

Selection of efficient machine learning algorithm on Bot-IoT dataset for intrusion detection in internet of things networks

Imane Kerrakchou, Adil Abou El Hassan, Sara Chadli, Mohamed Emharraf, Mohammed Saber

Laboratory of Smart Information, Communication and Technologies, National School of Applied Sciences, Mohammed First University
Oujda, Morocco

Article Info

Article history:

Received Aug 10, 2023

Revised Jun 14, 2023

Accepted Jun 17, 2023

Keywords:

Artificial intelligence

Bot-IoT dataset

Internet of things

Intrusion detection system

Machine learning

Supervised learning

ABSTRACT

With the growth of internet of things (IoT) systems, they have become the target of malicious third parties. In order to counter this issue, realistic investigation and protection countermeasures must be evolved. These countermeasures comprise network forensics and network intrusion detection systems. To this end, a well-organized and representative data set is a crucial element in training and validating the system's credibility. In spite of the existence of multiple networks, there is usually little information provided about the botnet scenarios used. This article provides the Bot-IoT dataset that embeds traces of both legitimate and simulated IoT networks as well as several types of the attacks. It provides also a realistic test environment to address the drawbacks of existing datasets, namely capturing complete network information, precise labeling, and a variety of recent and complex attacks. Finally, this work evaluates the confidence of the Bot-IoT dataset by utilizing a variety of machine learning and statistical methods. This work will provide a foundation to enable botnet identification on IoT-specific networks.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Imane Kerrakchou

Laboratory of Smart Information, Communication and Technologies, National School of Applied Sciences

Mohammed First University

BP 669, Oujda, 60000, Morocco

Email: i.kerrakchou@ump.ac.ma

1. INTRODUCTION

Fast internet development is driving the growth of the internet of things (IoT), the most notable examples of which are smart cities and homes, cyber-physical and healthcare systems. The IoT is a set of daily devices interconnected with lightweight cpu and network cards that can be controlled via the web interfaces, applications, or some other kinds of interfaces [1]. Understandably, any widely adopted new technology by the public will attract the attention of cyber-attackers who exploit technology vulnerabilities, utilizing a variety of complex techniques for hacking like botnets attacks. This issue has been exacerbated due to the limited standardization of the IoT systems along with the cheap, low-power, and light-weight devices comprising many of those systems. The spread of botnet malware that was demonstrated to launch distributed denial-of-service (DDoS) attacks of up to 1.1 TBps has been one of the ways that IoT networks have been explored for criminal purposes [2]. These botnets have resulted in a rise in network attacks, as evidenced by a recent report from Forbes indicating a three-fold increase in the number of events, reaching over 2.9 billion in 2019. Additionally, SonicWall reported a significant surge in IoT malware attacks, with a growth of 215.7% from 10.3 million in 2017 to 32.7 million in 2018 [3].

Given novel and unique technologies that can compromise a network of IoT and insufficient existing techniques to deal with them, developing an advanced digital forensic approach to identify and investigate

adversary behavior is essential. Network forensics techniques have been widely utilized in analysis of network traces and identification of the infected devices that participate in major cyber attacks [4]. Considering the nature and number of internet of things devices, the burden of treatment of the data collected could be a perfect application for analyzing big data. The last one being a complex set of processes conceived for three major issues: volume, variety, and velocity [5]. With the immense volumes of data generated by IoT networks, employing analytical techniques capable of processing the huge volumes of data in near real time is essential. In addition, forensics requires large data sources to validate their credentials in the IoT networks.

For the further evolution of forensic and intrusion detection solutions which recognize and examine cyber attacks, the development of a realistic dataset remains an important research subject [6]. Over time, several datasets have been developed, each with its advantages and disadvantages. The existing datasets, despite being applied to certain scenarios, have various problems, e.g., lack of reliably tagged data, low diversity of attacks like botnet scenarios, redundancy of traces, and lack of ground truth [7]. Nevertheless, a feasible botnet trace dataset in IoT networks was not designed effectively. The Bot-IoT dataset solves the challenges noted previously [8].

