrence of model bias.

4.2 Dataset used in the development of the classification model

The classification model, which aims to determine the class to which a device belongs based on the traffic flow characteristics it generates, is based on traffic flow characteristics collected over a period of 10 days for each device. The traffic flow feature vectors extracted for SHIoT devices are labeled with the appropriate class (Table 6). The number of traffic flows generated in the observed period depends on the characteristics of each SHIoT device [54].

The initial dataset, according to the above, has the characteristics of an unbalanced dataset and contains a total of 681,684 feature vectors distributed in four classes, according to Fig. 8. Therefore, before the development of the classification model, the number of traffic flows in the used dataset was balanced by stratification with the under-sampling of the majority represented class. Representation of traffic flows of an individual device in the initial dataset has been taken into account. The reason for this approach is the possibility of model bias occurring to the class that contains the largest number of feature vectors, and according to [55] it is necessary to stratify the classes before the model development. Following the stratification, the dataset contains 117,423 feature vectors used to further develop the classification model.

Table 7 Information gain values as the basis for selecting a subset of relevant independent features

No.#	IG value	Feature label	No.#	IG value	Feature label	No.#	IG value	Feature label
1	1.831	z71	27	1.2383	z27	53	0.171	z51
2	1.831	z12	28	1.2116	z15	54	0.1671	z54
3	1.7287	z11	29	1.2116	z60	55	0.1101	z58
4	1.7287	z69	30	1.2021	z20	56	0.1058	z53
5	1.7279	z17	31	1.2015	z80	57	0.1058	z38
6	1.7133	z46	32	1.1743	z31	58	0.0233	z50
7	1.6839	z13	33	1.1615	z29	59	0	z66
8	1.6836	z49	34	1.1105	z83	60	0	z37
9	1.6178	z25	35	1.1008	z36	61	0	z52
10	1.5472	z19	36	1.0894	z26	62	0	z39
11	1.5472	z61	37	1.0516	z42	63	0	z55
12	1.5247	z47	38	1.0341	z41	64	0	z40
13	1.5182	z30	39	0.997	z18	65	0	z75
14	1.5179	z35	40	0.9598	z21	66	0	z56
15	1.501	z48	41	0.9487	z79	67	0	z65
16	1.4204	z59	42	0.726	z78	68	0	z64
17	1.374	z73	43	0.6428	z68	69	0	z67
18	1.318	z23	44	0.6428	z9	70	0	z62
19	1.3006	z16	45	0.5977	z14	71	0	z72
20	1.29	z28	46	0.5942	z45	72	0	z57
21	1.2837	z33	47	0.5809	z74	73	0	z63
22	1.2761	z8	48	0.4579	z70	74	0	z81
23	1.2727	z24	49	0.4579	z10	75	0	z77
24	1.2675	z82	50	0.4032	z22	76	0	z76
25	1.2488	z32	51	0.3145	z44			
26	1.2467	z34	52	0.3103	z43			

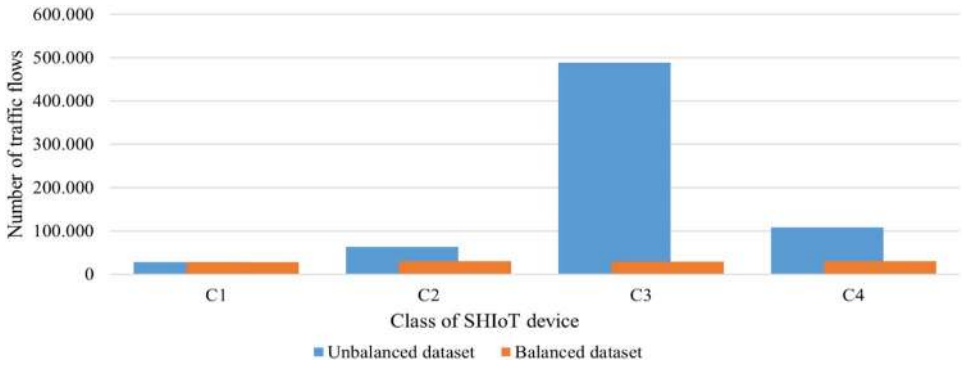

Fig. 8 Distribution of traffic flows according to SHIoT device classes

4.3 Application of the additive logistic regression method for multiclass classification of SHIoT devices

Additive logistic regression (logitboost) is a method of controlled machine learning that can be viewed as a generalization of the classical statistical method of logistic regression. The Logitboost method was developed in the year 2000 and presented in the research [56].

4.3.1 Logistic regression method

The logistic regression method models the conditional probability of belonging of the observed example to a particular class $Pr(G=j|X=x)$ for the J class, where it is possible to determine the classes of unknown examples according to expression (8).

$$j = \underset{j}{\operatorname{argmax}}\, Pr(G = j|X = x) \tag{8}$$

where:
j *j*-th class from the set of classes *G;*
G set of classes (1,…,*J*);
x independent feature from set *X;*
X a set of independent features.

