ISSN: 0067-2904

An Internet of Things Botnet Detection Model Using Regression Analysis and Linear Discrimination Analysis

Manar J. Gatea , Sarab M. Hameed *
Department Computer Science, College of Science, University of Baghdad, Iraq

Received: 29/10/2021 Accepted: 9/1/2022 Published: 30/10/2022

Abstract:

The Internet of Things (IoT) has become a hot area of research in recent years due to the significant advancements in the semiconductor industry, wireless communication technologies, and the realization of its ability in numerous applications such as smart homes, health care, control systems, and military. Furthermore, IoT devices inefficient security has led to an increase cybersecurity risks such as IoT botnets, which have become a serious threat. To counter this threat there is a need to develop a model for detecting IoT botnets.

This paper's contribution is to formulate the IoT botnet detection problem and introduce multiple linear regression (MLR) for modelling IoT botnet features with discriminating capability and alleviating the challenges of IoT detection. In addition, a linear discrimination analysis (LDA) model for distinguishing between normal activities and IoT botnets was developed.

Network-based detection of IoT (N-BaIoT) dataset was used to evaluate the performance of the proposed IoT botnet detection model in terms of accuracy, precision, and detection rate. Experimental results revealed that the proposed IoT botnet detection model provides a relevant feature subset and preserves high accuracy when compared with state-of-the-art and baseline methods, particularly LDA.

Keywords: Botnets, botnet detection, IoT botnet, linear discrimination analysis, regression analysis.

نموذج اكتشاف الروبوتات لإنترنت الأشياء باستخدام تحليل الانحدار وتحليل التمييز الخطي

منار جبار، سراب مجيد حميد *
قسم علوم الحاسوب ، كلية العلوم ، جامعة بغداد ، العراق

الخلاصة:

أصبحت إنترنت الأشياء (IoT) مجال بحث فاعل في السنوات الأخيرة بسبب التقدم الكبير في صناعة أشباه الموصلات وتقنيات الاتصالات اللاسلكية وإدراك قدرتها في العديد من التطبيقات مثل المنازل الذكية والرعاية الصحية وأنظمة التحكم ، والجيش. علاوة على ذلك ، أدى الافتقار إلى أمان أجهزة إنترنت الأشياء إلى زيادة مخاطر الأمن السيبراني مثل شبكات روبوت إنترنت الأشياء التي أصبحت تمثل تهديدًا خطيرًا. لمواجهة هذا التهديد ، هناك حاجة لتطوير نموذج للكشف عن شبكات روبوت إنترنت الأشياء.

*Email: sarab.m@sc.uobaghdad.edu.iq

مساهمة هذا البحث في صياغة مشكلة اكتشاف شبكة الروبوتات لإنترنت الأشياء وإدخال الانحدار الخطي المتعدد (MLR) لنمذجة ميزات الروبوتات الخاصة بإنترنت الأشياء مع القدرة على التمييز وتخفيف تحديات اكتشاف إنترنت الأشياء. بالإضافة إلى ذلك ، يتم تطوير نموذج تحليل التمييز الخطي (LDA) للتمييز بين الأنشطة العادية وشبكات روبوت إنترنت الأشياء.

يتم استخدام الاكتشاف المستند إلى الشبكة لمجموعة بيانات إنترنت الأشياء (N–BaIoT) لتقييم أداء النموذج المقترح لاكتشاف الروبوتات بإنترنت الأشياء من حيث الصحة والدقة ومعدل الكشف. تكشف النتائج التجريبية أن نموذج اكتشاف الروبوتات IoT المقترح يوفر مجموعة من الميزات ذات الصلة ويحافظ على دقة عالية عند مقارنتها بأحدث الطرق وطريقة خط الأساس وخاصة LDA.

1. Introduction

In recent years, the Internet of Things (IoT) has grown dramatically. This growth is the result of significant advancements in the semiconductor industry, wireless communication technologies, and the appearance of IoT applications in a variety of fields, such as smart homes, health care, control systems, and military. However, the increased deployment of IoT devices increases security concerns, vulnerabilities, and threats. The IoT has become an effective tool for attackers to conduct cyberattacks, serving as a weak point of entry to penetrate a chain of computer networks and exploiting several vulnerabilities to form IoT botnets [1].

A botnet is a network of compromised machines, such as computers, smartphones, and IoT devices that are controlled by a botmaster. The botmaster has complete control over these machines, which transform into robots and perform a variety of malicious acts such as distributed denial of service (DDoS) attacks, spamming, click fraud, and phishing attacks [2] [3].