In this paper, the performance of network forensics methods based on the algorithms of machine learning will be evaluated using the Bot-IoT dataset. The use of machine learning enables a lighter and less complex implementation compared to deep learning. Machine learning can provide satisfactory results with simpler models and a moderate amount of training data. This allows for faster detection and more efficient use of resources, which is crucial in combating botnet attacks that can have devastating effects on networks and computer systems. The remainder of this document is divided into the following: Section 2 gives an introduction of IoT and botnets. Section 3 defines Bot-IoT dataset. A complete overview of the different machine learning algorithms is presented in section 4. Section 5 lays out the proposed method to analyse our dataset. Finally, in section 6, the results of the experiments are presented and then discussed in section 7, accompanied with the conclusion of this document.

2. BACKGROUND AND RELATED WORK

2.1. IoT security

Through the years, internet of things has developed in terms of complexity and functionality, becoming a mature and integrated part of the society, covering many areas of application. The concept of IoT has been around for some time, occasionally under different names such as "web of things", and "internet of things", "embedded intelligence". In a presentation in 1999, Kevin Ashton introduced the term "internet of things" which described the interest in the use of automated and contextual computers to collect and use data [9]. Therefore, we propose the full definition of the IoT as: "The internet of things is a system of large and small devices called 'objects', which have been endowed with limited communication capabilities and processing power, and which provide services, to include the platforms, software, and the infrastructure, to an organization or user remotely, on-demand, and at a cost lower than buying physical systems."

Promising the innovation, automation, and optimization of both industrial and commercial systems, the global growth of IoT should not be surprising, as many studies and predictions are made about it. One survey predicts the number of devices that will be deployed as part of the IoT will skyrocket to more than 20.4 billion in future. The IoT's economic growth predictions are understood more clearly if we consider that IoT isn't just for one part of the global market but is gradually becoming an integral part of most areas of current society, such as the healthcare field [10], industry, and critical infrastructure, agriculture [11].

In the IoT, there is no widely accepted single framework or set of standards. On the contrary, providers are freed to build their systems using their preferred technologies, giving rise to a heterogeneous environment for the IoT. Due to its design, the IoT is vast and encompasses many technologies, that must cohere harmoniously for the entire system to function. Due to the open communication loop of the IoT and the heterogeneity of their services and protocols, those technologies are vulnerable to cyber-attacks. As explained in the next section, we primarily focus on botnets, as they pose significant harm to IoT devices and applications [12].

2.2. Botnets

Botnets come with a rich development and history throughout the years, causing corruption and disruption to computer systems and networks. Initially, botnets were designed for benign purposes. Their primary function was to offer administrative assistance to internet relay chats (IRC), a type of communication that was very popular in the 1990s. In 1993, the primary IRC bot appeared. It is called Eggdrop and it assisted with communication via IRC channels. After Eggdrop, the first malicious bots appeared in 1998 named global threat bot (GTBot), the first of its kind, capable of executing scripts when invited by its IRC command and control (C&C) channel. Over time, a number of sophisticated bots have emerged, such as ZBot and Gameover Zeus, apt to perform a wide variety of malicious activities, such as spam and DDoS attacks that disrupt the flow of internet communication, and bank account theft [13]. Botnets composed of IoT devices have been the

next evolutionary stage of botnets. The most famous one, Mirai, emerged during September 2016 and carried out a few of the more violent DDoS attacks in the history of the Internet, including 620 Gbps against Brian Krebs's website, and 1.1 Tbps against the French provider of cloud services OVH.