Logistic regression models probabilities using linear functions in *x* while at the same time ensuring that their sum remains within limits [0,1]. The model is specified in terms of *J—1* log-odds that separate each class from the "basic" class *J* according to expressions (9, 10, 11).

$$\log \frac{Pr\left[G = j|X = x\right]}{Pr[G = J|X = x]} = \beta_j^T x_i; j = 1, .., J-1 \tag{9}$$

where:
β_j logistic coefficient of the independent feature for class *j;*

$$\Pr(G = j|X = x) = \frac{e^{\beta_j^T x_i}}{1 + \sum_{l=1}^{J-1} e^{\beta_l^T x_i}}; j = 1, .., J-1 \tag{10}$$

$$\Pr(G = J|X = x) = \frac{1}{1 + \sum_{l=1}^{J-1} e^{\beta_l^T x_i}} \tag{11}$$

Expression (9) implies a multiclass classification model in which x_i is the *i*-th feature vector, and *J* represents a class where $j \in \{0,1,2,..,J\text{-}1\}$ under condition $J \geq 3$. This model sets linear boundaries between areas corresponding to different classes. Thus, examples (x_i) that lie on the boundary between two classes (*j* and *J*) are those for which implies $Pr(G = j|X = x) = Pr\ (G = J|X = x)$ which is also the equivalent of $log\ odds = 0$. Adaptation of the logistic regression model involves estimating parameter β_j where the standard statistical procedure is to find the maximum of the likelihood function [57].

4.3.2 Logitboost method

In models based on logistic regression, there is no single method for estimating parameter β_j that would result in maximizing the plausibility function, but it is necessary to use the optimization methods. In this way, the maximum of the likelihood function is reached by an iterative procedure. Logitboost is one such method used in this study, and is based on the multinomial ordinal logistic regression method due to the existence of more than two dependent features whose values follow a natural sequence. In general, logitboost takes the form shown by expression (12).

$$\Pr(G = j|X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}; \sum_{k=1}^{J} F_k(x) = 0 \tag{12}$$

where:
$F_j(x)$ independent feature function (x).
Functions $F_j(x) = \sum_{m=1}^{M} f_{mj}(x)$ and f_{mj} are functions of independent features. In each iteration *m* ($m \in \{1,2,\ldots, M\}$), for example (x_i) that is misclassified, the weighting factor (w) increases while in the correctly classified example, the weighting factor decreases. In this way, the *m*-th "weak" classifier f_m focuses on examples that have been misclassified in previous iterations.

Logitboost method

1. Assign initial weights to examples $w_{i,j} = 1/N$, $i = 1,2,..,N$, $j = 0,1,..,J$, $F_j(x)=0$, $p_j(x) = 1/(J+1)\ \forall j$.
2. Repeat for $m=1,2,..,M$:
   - Repeat for $j=0,1,..,J$:
     - Calculate work responses and weights for class *j*

$$z_{ij} = \frac{y_{ij} - p_j(x_i)}{p_j(x_i)(1 - p_j(x_i))};$$

$$w_{ij} = p_j(x_i)(1 - p_j(x_i)).$$

     - Adjust the function $f_{mj}(x)$ with least squares weight regression of z_{ij} for x_i with weights w_{ij}
   - Set $f_{mj}(x) \leftarrow \frac{J-1}{J}\left(f_{mj}(x) - \frac{1}{J}\sum_{k=1}^{J} f_{mk}(x)\right)$, $F_j(x) \leftarrow F_j(x) + f_{mk}(x)$
   - Update $p_j(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}$
3. Output classification model $argmax_j\ F_j(x)$

Fig. 9 Logitboost method [46]

The output of the logitboost method is a set of $J+1$ response functions $\{F_j(x);\ j=0,1,\ldots,J\}$ as shown in Fig. 9. Each $F_j(x)$ is a linear combination of a set of "weak" classifiers.

5 Results analysis and discussion

Model development, testing, and validation were performed using the WEKA software tool with the support of MS Excel 2016 during the preparation of the dataset for model development. Because a total of 59 features were selected in the feature selection process using the information gain method, the number of features was gradually reduced during the model development when validation measures for each model were compared. This process aims to develop a model that will use the least possible number of independent features that will not significantly negatively affect its performance.

Each model was validated by k-fold cross-validation with $k=10$. The principle of operation of the k-fold cross-validation for $k=5$ is shown in Fig. 10. Cross-validation is a statistical method intended to assess the performance of machine learning models on new, unseen data. This method is used to assess the behavior of the model over data that was not used in the learning phase. In doing so, the model is applied k times iteratively over the dataset. In each iteration, the dataset is divided into k parts. One part of the set is used to validate the model, while the remaining k-1 parts of the set are combined into a subset for model learning.