A typical bot lifecycle has five stages. The first stage is the creation stage, in which the botmaster scans the target network for vulnerabilities to infect the device. In the second stage, the infected device generates a shellcode, also known as a script. The actual binary bot images are downloaded from specific websites via peer-to-peer (P2P) and hypertext transfer protocol (HTTP) and file transfer protocol (FTP) and installed on the infected machine. After the botnet is installed, the targeted computer is transformed into a zombie. For the third stage, the botnet creates a Command and Control (C&C) to which zombies are connected. The attacker then performs malicious activities on the zombie machine in the fourth stage. Finally, botmasters update and maintain bots for a variety of reasons, such as avoiding detection techniques or adding new operations to the bot army [4].

Numerous methods and technologies can be used to improve the detection of IoT botnets and the security of the IoT system. This paper proposes a regression-based model for determining the characteristics of an IoT botnet, as well as a linear discrimination analysis (LDA) for detecting an IoT botnet.

This paper's contribution is to analyze the IoT botnet and introduce a set of features capable of representing the IoT botnet with discriminative capability. In addition to alleviating the challenges of IoT detection using multiple linear regression (MLR), as well as developing a linear discrimination analysis (LDA) model for detecting IoT effectively. To the best of our knowledge, this is the first time MLR has been used for IoT botnet detection to select the relevant features and LDA for detecting botnet attacks.

The rest of this paper is organized as follows: Section 2 summarizes the related work. In Section 3, an introduction for preliminary concepts related to regression analysis and LDA is presented. Section 4 introduces the proposed methodology. The results and discussion are

clarified in Section 5. Finally, Section 6 presents the conclusion and suggestions for future work.

2. Related work

Different IoT botnet detection approaches have been introduced in the literature. McDermott et al. presented a botnet detection method for IoT devices using a deep bidirectional long short-term memory-based recurrent neural network (BLSTM-RNN) within word embedding. The BLSTM-RNN botnet detection model was evaluated against unidirectional LSTM-RNN. Even though BLSTM-RNN introduced overhead to each epoch and increased processing time, the results proved BLSTM- RNN effectiveness for botnet detection in the IoT [5].

Maidan et al. suggested a model for anomaly identification based on network behaviour snapshots and deep autoencoders. In addition, they built the network-based detection of IoT (N-BaIoT) dataset . The results show that the proposed method is capable of detecting botnet attacks as they originate from compromised IoT devices accurately and instantly. The high positive rate for Philips B120N/10 as compared to other devices is a shortcoming of this model. [6].

Prokofiev et al. presented a logistic regression model for detecting IoT botnets at the propagation stage and determining the likelihood that an IoT device making a connection is running a bot. The results demonstrated that the suggested model is effective for detecting IoT botnets [7].

For detecting cyber-attacks on IoT networks, Anthi et al. proposed a three-layer intrusion detection system (IDS). First, the system classifies each IoT device connected to the network and determines the activity of the normal profile. Once an attack occurs, it identifies malicious packets on the network and categorizes the type of attack that has been launched [1].

Alhajri et al. investigated the feasibility of using auto-encoders for detecting IoT botnets preventing distributed denial of service attacks. The results reveal that the auto-encoder is effective for network security threats detection and improved detecting distinct threats to IoT networks. The desired properties of an auto-encoder and mapping the security criteria for a botnet detection system were not considered in this study [8].

Jung et al. presented a study on the detection of IoT bots by utilizing energy consumption with a convolutional neural network (CNN) model. First, the power consumption data was segmented and normalized. Then, the CNN model was applied to sense the differences in power consumption patterns and classify IoT botnet attacks. The proposed model achieved high accuracy for botnet detection [9].

The models in [7] that used logistic regression for detecting IoT botnet attacks were adopted for comparison with the proposed model. The suggested work in this research employs Linear Discriminant Analysis (LDA) for IoT botnet detection, and it differs from previous works in that it proposes a new feature selection model that is based on Multiple Linear Regression (MLR).

3. Preliminary concepts

3.1 Regression analysis

Regression analysis is a statistical model for estimating the relationship between variables that cause and effect a relationship. There are various types of regression analysis models that differ in terms of the number of independent variables, the type of dependent variables, and the shape of the regression analysis. The main goal of univariate regression is to examine the relationship between a dependent variable and one independent variable. In this paper, Multiple Linear regression is adopted, which is a type of regression model that has one dependent variable and multiple independent variables. Eq. (1) illustrates the formula of MLR [10] [11] [12].

$$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots \omega_n x_n + \epsilon \quad (1)$$

Where
y is dependent variable.
$\forall\, i, 1 \leq i \leq n, x_i$ are independent variables.
ω_i are regression coefficients.
n is number of independent variables.
ϵ is error.