Botnet architectures consist of various components. First, a bot refers to a program that infects a vulnerable host and incorporates it into the botnet [14]. What sets bots apart from other malicious software is their ability to establish a communication channel with their creators. This channel allows the bot to receive commands from its creators and enables botnets to be flexible in terms of functionality. To spread the malware within the botnet, a propagation mechanism is used to target vulnerable destinations. However, the defining characteristic of botnets lies in the capability of their controller, known as a botmaster, to provide instructions and receive feedback from the network of infected devices. This control is made possible through a (C&C) infrastructure. Figure 1 shows the basic elements of a botnet.

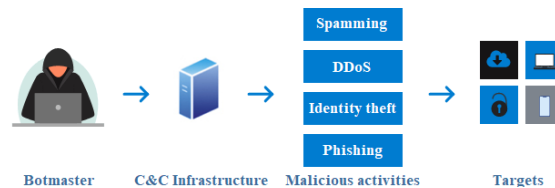


Figure 1. Botnet command and control architecture

One of the more pervasive elements of the code that traverses the internet are botnets. The primary reason for such attention is not the ability of botmasters that hide their bots to law enforcement, but the botnets' convenient capacities and the services provided by them to the botmasters and their customers. The botnets use a variety of hacking techniques, including keylogging, phishing, spamming, distributed denial of service attacks, identity theft, and possibly the propagation of malware [15].

2.3. Literature review

This section is aimed at addressing the objectives of the review and summarization of the research work on the Bot-IoT dataset. In this work, frameworks have been presented for both the identification and security of networks against adversarial attacks. The work of Djenna concentrated on the efficient detection of DDoS attacks on the IoT networks. Authors proposed a minimal human intervention system that could perform its function.

The system is composed of several stages taking the raw traffic, processing it, thereafter using an artificial neural network (ANN) to identify its classification. This model achieved an F-measure of 99.05% when classifying DDoS traffic [16]. A fog computing-based framework for detecting attacks in the cloud was developed by Lawal *et al.* [17]. Their framework supports the computational load of IoT devices by using resources in the fog of edge computing services. In terms of their framework, the eXtreme gradient boosting algorithm (XGBoost) has been tested for binary classification. The result was high, with an F-measure of 99.5%.

A framework for intrusion detection was created by Ge *et al.* that uses deep learning models, such as the feedforward neural network (FNN) model, to categorize attacks based on information at the packet level. These authors found that the performance of the FNN model ranged from 99 to 100% accuracy for binary classification of every type of attack [18]. In their IDS, Jithu *et al.* [19] used a deep neural network (DNN) as an initial approach of classification. The model DNN was typical, utilizing several hidden node layers and a rectified linear unit (ReLU) activation function. According to results of their tests, an average F-measure of 94% was achieved by their model for binary classification.

Filus *et al.* [20] have aimed to evolve a low computational costs intrusion detection system by utilizing a randomized neural network (RNN) model. Utilizing the smaller subset of the Bot-IoT dataset with just the "top 10 features", a number of versions of the RNN model were trained and tested by the authors, using different numbers of nodes and initialization of weights. Several tests showed that the RNN model with 12-node performed well with 96.09% accuracy.

3. BOT-IOT DATASET

In 2018, the Bot-IoT dataset [21] has been developed and released by University of New South Wales (UNSW) Canberra in 2019 like a realistic and modernized dataset which is mainly centered on the IoT infrastructure. Figure 2 illustrates the Bot-IoT testbed, which consists of four Kali virtual machines belonging to the attacking machines, along with an Ubuntu server, Windows 7, Ubuntu mobile, Metasploitable, and an

Ubuntu tap machine. Their internet connection is established through the PFSense machine, which functions as a packet filtering device. The traffic is then passed through a switch and a second firewall before being routed to the internet. The Bot-IoT dataset bundles authentic and fabricated web traffic with three types of cyber attacks: Information theft, denial of service, and probing attacks.