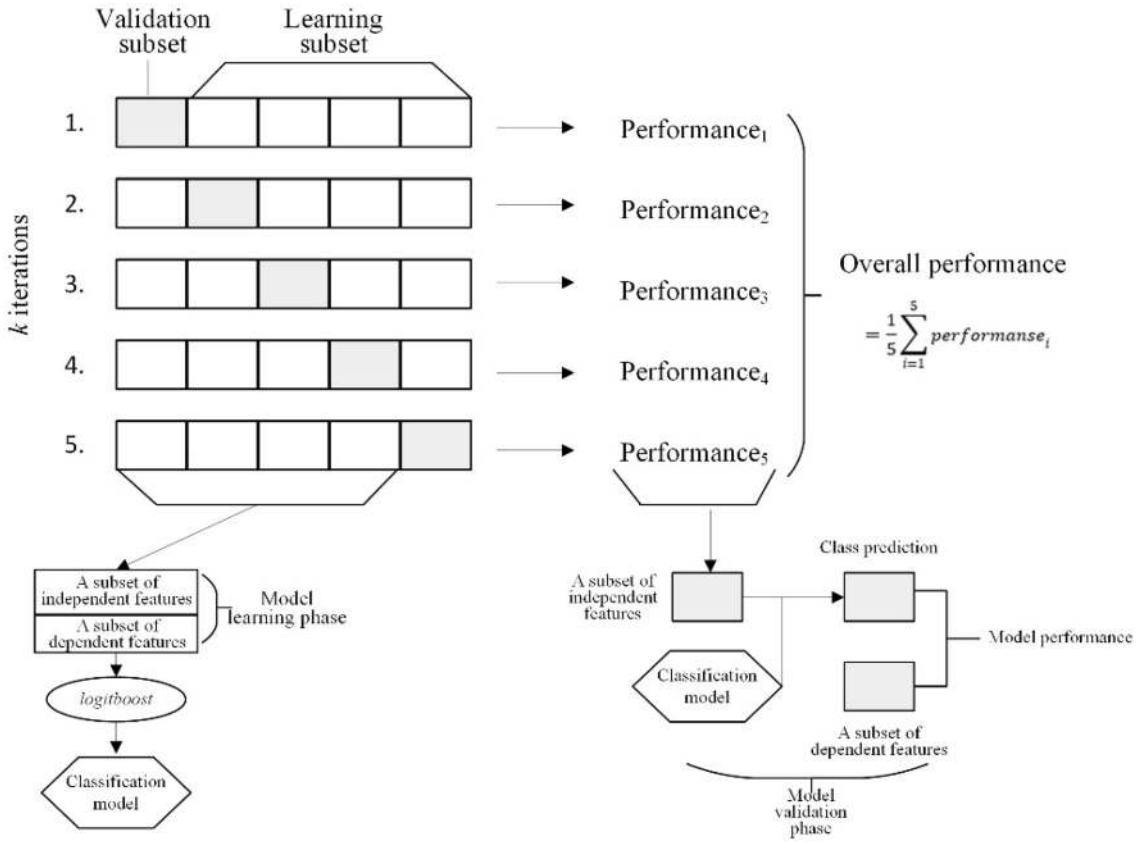

Fig. 10 Representation of k-fold cross-validation with $k=5$ [58]

Table 8 Performance representation of the SHIoT device classification model

	M1–59	M2–48	M3–33	M4–13	M5–8
Accurately classified examples	117,197 (99.8075%)	117,188 (99.7999%)	117,178 (99.7914%)	117,183 (99.7956%)	117,087 (99.7139%)
Misclassified examples	226 (0.1933%)	235 (0.2001%)	245 (0.2086%)	240 (0.2044%)	336 (0.2861%)
Kappa coefficient (κ)	0.9974	0.9973	0.9972	0.9973	0.9962
Total examples	117,423				

Table 9 Confusion matrix for classification model M4

Predicted class affiliation					
C1	C2	C3	C4		
28,068	7	16	54	C1	Actual class affiliation
12	29,831	3	7	C2	
18	3	29,331	28	C3	
57	12	23	29,953	C4	

Tables 8, 9, 10, 11 show the performance and results of validation measures for a total of five models (M1,…, M5) with a different number of independent features used (M1–59 features, M2–48 features, M3–33 features, M4–13 features and M5–8 features). Features were reduced to the lowest IG value (Table 7). The initial dataset was divided into a 70/30 ratio, where 70% of the examples in the set were used for model learning, while 30% were used for model testing. This division, along with 60/40 and 80/20, is common in the development of models based on machine learning methods [58].

The performance of a classification model based on machine learning needs to be expressed through several different measures, especially when the model is multiclass, given that each measure has advantages and limitations [59]. Accuracy is one of these measures that represents the share of accurately classified examples in the set of all examples according to expression (13) where TP (true positive examples), TN (true negative examples), FP (false positive examples) and FN (false negative examples).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where:

Acc proportion of accurately classified examples in the set of all examples;

TP number of true positive examples;

TN number of true negative examples;

FP number of false positive examples;

FN number of false negative examples.

Table 8 shows that all models have approximately the same classification accuracy (≈ 99.8%). The drop in accuracy is only noticeable with the M5 model, which uses eight independent features. The accuracy of the classification is 99.71% or 336 misclassified examples. The table shows a slight decrease in the accuracy of the classification for the M4 model (99.7956%) compared to the M1 model (99.8075%) which uses all 59 features.

Kappa coefficient (κ) expresses the measure of the success of the observed model according to the ideal with the correction of random selection [60]. The values of the kappa coefficient range from [0,1] where κ = 0–0.2 is an extremely bad model, κ = 0.2–0.39 is a bad model, κ = 0.4–0.59 is a moderate model, κ = 0.6–0.79 a good model, κ = 0.8–0.9 a very good model and κ = 0.9–1 an excellent model. According to the scale shown and the results seen in Table 8, all models show excellent characteristics according to the *kappa* coefficient, whereas the M4 model shows a minimal deviation from the M1 model (0.0001) with a significant reduction in the independent features used.