3.2 Linear discriminant analysis

Linear discriminant analysis is a statistical, pattern recognition, and machine learning algorithm that classifies patterns into two or more groups. LDA attempts to find a linear projection that maximizes the ratio of the variance between-class and the variance within-class. It computes the mean and standard deviation for the input example based on the class label. These statistics represent the model that was produced using the training data. Predictions are formed by calculating the probability of a new example corresponds to each class label. After that, the example is assigned to the class with the highest probability [13] [14].

4. Model for IoT Botnet detection

This section introduces the suggested IoT Botnet detection method, which detects one of the most serious concerns facing IoT.

4.1 Problem formulation

Solving IoT botnet detection problem needs two participants: a model for selecting the relevant features, and a model for detecting IoT botnet and evaluating the selected features. The contribution of this paper is to develop a model for these two participants. The proposed model: LDA_{MLR}consists of two stages: the first stage adopts MLR to select the most relevant feature subset from a given feature while keeping classification performance. The second stage uses LDA to detect IoT botnets. Network-based Detection of IoT (N-BaIoT) dataset was used as an evaluation data. The N-BaIoT dataset has 115 features and 7062606 network traffic data that were captured by running the most dangerous IoT malware, such as Mirai and Gafgyt. It comprises real traffic statistics from 9 commercial IoT devices infected with Mirai and Gafgyt. The devices are Danmini, Ecobee, Philips B120N/10, Provision PT-737E, Provision PT-838, SimpleHome XCS7-1002-WHT, SimpleHome XCS7-1003-WHT, Ennio, and Samsung SNH1011N [6].

N BaIoT dataset can be formally described as $\mathbb{B} = \{B_1, B_2, \ldots, B_n\}$, where n is the number of network traffic data. Moreover, each network traffic data, $B_j \in \mathbb{B}$ has 115 features. Suppose $\mathbb{F} = \{\mathrm{F}_1, \mathrm{F}, \ldots, \mathrm{F}_{115}\}$ is the whole feature set, then, the aim of MLR to find a subset of feature, $\mathcal{F}$ $(\mathcal{F} \subseteq \mathbb{F})$ that can differentiate between benign and IoT botnet as well as improve or preserve the performance of the proposed IoT detection, LDA_{MLR} that classifies set $\mathbb{B}$ into two classes $\{c_o, c_1\}$ corresponding to benign and IoT botnet.

4.2 MLR based feature selection

The features of botnet play a crucial role in detection botnet, so it is vital to provide a new model for feature selections by utilizing MLR. This can be accomplished by removing irrelevant features and selecting a small appropriate feature subset that is most effective for distinguishing between classes based on the factors related to the target attribute. MLR is

carried out for selecting a relevant feature set, $\mathcal{F}$, out of the whole set of 115 features, $\mathbb{F}$. The strength of the correlation between variables is employed to weight the features and the features that contain more correlation for separating the classes. In the proposed feature selection model, as clarified in algorithm 1, given N BaIoT training dataset containing n_t network traffic data : $\mathbb{B}_{trn} = \{B'_1, B'_2, \dots, B'_{nt}\}$, and their corresponding labels $L = \{c_o, c_1\}$, the estimated parameters (ω) for each feature in Eq. (1) can be calculated using Eq.(2).

$$\omega = (\mathbb{B}' \times \mathbb{B}'^T)^{-1} \times (\mathbb{B}'^T * L) \quad (2)$$

The resultant ω vector is of size 115×1, which contains the weight for each feature. In other words, $\forall i, 1 \leq i \leq 115$, ω_i is the weight of the feature. After that, $\mathbb{F}$ is sorted in descending order based on the values of the weight. Finally, a percentage from the whole features is selected. Formally speaking:

Let $\mathcal{P}$ is selection percentage, then the feature set $\mathcal{F}$ is selected such that $|\mathcal{F}| = \mathcal{P}|\mathbb{F}|$

4.3 LDA based IoT botnet detection

IoT botnet can be detected according to the signatures of executable malware or signatures of malicious network traffic made by malware. The role of LDA is to detect IoT botnets. Features resulting from MLR based feature selection is used as input to LDA. LDA involves two phases: learning phase and testing phase. The goal of the learning phase is to compute the mean and covariance of benign class and IoT botnet class. The purpose of the testing phase is to classify incoming network traffic data as benign or IoT botnet.

In learning phase, as clarified in Algorithm 2, given $\mathbb{B}_{trn} = \{B'_1, B'_2, \dots, B'_{nt}\}$, and their corresponding labels $L = \{c_o, c_1\}$, the mean μ_I , covariance matrix I of each feature for I_i and pooled covariance matrix C are calculated.