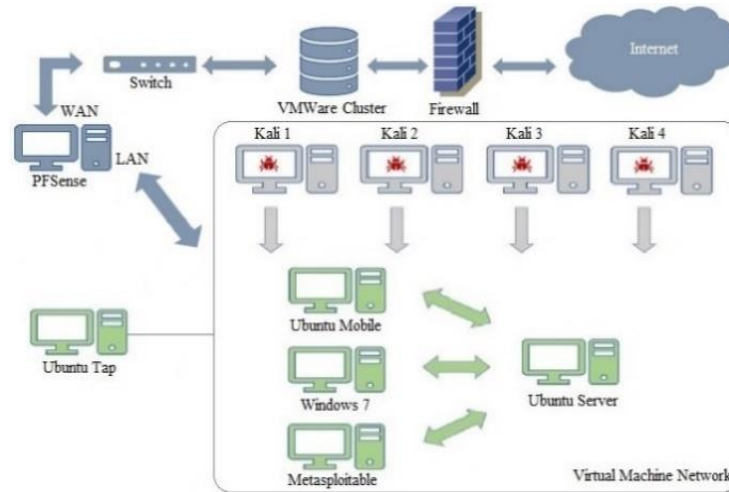


Figure 2. The Bot-IoT dataset's testbed environment

The dataset consists of approximately 70 GB packet CAPture files (PCAP). Those files include the collected network data from network taps in the test-bed environment. The Bot-IoT authors utilized an Argus tool, which is a tool for network data auditing and security, to analyze the traffic on the network and generate their processed data set from the PCAP file. Then, Argus PCAP files were fed and the resultant data were inserted into their structured query language (MySQL) database using a client Argus program. Following this, the information was extracted as a comma-separated value (CSV) file from the MySQL database without any further modification to the complete set [22]. Bot-IoT is comprised of two distinct subsets: a 5% subset, which is used in this work, and a 10-best subset. The 5% subset was provided as a shorter and easily managed version of the dataset. As indicated by its name, this includes 5% of the original set's instances, or about 3.6 million, and looks to be a sample representative of the entire set in regard to the category of attack. In the 5% subset, 43 independent features and 3 dependent features are represented. The 43 independent features include the features of the Argus network traffic and the supplemental features computed. The 5% subset is split in four CSV files, each with a header row of the names of the features.

4. INCORPORATE MACHINE LEARNING CLASSIFIERS INTO BOT-IOT

In this section, we will present a concise overview of the various models studied using the Bot-IoT dataset:

- Random forest classifier: It is an algorithm for supervised learning. It is utilized for regression as well as for classification. It's also the easiest and most flexible algorithm to use. The forest is composed of many trees. It is stated that the more trees there are, the more powerful a forest is. The random forests generate decision trees on random samples of data, get a prediction from these trees, as well as choose the best solution through a vote. It also gives a fairly accurate indicator of the importance of the feature [23].
- Gradient boosting classifier: It is a technique that tackles the problem of minimizing the model loss function through numeric optimization. It achieves this by employing weak learners and utilizing gradient descent. This iterative algorithm finds local minimums of a distinct function. The technique is versatile as it can be applied to regression, multiclass classification, and more, thanks to its ability to accommodate different loss functions. Importantly, gradient boosting does not alter the sample distributions, but instead trains weak learners on the residuals of the strong learner's remaining error, known as pseudo-residuals [24]. This approach allows for the weighting of misclassified observations. Additionally, new strong learners can be added to address areas where existing learners are underperforming. The contributions of each weak learner are determined through gradient optimization, aiming to minimize the overall error of the strongest learner.