The accuracy of the model in predicting SHIoT device classes according to the traffic flow characteristics is also given by the confusion matrix shown in Table 9. Confusion matrix represents the performance measure for machine learning classification models where output can be two or more classes, representing the basis for other performance measures. In the confusion matrix shown in Table 9, the relation between traffic flow class affiliation

Table 10 Overview of model validation measures (TPR and FPR)

Class	True positive rate (TPR)					False positive rate (FPR)				
	M1–59	M2–48	M3–33	M4–13	M5–8	M1–59	M2–48	M3–33	M4–13	M5–8
C1	0.998	0.998	0.997	0.997	0.997	0.001	0.001	0.001	0.001	0.001
C2	0.999	0.999	0.999	0.999	0.999	0	0	0	0	0.001
C3	0.998	0.998	0.998	0.998	0.997	0.001	0	0.001	0	0.001
C4	0.997	0.997	0.997	0.997	0.996	0.001	0.001	0.001	0.001	0.001

Table 11 Overview of model validation measures (F-measure and precision)

Class	Precision					*F*-measure (*F1* score)				
	M1–59	M2–48	M3–33	M4–13	M5–8	M1–59	M2–48	M3–33	M4–13	M5–8
C1	0.998	0.997	0.997	0.997	0.995	0.998	0.997	0.997	0.997	0.996
C2	0.999	0.999	0.999	0.999	0.998	0.999	0.999	0.999	0.999	0.999
C3	0.998	0.999	0.998	0.999	0.998	0.998	0.998	0.998	0.998	0.998
C4	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.996

predicted by the developed model and actual class affiliation of observed traffic flow is visible. Accordingly, a high number of traffic flows whose class affiliation is accurately predicted in relation to the number of instances whose class affiliation is incorrectly predicted is observed.

Additional validation measures are expressed through sensitivity, i.e., the rate or frequency of TPR and the rate or frequency of FPR are shown in Table 10.

The true positive rate results represent accurately classified examples of a class in the set of all examples assigned to that class according to expression (14). The false positive example rate represents the ratio of misclassified class examples in the set of all examples assigned to that class to expression (15)

$$TPR = \frac{TP}{TP + FN} \tag{14}$$

where:
TPR true positive rate;

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

where:
FPR false positive rate;

Table 10 shows that models M1 and M2 provide the best results according to the TPR measure for class C1, for class C2 all observed models provide the same results, while for classes C3 and C4 model M5 provides slightly worse results compared to the others. From the aspect of FPR measures, the M2 and M4 models provide better or equally good results compared to other models, with the M5 model providing the worst results.

Additional validation measures that show the quality of the classification model are the precision or positive prediction value (PPV) and the *F*-measure shown in Table 10, as well as ROC and PRC (Precision-Recall Curve) curves whose values are shown in Table 12.

The measure of precision is used to express the number of correctly classified examples in relation to the total number of examples belonging to that class according to expression (16).

$$PPV = \frac{TP}{TP + FP} \tag{16}$$

where:
PPV positive prediction value

According to the values expressed in Table 11, it can be seen that for class C1 the best results are given by model M1 while the worst results are visible with model M5. For classes C2 and C4, equally good results are observed for all models with the exception of the M5 model for the C2 class, while for the C3 class, the M2 and M4 models provide the best results.

The *F*-measure or *F1* score represents the harmonic mean of the precision measures and the TPR according to expression (17) [59]. According to [61], the harmonic mean is more intuitive than the classical arithmetic mean to calculate the mean of the ratio.

$$F1 = \frac{2(PPV \cdot TPR)}{PPV + TPR} \tag{17}$$

The calculated values of the F-measure shown in Table 11 indicate the M5 model as the worst observed from the aspect of classes C1 and C4 while the other models show approximately the same results.

Table 12 shows the values of ROC and PRC validation measures. The ROC curve, or AUROC (Area Under the ROC Curve), is one of the most important and most frequently used measures that show the quality of the classification model.

ROC is, although in this case, expressed in tabular form, a graphical representation of the relationship between the rate of true positive classifications (TPR) and specificity, i.e. the rate of true negative classifications (TNR = 1–FPR). An example of a graphical representation of the ROC curve is seen in Fig. 11 for the M4 model. The area under the curve, AUROC, is interpreted as the average TPR value for all TNR values in the interval [0,1]. The closer the AUROC is to the value of 1, the better the performance of the classification model. A lower value of 0.5 represents the performance of the model equal to random guessing [8]. The values shown in Table 12 indicate the excellent performance of all observed models, i.e. almost all AUROC values are very close to the value of 1.

Table 12 Overview of model validation measures (ROC and PRC)

Class	ROC					PRC				
	M1–59	M2–48	M3–33	M4–13	M5–8	M1–59	M2–48	M3–33	M4–13	M5–8
C1	0.9999	0.9999	0.9998	0.9997	0.9998	0.999	1	0.999	0.998	0.999
C2	1	1	1	0.9999	0.9998	1	1	1	1	1
C3	0.9997	0.9999	0.9999	0.9998	0.9998	1	1	1	1	1
C4	0.9998	0.9999	0.9998	0.9996	0.9997	1	1	0.999	0.999	0.999