In testing phase, as clarified in Algorithm 3, the mean and covariance matrix values from the learning phase are utilized as input to the testing phase. Then, the discriminant function of each class $c_j \in C, \forall j \in \{0,1\}$ in N-BaIoT testing dataset, $\mathbb{B}_{tst} = \{T_1, T_2, \dots, T_{ns}\}$.

Finally, a label is assigned for the network traffic T_i. The network traffic T_i is classified as a benign when the discriminant function of c_0 is greater than the discriminant function of c_2. Otherwise, T_i is classified as an IoT botnet. In other words, T_i belongs to class with highest discriminant function.

Algorithm 1: Feature selection by MLR	
Input:	
• $\mathbb{B}_{trn} = \{B'_1, B'_2, \dots, B'_{nt}\}$: N-BaIoT training dataset • $L = \{c_o, c_1\}$ *:* Label of training samples • $\mathcal{P}$: Selected features percentage	
Output:	
• $\mathcal{F} \subseteq \mathbb{F}$: set of selected features	
1.	Calculate the weights of all features $\omega = (\mathbb{B}' \times \mathbb{B}'^T)^{-1} \times (\mathbb{B}'^T * L)$
2.	Sort the weights of the features in descending order Sort F_i , $\forall i, 1 \leq i \leq 115$ according to ω_i
3.	Select feature set $\mathcal{F}$ according to the specified percentage, $\mathcal{P}$ $\|\mathcal{F}\| = \mathcal{P}\|\mathbb{F}\|$

Algorithm 2: Learning phase of LDA

Input:

- $\mathbb{B}_{trn} = \{B'_1, B'_2, \dots, B'_{nt}\}$:N-BaIoT training dataset
- $L = \{c_o, c_1\}$ *:* Label of training samples
- nt: Number of network traffic data in $\mathbb{B}_{trn}$

Output:

- μ : mean of each feature for each class
- C : pooled covariance matrix
- p: Prior probability

1.	Calculate the mean of each feature for each class $$\mu_i = \frac{1}{n_i}\sum_{B'_j \in c_i} B'_j, \forall i, 0 \le i \le 1$$ Where n_i represents the number of network traffic data in class c_i
2.	Calculate the total mean of the entire training dataset, $\mathbb{B}_{trn}$. $$\mu = \frac{1}{n_t}\sum_{i=1}^{n_t} B'_i$$
3.	Calculate the mean corrected data for each class $\tilde{\mu}_i = B'_j - \mu \ , \forall i, 0 \le i \le 1, \ B'_j \in c_i$
4.	Compute the covariance matrix for each class $cov_i = \frac{\tilde{\mu}_i{}^T \tilde{\mu}_i}{n_i}$, $\forall i, 0 \le i \le 1$
5.	Calculate of pooled covariance matrix $C = \frac{1}{n_t}\sum_{i=1}^{2} n_i * cov_i;$
6.	Calculate the prior probability for each class $p_i = \frac{\sum_{j=1}^{nt} c_i = c_j}{nt}, \ \forall i \in \{0,1\}$

Algorithm 3: IoT botnet detection by LDA

Input:

- $\mathbb{B}_{tst} = \{T_1, T_2, \dots, T_{ns}\}$: N-BaIoT testing dataset
- n_s: Number of network traffics in $\mathbb{B}_{tst}$
- μ : mean of each feature for each class
- C : pooled covariance
- p**:** Prior probability

Output:

- Classified $\mathbb{B}_{tst}$

1.	Compute the inverse of the pooled covariance matrix $C^{-1} = inverse(C);$
2.	Compute the discriminant function $\forall i, 0 \le i \le 1$ $$f_i = (\mu_i \times C^{-1} \times \mathbb{B}_{tst}{}^T) - \frac{1}{2}(\mu_i \times C^{-1} \times \mu_i + log(p_i)$$

3.	Assign label benign or IoT to the network traffic T_i $\forall i, 0 \leq i \leq ns$ $T_i = \begin{cases} \text{benign } if\ f_0 > f_1 \\ \text{IoT bonet}, otherwise \end{cases}$

5. Experiments and results

In this section, the performance of the proposed LDA_{MLR} model is tested against baseline feature selection model, namely information gain (IG) and [7] that used logistic regression model (LR) for classification. Moreover, the accuracy (Acc),precision(P), and detection rate (DR) are used as evaluation measurements [15].

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$P = \frac{TP}{TP+FP} \quad (4)$$

$$DR = \frac{TP}{TP+FN} \quad (5)$$

Where
TP is true positive,
TN true negative,
FP is false positive,
and
FN is false negative.