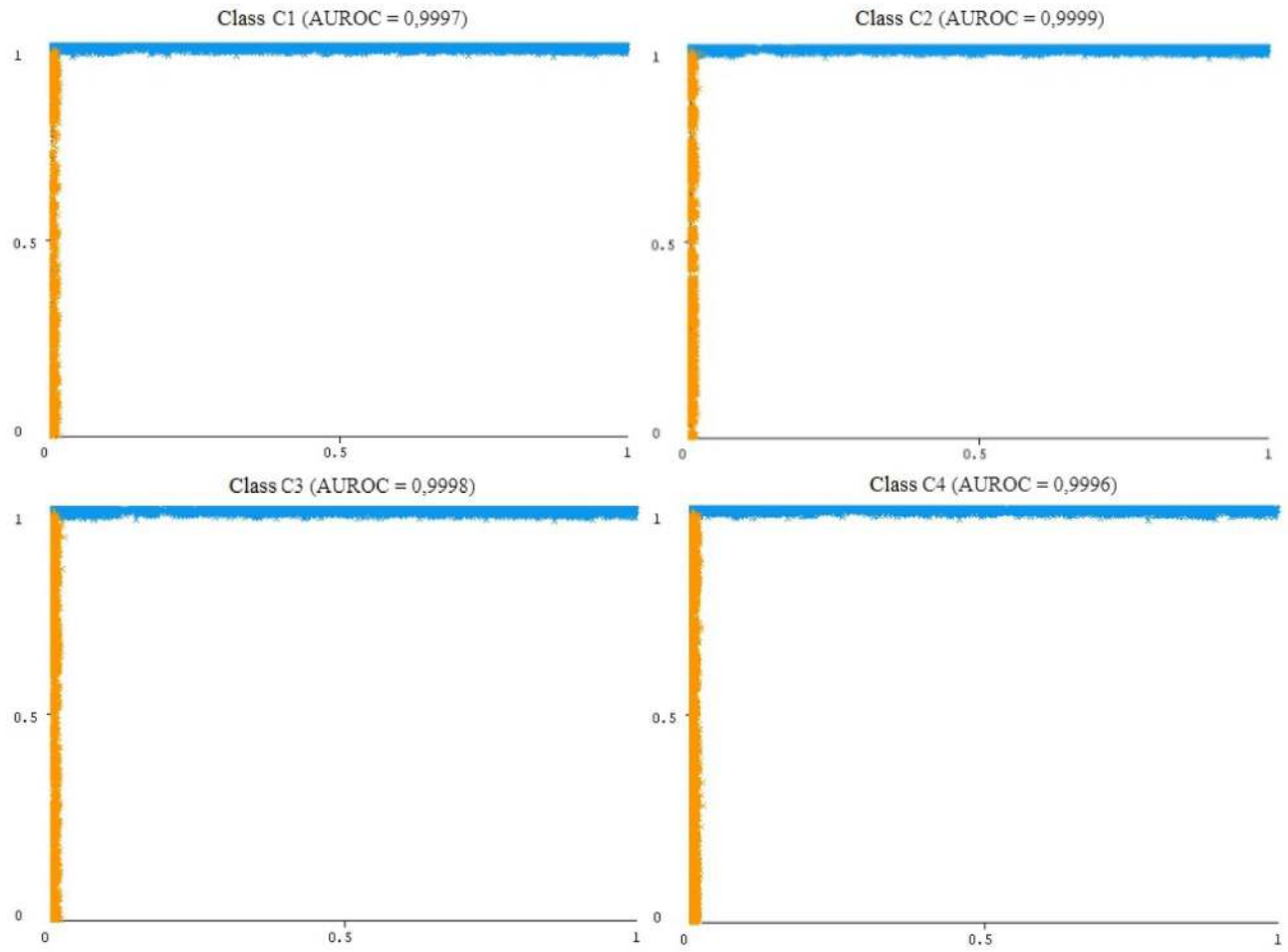

Fig. 11 ROC curve representation for classification model M4

An alternative measure to the ROC is the PRC (Precision-Recall Curve), which is often used in cases of unbalanced datasets, whereas in the ROC measure a significant change in the number of false positively classified examples may result in small change in the rate of false positively classified examples. Therefore, since PRC uses the ratio of PPV and TPR, i.e. focuses on positively classified examples (TP and FP), it can better demonstrate the impact of many negative examples on the model performance. Because the dataset was stratified in this research, the PRC measure gives almost the same results as the ROC measure for all the observed models.

According to the analysis of the results, the M4 model was selected as the optimal model of SHIoT device classification, considering that its performance according to all presented measures does not deviate significantly from other observed models (TPR 0–0.001, FPR 0–0.001, PPV 0–0.001, F1 0–0.001, ROC 0.0001–0.0003 and PRC 0–0.002) with a significant reduction in the independent features used. In the M4 model, a total of 13 independent features were used compared to the initial 59, which, according to the IG value, had some influence on the dependent feature. The independent features used are shown in Table 13.

As can be seen from the table, the most relevant features are information-related to the length of packets in the observed traffic flow (z11–total length of sent packets in traffic flow; z12–total length of received packets; z13–the maximum length of sent packets; z17–the maximum length of received packets, z19–mean value of received packet length, z46–maximum packet length, z47–mean packet length value, z49–packet length variation), then information on interarrival packet times in traffic flow (z25–maximum interarrival packet time in traffic flow; z30–the maximum time between two consecutive packets sent in traffic flow). Features that provide information on the segments in the traffic flow (z61–the average size of the received segment), as well as features that provide information on the amount of data transmitted in the sub-stream (z60–the amount of data sent in the sub-stream; z71–the amount of data received in the sub-stream), proved to be relevant (Table 14).