5.1 N- BaIoT dataset preparation

As mentioned previously, N-BaIoT dataset was used in this paper. First, duplicate instances from N- BaIoT dataset are removed. After that, the N- BaIoT dataset must be split into two parts: training and testing. A 3-fold cross-validation method was adopted to define a random partition on N-BaIoT dataset and validate the proposed LDA_{MLR} model. The size of the training and test datasets is depicted in Figure 1

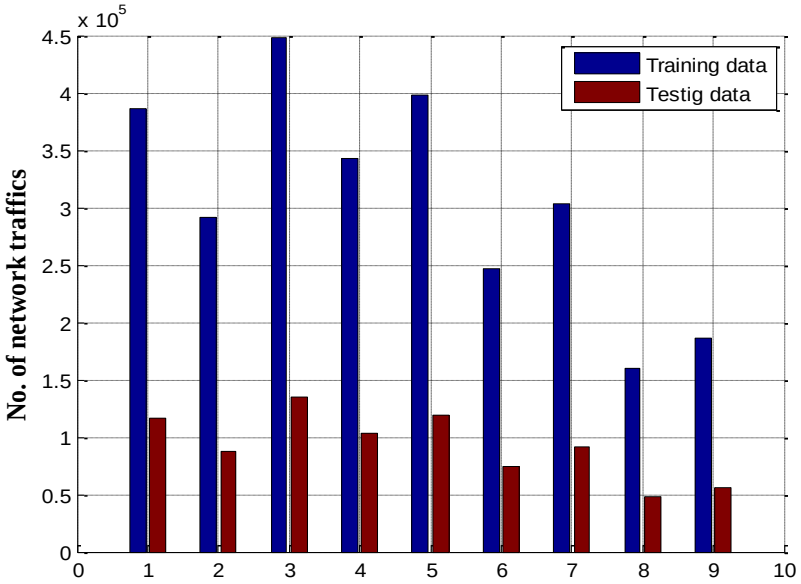

Figure 1: Number of network traffics in training dataset and testing dataset for each device

5.2 Evaluation of the proposed LDA_{MLR} model performance

To demonstrate the value of selected features on the performance of the proposed model, three percentages: 25%, 50%, and 75% of the highest overall features in the N-BaIoT dataset were chosen. Tables 2- 10 report the average accuracy (Acc'), average precision (P'), and average detection rate (DR') over 3 folds.

Tables 1-9 demonstrate the results of the proposed LDA_{MLR} model against the well-known features selection information gain (IG) and [7].

Table 1: Acc', P' *and* DR' of Danmini device

Model	Selected features%	Acc'	P'	DR'
LDA	**100**	99.82	99.94	99.84
LDA_{MLR}	**75**	**99.80**	**99.94**	**99.82**
	50	99.68	99.91	99.71
	25	86.20	99.86	83.37
LDA_{IG}	**75**	**99.80**	**99.94**	**99.83**
	50	99.62	99.94	99.60
	25	99.45	99.95	99.39
[7]	**100**	37.01	76.23	30.52
	75	63.30	78.83	75.92
	50	82.56	82.56	100
	25	82.56	82.56	100

Table 2: Acc', P' *and* DR' of Ecobee device

Model	Selected features%	Acc'	P'	DR'
LDA	**100**	59.31	94.78	58.99
LDA_{MLR}	**75**	**93.52**	**99.96**	**96.43**
	50	85.62	83.29	84.45
	25	80.37	96.73	80.47
LDA_{IG}	**75**	61.13	93.36	60.69
	50	35.82	61.53	33.35
	25	54.08	64.41	51.73
[7]	**100**	74.31	91.76	78.96
	75	89.23	92.30	96.38
	50	89.23	92.30	96.38
	25	92.51	92.51	100

Table 3: Acc', P' and DR' of Philips B120N//10 device

Model	Selected features%	Acc'	P'	DR'
LDA	**100**	99.70	99.93	99.55
LDA_{MLR}	**75**	**99.68**	**99.94**	**99.51**
	50	99.66	99.95	99.46
	25	99.00	99.97	98.31
LDA_{IG}	**75**	99.66	99.93	99.50
	50	**99.68**	**99.94**	**99.52**
	25	99.55	99.91	99.31
[7]	**100**	63.89	86.42	49.04
	75	44.76	51.68	71.91
	50	58.33	58.33	100
	25	58.33	58.33	100

Table 4: Acc', P' and DR' of Provision PT-737E device

Model	Selected features%	Acc'	P'	DR'
LDA	**100**	68.67	99.11	57.54
LDA_{MLR}	**75**	96.19	97.92	96.88
	50	88.73	99.91	84.69
	25	**96.41**	**99.93**	**95.16**
LDA_{IG}	**75**	59.57	84.27	66.66
	50	57.80	84.30	54.55
	25	91.67	99.85	88.78
[7]	**100**	44.56	91.23	31.81
	75	67.22	71.71	90.99
	50	73.18	73.18	100
	25	73.18	73.18	100

Table 5: Acc', P' and DR' of Provision PT-838 device

Model	Selected features%	Acc'	P'	DR'
LDA	**100**	46.57	66.24	53.85
LDA_{MLR}	**75**	60.92	77.38	51.29
	50	**98.08**	**99.84**	**97.00**
	25	97.18	99.88	95.46
LDA_{IG}	**75**	65.84	78.75	66.27
	50	59.62	75.85	57.25
	25	53.06	78.53	45.36
[7]	**100**	39.29	45.71	50.35
	75	60.71	60.71	100
	50	60.71	60.71	100
	25	60.71	60.71	100

Table 6: Acc', P' and DR' of SimpleHome XCS7-1002-WHT device

Model	Selected features %	Acc'	P'	DR'
LDA	**100**	99.72	99.82	99.80
LDA_{MLR}	**75**	**99.69**	**99.82**	**99.74**
	50	99.47	99.84	99.42
	25	98.88	99.82	98.61
LDA_{IG}	**75**	99.48	99.81	99.46
	50	99.47	99.82	99.44
	25	98.94	99.48	99.02
[7]	**100**	49.54	69.75	42.82
	75	65.58	69.69	91.16
	50	71.14	71.14	100
	25	71.14	71.14	100

Table 7: Acc', P' and $DR'$of SimpleHome XCS7-1003-WHT device

Model	Selected feature %s	Acc'	P'	DR'
LDA	**100**	52.66	90.71	52.64
LDA_{MLR}	**75**	85.15	97.51	85.10
	50	**99.78**	**99.92**	**99.83**
	25	99.67	99.96	99.67
LDA_{IG}	**75**	50.15	81.24	54.13
	50	79.43	90.95	86.03
	25	98.25	99.95	98.11
[7]	**100**	26.73	57.44	22.37
	75	83.00	89.35	91.96
	50	90.13	90.13	100
	25	90.13	90.13	100

Table 8: Acc', P' and DR' of Ennio device

Model	Selected features %	Acc'	P'	DR'
LDA	**100**	57.83	81.39	41.93
LDA_{MLR}	**75**	**99.71**	**99.74**	**99.80**
	50	99.70	99.69	99.83
	25	99.61	99.71	99.67
LDA_{IG}	**75**	53.49	51.92	45.79
	50	66.51	54.30	66.49
	25	63.14	42.09	100
[7]	**100**	31.59	12.09	11.04
	75	63.14	63.14	100
	50	63.14	63.14	100
	25	63.14	63.14	100

Table 9: Acc', P' and $DR'$of Samsung SNH 1011 N device

Model	Selected features %	Acc'	P'	DR'
LDA	**100**	77.66	86.31	76.95
LDA_{MLR}	**75**	86.81	99.89	76.37
	50	**99.88**	**99.84**	**99.94**
	25	99.20	99.85	98.71
LDA_{IG}	**75**	53.25	51.85	42.51
	50	54.00	33.70	33.39
	25	69.74	50.71	66.44
[7]	**100**	51.85	37.04	66.67
	75	55.56	55.56	100
	50	55.56	55.56	100
	25	55.56	55.56	100

The Danmini device results in Table 1 demonstrate that the classification using LDA alone without feature selection has an accuracy result of 99.82%, which is very high, so it is difficult to improve but aspire to keep it as possible. The accuracy results of LDA_{MLR} with 75%, 50%, and 25% are 99.80%, 99.68%, and 86.20%, respectively that surpasses LDA_{IG} and [7, 10]. The results of the Ecobee in Table 2 demonstrate that LDA_{MLR} results with percentages of 75%, 50%, and 25% are better than LDA, LDA_{IG} , and [7].

The results of the Philips B120N/10, as given in Table 3, reveal that the classification using LDA, the accuracy result is 99.70%, which is also high result and difficult to improve but needed to be maintained as much as possible. Accuracy results of LDA_{MLR} with percentages of 75%, 50%, and 25% are 99.68%, 99.66%, and 99.00%, respectively. LDA_{IG} with the same proportions of features, the accuracy are 99.66%, 99.68%, and 99.55%, respectively. The results of [7] are less than the previous models and errors happen frequently within a normal class.