Table 13 Independent features used in the process of developing the classification model

A subset of features	Number of features	Feature label
Initial set	83	z1, z2, z3, z4, z5, z6, z7, z8, z9, z10, z11, z12, z13, z14, z15, z16, z17, z18, z19, z20, z21, z22, z23, z24, z25, z26, z27, z28, z29, z30,z31, z32, z33, z34, z35, z36, z37, z38, z39, z40, z41, z42, z43, z44, z45, z46, z47, z48, z49, z50, z51, z52, z53, z54, z55, z56, z57, z58, z59, z60, z61, z62, z63, z64, z65, z66, z67, z68, z69, z70, z71, z72, z73, z74, z75, z76, z77, z78, z79, z80, z81, z82, z83
Information gain	59	z9, z10, z11, z12, z13, z14, z15, z16, z17, z18, z19, z20, z21, z22, z23, z24, z25, z26, z27, z28, z29, z30, z31, z32, z33, z34, z35, z36, z38, z41, z42, z43, z44, z45, z46, z47, z48, z49, z50, z51, z53, z54, z58, z59, z60, z61, z68, z69, z70, z71, z73, z74, z78, z79, z8, z80, z82, z83
Model M4	13	z11, z12, z13, z17, z19, z25, z30, z46, z47, z49, z61, z69, z71

Table 14 Notation used in paper

Notation	Description
C_u	Coefficient of variation of the received and sent data
$\mathrm{CVar_u}$	Traffic predictability level index for SHIoT device
x_i	The amount of the mean value of the ratio of received and sent traffic volume for consecutive traffic flows
$C_{u(norm)}$	Normalized value of a logarithmically transformed value $\mathrm{C_u}$ in the interval [0,1]
$\log(C_u)$	Logarithmic value of C_u for device u;
$\log\left(C_{u_{min}}\right)$	Minimum logarithmic value of C_u of all devices
$\log\left(C_{u_{max}}\right)$	Maximum logarithmic value of C_u of all devices
$\mathrm{A}_{C_{u(norm)}}$	Arithmetic mean of the coefficients of variation of all devices
$\sigma_{C_{u(norm)}}$	Standard deviation of the coefficients of variation of all devices
$\overline{C}$	Arithmetic mean of the coefficients of variation of all devices
$H(X)$	Entropy of dependent feature X
$P(x_i)$	Probability of occurrence of value x_i for feature X
$P(y_i)$	Probability of occurrence of value y_j for feature Y
$P(x_i\|y_j)$	Conditional probability of feature X concerning values of feature Y
j	j-th class from the set of classes G
G	set of classes (1,…,J);
β_j	Logistic coefficient of the independent feature for class j
$F_j(x)$	Independent feature function (x)

6 Conclusion and future work

The research presented in this paper provides a new approach in observing the behavior of IoT devices based on the generated network traffic. The goal, which was achieved by this research, was to develop an effective model of IoT device classification in a smart home environment, which is based on the outgoing and incoming traffic ratio coefficient of variation as a measure of device behavior predictability. The basis of the research and achieving the defined goal is the scientific hypothesis that it is possible to define classes of IoT devices and develop an effective classification model based on supervised machine learning methods acknowledging traffic characteristics generated by IoT devices in a smart home environment. The scientific hypothesis was proved by defining four classes of devices based on the coefficient of variation ratio of received and sent traffic. The defined classes conditioned the development of the classification model of IoT devices.

The mentioned coefficient was named C_u index, which was chosen as a dependent feature used for the purpose of defining a total of four classes of SHIoT devices using the method of the coefficient of variation classification. Based on the defined classes of SHIoT devices, a multiclass classification model based on the boosting method of additive logistic regression as a machine learning method was developed, which according to all validation measures,

shows high performance. The accuracy of SHIoT device recognition to one of the defined classes based on independent traffic flow features is 99.79%. Relevant independent traffic flow features used in the development of the model were selected using the information gain method.

The research proved that it is possible to assign, with high accuracy, new and unseen devices, and traffic flow that they generate into predefined classes with the application of boosting methods of machine learning. Besides, this approach responds to the needs of the newly created IoT environment in which the number of devices is growing exponentially, and it is not possible (or requires substantial resources) to know traffic profiles for each device, but it is sufficient to identify which class the device belongs to. This innovative approach has the potential to lay the foundations for many other activities and research in the IoT concept problem domain. The detection of anomalies in the communication network caused by IoT devices is one of the future research directions that will use findings and conclusions gathered in this research. Besides further research, developed model can have real-life applications as a software solution that can upgrade functionalities of the existing solutions for the device and network monitoring and management in an environment where many various IoT devices exist. This solution can help to monitor device groups that have similar communication patterns, manage their behavior in the network, plan future communication capacities for various device classes or similar activities. This kind of a solution is only usable as a support process for various other activities. That is why future research, as well as future real-life applications and service, will have a foundation in the results achieved in this research.

Author contributions Conceptualization: IC and DP; Methodology: IC, DP, MP and BG; Data curation: IC; Formal analysis: IC, DP and MP; Funding acquisition: DP and MP; Investigation: IC; Supervision: DP, BG; Validation: IC, DP and BG; Visualization: IC and MP; Writing – original draft: IC; Writing – review & editing: DP, MP and BG.

Funding This research is funded by the University of Zagreb through the Grants for core financing of scientific and artistic activities of the University of Zagreb in academic year 2019/2020 under the project (555-1) "Challenges of Information and Communication Networks and Technologies, services, and user equipment in establishing the Society 5.0 environment."

Data availability Collected and preprocessed dataset segment can be found on Kaggle (www.kaggle.com/dataset/5cae54f093f90a5a0613542573a16ab7c3f5dbf271cac549578266e7c632d7f9).

Code availability Not applicable.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. DigiCert Inc (2018) State of IoT Security Survey 2018. https://www.digicert.com/wp-content/uploads/2018/11/StateOfIoTSecurity_Report_11_02_18_F_am.pdfAccessed 18 Mar 2020
2. Vodafone Business (2019) Your IoT-Driven Future [Internet] 2019. https://www.vodafone.com/business/news-and-insights/whitepaper/vodafone-iot-barometer-2019 Accessed 14 May 2020
3. Kejriwal S, Mahajan S (2016) Smart buildings: how IoT technology aims to add value for real estate companies. Deloitte University Press, London
4. HIS (2017) The Internet of Things : a movement, not a market Start revolutionizing the competitive landscape. IHS Markit. https://ihsmarkit.com/Info/1017/internet-of-things.html Accessed 21 Mar 2020
5. Adat V, Gupta BB (2018) Security in Internet of Things: issues, challenges, taxonomy, and architecture. Telecommunication Systems 67(3):423–441
6. Li D, Deng L et al (2019) A novel CNN based security guaranteed image watermarking generation scenario for smart city applications. Information Sciences 479:432–447
7. Hossain MS, Muhammad G, Abdul W, Song B et al (2018) Cloud-assisted secure video transmission and sharing framework for smart cities. Future Generation Computer Systems 83:596–606
8. Sedgwick P (2011) Receiver operating characteristic curves. BMJ 343:d4302–d4302. https://doi.org/10.1136/bmj.d4302
9. Tewari A, Gupta BB (2020) Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework. Future generation computer systems 108:909–920
10. Statista (2018) Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions) [Internet]. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. Accessed Jun 24 2018
11. Statista (2018) The Internet of Things (IoT)* units installed base by category from 2014 to 2020 (in billions): https://www.statista.com/statistics/370350/internet-of-things-installed-base-by-category/. Accessed 24 Jun 2018
12. Cvitić I, Vujić M, Husnjak S (2016) Classification of security risks in the IoT environment. In: DAAAM International Symposium on Intelligent Manufacturing and Automation, pp 0731–40
13. Ali B, Awad A (2018) Cyber and physical security vulnerability assessment for IoT-based smart homes. Sensors 18:817
14. Al-Shammari BKJ, Al-Aboody N, Al-Raweshidy HS (2018) IoT traffic management and integration in the QoS supported network. IEEE Internet Things J 5:352–370
15. Meidan Y, Bohadana M, Mathov Y, Mirsky Y, Breitenbacher D, Shabtai A et al (2018) N-BaIoT: network-based detection of IoT botnet attacks using deep autoencoders. IEEE Pervasive Comput 13:1–8