The results of the Provision PT-737E, as shown in Table 4, is 68.67% when using LDA with all features (115). While LDA_{MLR} accuracy results at percentages of 75%, 50% and 25% are 96.19%, 88.73%, and 96.41%, respectively. The LDA_{MLR} accuracy results are high in all percentages, especially at 25%. For LDA_{IG} with the same previous percentages, the accuracy results are 59.57%, 57.80%, and 91.67%, respectively. The results of LDA, LDA_{IG}, and [7] are significantly lower than the proposed LDA_{MLR}. Furthermore, LDA_{IG} and [7] are unable to effectively classify normal activity.

The results of the Provision PT-838, as shown in Table 5, shows that classification with LDA is 46.57%, which is a poor result. However, LDA_{MLR} accuracy results are 60.92%, 98.08%, and 97.18% for 75%, 50%, and 25% features percentages respectively. Results clearly show the superiority of LDA_{MLR} over LDA_{IG} and [7] that cannot accurately classify normal activities. The SimpleHome XCS7- 1002-WHT device results, as reported in Table 6, reveal that LDA_{MLR} outperforms LDA, LDA_{IG} and [7]. The best accuracy result of LDA_{MLR} is 99.69% at 75% feature percentage (i.e., the number of features is 86).

The LDA_{MLR} accuracy results for the SimpleHome XCS7-1003-WHT, as quantified in Table 7, are 85.15%, 99.78%, and 99.67% for corresponding 75%, 50%, and 25% feature percentages respectively, which are better than LDA, LDA_{IG} and [7]. The best accuracy result of LDA_{MLR} is 99.78% at 50% feature percentage (i.e., the number of features is 58). Table 8 shows the accuracy results of LDA_{MLR} on Ennio device. We can see that in almost all feature selection percentages (i.e., 75%, 50%, and 25%), the proposed LDA_{MLR} model beats LDA, LDA_{IG} and [7].

Table 9 summarizes the performance of the LDA_{MLR} model against LDA, LDA_{IG} and [7] when applied to Samsung SNH 1011 N device. In three settings of features selection (i.e., 75%, 50%, and 25%), the observation is that the performance of LDA_{MLR} succeeds in correctly classifying IoT botnet and normal activities. While LDA_{IG} and [7] accuracy results are significantly lower, and the misclassification happens frequently in normal activities classification.

In summary, the results obtained by LDA_{MLR} have high accuracy, precision, and detection rate than those obtained by LDA_{IG} and [7]. The results clearly show the positive impact gained by collaborating the proposed MLR feature selection model with LDA for the three distinct percentages, indicating that it is more efficient than IG when combined with LDA. LDA_{MLR} provides high accuracy, precision, and detection rate with a fewer number of features due to LDA_{MLR} basing its calculations on the strength of the association between the features. Whereas, the information gain model is based on the concept of entropy, as it seeks to reduce the entropy by calculating the information gain value for each feature. Furthermore, the proposed MLR feature selection model could improve the performance of [7].

6. Conclusions

The internet's rapid growth over the last decade has certainly contributed to an increase in cyber-attack occurrences. The vast majority of IoT devices are more vulnerable than standard desktop computers. As a result, they constitute a security risk to information systems. Thus, detecting an IoT botnet is one way to handle this issue.

In this paper, a model for detecting botnets on IoT devices has been developed by using a collaboration between the MLR and LDA. The MLR has been designed for feature selection to cope with relevant features and discard irrelevant features. The comparison was performed with baseline model and models in [7]. Experiments on N-BaIoT show that the suggested LDA_{MLR} model is effective in defining the IoT botnet detection problem. Furthermore, the results demonstrate the superiority of LDA_{MLR} over the baseline model and models in [7].
The proposed LDA_{MLR} model accuracy result of Danmini device with 75%, 50%, and 25% feature percentages achieved 99.80%, 99.68%, and 86.20%, respectively. The accuracy results of LDA_{MLR}model of the Ecobee device with percentages of 75%, 50%, and 25% reached 93.52%, 85.62% and 80.37% respectively.

Regarding Philips B120N/10, the accuracy results of LDA_{MLR} with percentages of 75%, 50%, and 25% are 99.68%, 99.66%, and 99.00%, respectively. The results of the Provision PT-737E accuracy results at percentages of 75%, 50% and 25% were 96.19%, 88.73%, and 96.41%, respectively. The results of the Provision PT-838, accuracy results were 60.92%, 98.08%, and 97.18% for 75%, 50%, and 25% features percentages respectively.