16. Sivanathan A, Sherratt D, Gharakheili HH, Radford A, Wijenayake C, Vishwanath A, et al (2017) Characterizing and classifying IoT traffic in smart cities and campuses. In: IEEE International Conference on Pervasive Computing and Communications Work (INFOCOM WKSHPS). IEEE, pp 559–64
17. Sivanathan A, Gharakheili HH, Loi F, Radford A, Wijenayake C, Vishwanath A et al (2019) Classifying IoT devices in smart environments using network traffic characteristics. IEEE Trans Mob Comput 18:1745–1759
18. Shafiq MZ, Ji L, Liu AX, Pang J, Wang J (2013) Large-scale measurement and characterization of cellular machine-to-machine traffic. IEEE/ACM Trans Netw 21:1960–1973
19. Bekerman D, Shapira B, Rokach L, Bar A (2015) Unknown malware detection using network traffic classification. In: 2015 IEEE Conference on Communications and Network Security. IEEE, pp 134–42
20. Cvitić I, Zorić P, Kuljanić TM, Musa M (2019) Analysis of network traffic features generated by IoT DEVICES. In: Radojičić V, Bojović N, Marković D, Marković G (eds) XXXVII Simp o novim Tehnol u poštanskom i Telekomun saobraćaju–PosTel 2019. Univerzitet u Beogradu-Saobraćajni fakultet, Beograd, pp 193–200
21. Trimananda R, Varmarken J, Markopoulou A, Demsky B (2020) Packet-level signatures for smart home devices. In: Proceedings 2020 Network and Distributed System Security Symposium. Reston, VA: Internet Society.
22. Charyyev B, Gunes MH (2020) IoT traffic flow identification using locality sensitive hashes. ICC 2020. IEEE International Conference on Communications [Internet]. IEEE, pp 1–6
23. Kotak J, Elovici Y (2020) IoT device identification using deep learning. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-030-57805-3_8
24. Anthi E, Williams L, Slowinska M, Theodorakopoulos G, Burnap P (2019) A supervised intrusion detection system for smart home IOT devices. IEEE Internet Things J 6:9042–9053
25. Salman O, Elhajj IH, Chehab A, Kayssi A (2019) A machine learning based framework for IoT device identification and abnormal traffic detection. Trans Emerg Telecommun. https://doi.org/10.1002/ett.3743
26. Karimi AM, Niyaz Q, Weiqing Sun, Javaid AY, Devabhaktuni VK (2016) Distributed network traffic feature extraction for a real-time IDS. 2016 IEEE International Conference on Electro Information Technology. IEEE, pp. 0522–6
27. Amar Y, Haddadi H, Mortier R, Brown A, Colley J, Crabtree A (2018) An Analysis of Home IoT Network Traffic and Behaviour. arXiv. http://arxiv.org/abs/1803.05368 Accessed 14 Mar 2020
28. Meidan Y, Bohadana M, Shabtai A, Ochoa M, Tippenhauer NO, Guarnizo JD, et al. (2017) Detection of unauthorized IoT devices using machine learning techniques. arXiv. http://arxiv.org/abs/1709.04647 Accessed 23 Sep 2019
29. Cvitic I, Perakovic D, Perisa M, Botica M (2020) Definition of the IoT device classes based on network traffic flow features. Springer, Cham, pp 1–17
30. Hamza A, Ranathunga D, Gharakheili HH, Roughan M, Sivaraman V (2018) Clear as MUD: generating, validating and applying iot behaviorial profiles. Proceedings of 2018 Work IoT Secur Priv - IoT S&P '18. ACM Press, New York, pp 8–14
31. Hamza A, Gharakheili HH, Sivaraman V (2018) Combining MUD policies with SDN for IoT intrusion detection. Proceedings of 2018 work IoT Secur Priv - IoT S&P '18. ACM Press, New York, pp 1–7
32. Iskon (2016) Iskon smarthome. https://smarthome.iskon.hr/statiс/pdf/Upute_basic_paket_SH31_5_B.pdf. Accessed 30 Sep 2019
33. Hrvatska A (2019) Smart home uređaji.. Available from: https://www.a1.hr/privatni/promocije/smarthome/uredaji. Accessed 30 Sep 2019
34. Smart Home-worldwide (2019) Statista market forecast. https://www.statista.com/outlook/279/100/smart-home/worldwide. Accessed 09 Mar 2019
35. Aghaei-Foroushani V, Zincir-Heywood AN (2015) A proxy identifier based on patterns in traffic flows. In: 2015 IEEE 16th International Symposium on High Assurance Systems Engineering. IEEE, pp 118–25
36. Habibi Lashkari A, Draper Gil G, Mamun MSI, Ghorbani AA (2017) Characterization of tor traffic using time based features. In: Proceedings 3rd International Conference on Information System Security and Privacy. SCITEPRESS-Science and Technology Publications, pp 253–62
37. Bai L, Yao L, Kanhere SS, Wang X, Yang Z (2018) Automatic device classification from network traffic streams of internet of things. 2018 IEEE 43rd Conference on Local Computing Networks. IEEE, pp 1–9
38. Doshi R, Apthorpe N, Feamster N (2018) Machine learning DDoS detection for consumer internet of things devices. 2018 IEEE Security and Privacy Work. IEEE, pp 29–35
39. Vaz MAB, Pacheco PS, Seidel EJ, Ansuj AP (2017) Classification of the coefficient of variation to variables in beef cattle experiments. Ciência Rural 47:9–12
40. Couto MF, Peternelli LA, Barbosa MHP (2017) Classification of the coefficients of variation for sugarcane crops. Ciência Rural 43:957–961
41. Romano FL, Ambrosano GMB, de Magnani MBB, Nouer DF (2005) Analysis of the coefficient of variation in shear and tensile bond strength tests. J Appl Oral Sci 13:243–246
42. de Ferreira AASN, Dourado LRB, Biagiotti D, da Santos NP, Nascimento DCN, Sousa KRS (2019) Methods for classifying coefficients of variation in experimentation with poultrys. Comun Sci 9:565–574
43. Ernst PA, Thompson JR, Miao Y (2017) Tukey's transformational ladder for portfolio management. Financ Mark Portf Manag 31:317–355. https://doi.org/10.1007/s11408-017-0292-1
44. Hanusz Z, Tarasińska J (2015) Normalization of the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality. Biometrical Lett 52:85–93
45. Rokach L (2006) Ensemble methods for classifiers. Data Min Knowl Discov Handb. https://doi.org/10.1007/0-387-25465-X_45
46. Zhou SK, Park JH, Georgescu B, Comaniciu D, Simopoulos C, Otsuki J (2006) Image-Based multiclass boosting and echocardiographic view classification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp 1559–65
47. Egea S, Rego Manez A, Carro B, Sanchez-Esguevillas A, Lloret J (2018) Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments. IEEE Internet Things J 5:1616–1624
48. Zainal A, Maarof MA, Shamsuddin SM (2006) Feature selection using rough set in intrusion detection. TENCON 2006 IEEE Reg 10 Conference. IEEE, pp. 1–4
49. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
50. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31:91–103
51. Jovic A, Brkic K, Bogunovic N (2015) A review of feature selection methods with applications. 2015 38th International Conference on Information and Communication Technology. IEEE, pp 1200–5
52. Osanaiye O, Choo K-KR, Dlodlo M (2016) Analysing feature selection and classification techniques for DDoS detection in cloud. Southern Africa Telecommunication Networks and Applications Conference, pp 198–203

53. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. Proceedings of the Twenty-First International Conference on Machine Learning, pp 856–63
54. Cvitić I (2020) Network traffic anomaly detection based on traffic characteristics and device class affiliation. Doctoral thesis. University of Zagreb.
55. Longadge R, Dongre S (2013) Class imbalance problem in data mining review. Eur J Intern Med 24:256
56. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28:337–407
57. Landwehr N, Hall M, Frank E (2005) Logistic Model Trees. Mach Learn 59:161–205
58. Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. arXiv. http://arxiv.org/abs/1811.12808 Accessed 13 Nov 2020
59. Hossin M, Sulaiman M (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5:01–11
60. Seletković A, Pernar R, Jazbec A, Ančić M (2008) Točnost klasifikacije satelitske snimke visoke prostorne rezolucije ikonos za potrebe šumarstva. Šumarski List 588:393–404
61. Sasaki Y (2007) The truth of the F-measure. Teach Tutor Mater 1:1–6