The LDA_{MLR} accuracy results for the SimpleHome XCS7-1002-WHT are **99.69**%, 99.47%, and 98.88% for corresponding 75%, 50%, and 25% feature percentages respectively. The LDA_{MLR} accuracy results for the SimpleHome XCS7-1003-WHT were 85.15%, 99.78%, and 99.67% for corresponding 75%, 50%, and 25% feature percentages respectively. The LDA_{MLR} accuracy results for the SimpleHome XCS7-1003-WHT were 85.15%, 99.78%, and 99.67% for corresponding 75%, 50%, and 25% feature percentages respectively. The accuracy results o LDA_{MLR} model when applied on Samsung SNH 1011 N device are 86.81, 99.88, and 99.20. The works presented in [7] used LR for classification without concentrate on distinguished features. However, the proposed model has adopted a MLR for selecting the distinguished features. From the results, it is clear that the proposed LDA_{MLR} model has identified the distinguished features in different features percentages. In addition, the proposed model can detect IoT botnet efficiently with a high accuracy, detection rate and precision as compared against [7].

In the future, a research into determining IoT botnets can be conducted using deep neural networks to attain more accurate botnet detection results.

References

[1] E. Anthi, L. Williams, M. Słowińska, G. Theodorakopoulos and P. Burnap, "A Supervised Intrusion Detection System for Smart Home IoT Devices", in *IEEE Internet of Things Journal*, Vol. 6, No. 5, pp. 9042-9053, Oct. 2019, doi: 10.1109/JIOT.2019.2926365.

[2] W. S. Hamza, H. M. Ibrahim, M. A. Shyaa, and J. J. Stephan, "IoT Botnet Detection: Challenges and Issues", *Test Engineering and Managemen*t, No. 15092, pp. 15092–15097, 2020

[3] K. Shinan, K. Alsubhi, A. Alzahrani, and M. U. Ashraf, "Machine learning-based botnet detection in software-defined network: A systematic review", *Symmetry.*, Vol. 13, No. 5, pp. 1–28, 2021, doi: 10.3390/sym13050866.

[4] S. Gaonkar, N. F. Dessai, J. Costa, A. Borkar, S. Aswale and P. Shetgaonkar, "A Survey on Botnet Detection Techniques", *Proc. of International Conf. on Emerging Trends in Information Technology and Engineering (ic-ETITE),* 2020, pp. 1-6, doi: 10.1109/ic-ETITE 47903.2020. Id-70.

[5] C. D. McDermott, F. Majdani and A. V. Petrovski, "Botnet Detection in the Internet of Things Using Deep Learning Approaches", *Proc. of International Conf. on Neural Networks (IJCNN),*

2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489489

[6] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, ad Y. Elovici, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," in *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12-22, Jul.-Sep. 2018, doi: 10.1109/MPRV.2018.03367731.

[7] A. O. Prokofiev, Y. S. Smirnova and V. A. Surov, "A method to detect Internet of Things botnets", *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2018, pp. 105-108, doi: 10.1109/EIConRus.2018.8317041.

[8] R. Alhajri, R. Zagrouba, and F. Al-Haidari, "Survey for Anomaly Detection of IoT Botnets Using Machine Learning Auto-Encoders", *International Journal of Applied Engineering Research.*, vol. 14, no. 10, pp. 2417–2421, 2019.

[9] W. Jung, H. Zhao, M. Sun, and G. Zhou, "IoT botnet detection via power consumption modeling", *Smart Health*, vol. 15, no. December 2019, p. 100103, 2020, doi: 10.1016/j.smhl.2019.100103.

[10] G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis", *Procedia - Social and Behavioural Science*s, Vol. 106, pp. 234–240, 2013, doi: 10.1016/j.sbspro.2013.12.027.

[11] W. Vogt and R. Johnson, "Correlation and Regression Analysis", SAGE Publications Ltd 2012, doi: 10.4135/9781446286104.

[12] P. Roback and J. Legler, " Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R", Chapman and Hall/CRC, 2020

[13] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial", *AI* Communications, Vol. 30, No. 2, pp. 169–190, 2017, doi: 10.3233/AIC-170729.

[14] A. Csikosova, M. Janoskova, and K. Culkova, "Application of Discriminant Analysis for Avoiding the Risk of Quarry Operation Failure," *Journal of Risk Financial Manag*. Vol. 13, No. 231, 2020.

[15] T. B. Alhijaj, S. M. Hameed, and B. A. Attea, "A decision tree-aware genetic algorithm for botnet detection," *Iraqi Journal of Science.*, Vol. 62, No. 7, pp. 2454–2462, 2021, doi: 10.24996/ijs.2021.62.7.34.