- Decision tree classifier: It is regarded as an easily understandable and highly efficient algorithm. It requires minimal effort for training and is capable of effectively classifying non-linear data that can be separated. Compared to other classification algorithms, it exhibits fast performance and high efficiency. When it comes to attribute selection in this type of classifier, the most commonly used measures are entropy and information gain. Entropy quantifies the level of uncertainty or randomness in an element, serving as an indicator of impurity. Information gain, on the other hand, measures the relative change in entropy when considering an independent attribute, aiming to estimate the informational value of each attribute. In constructing the decision tree, the algorithm searches for the attribute that yields the greatest information gain [25].
- Artificial neural network classifier: It is a computational model made up of units called neurons, organized in layers. These neurons take an input vector, apply a non-linear function to it, and transmit the output to the next layer. The network is usually feedforward, i.e. there are no connections back to the preceding layer. The connections between neurons have weights, which are adjusted during the learning phase to solve a specific problem. The network consists of three layers: input, hidden, and output. The input layer contains input values, the hidden layer contains neurons, and the output layer has nodes representing each class. The network assigns a value to each output node and assigns the record to the class node with the highest value [26].
- Naive bayes classifier: It is a simple probabilistic machine learning algorithm that is commonly used for classification tasks. It is based on the Bayes' theorem and assumes that the features used to classify the data are independent of each other, hence the term "naive". The classifier calculates the probability of a particular instance belonging to a certain class by combining the prior probability of the class and the conditional probabilities of the features given the class. It assumes that each feature contributes independently to the probability, which is where the "naive" assumption comes into play. This classifier is known for its simplicity, efficiency, and ability to handle high-dimensional feature spaces [27].
- Logistic Regression Classifier: It aims to generate the best model to determine the dependence or relationship between the class variable and the features [28]. In the case of the test with just two classes: 0 and 1, it essentially predicts a value between 0 and 1 as the probability that the class is 1 for observation. Simple model Logistic Regression is suitable only for binary classification, but with some effort, it could be extended to a multiclass objective [29].

5. METHOD PROPOSED

Our suggested system contains five fundamental method steps: feature extraction, data preprocessing, feature selection, data splitting, and the implementation of Machine Learning (ML) algorithms. These steps, depicted in Figure 3, play a crucial role in the system by enabling effective data analysis and model training. By following these method steps, the suggested system facilitates the transformation of raw data into meaningful features, ensuring optimal data preparation and model performance for successful machine learning outcomes.

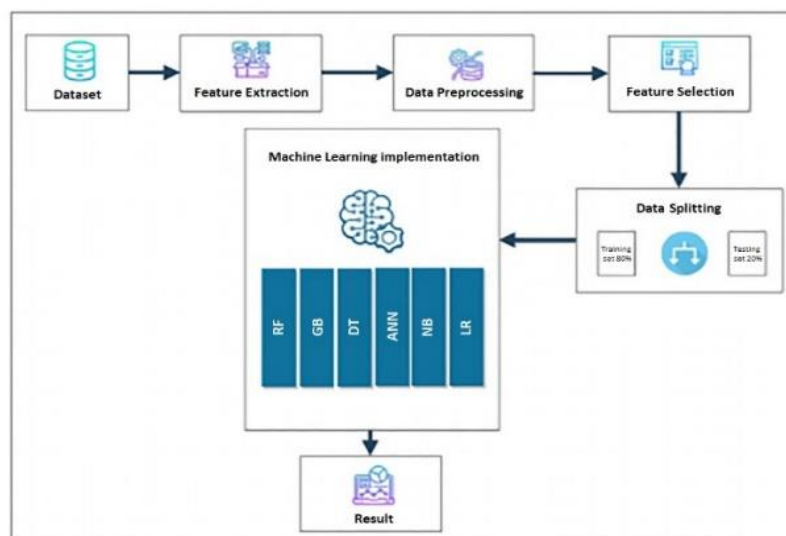


Figure 3. Overview of the proposed method

- Feature Extraction: Multiple virtual machines were connected to generate traffic like normal with Ostinato. The IoT devices were mapped into the network using Node-Red. By employing T-shark, pcap files of about 69 GB were captured. Then, using Argus and the usage of scripts in the MySQL database, network flows information was extracted from the original pcaps, consisting of 72,000,000 records and 43 features. Finally, about 5% of the original records were extracted into CSV files.
- Data preprocessing: is a method of machine learning, involving the conversion of raw information into readable format. In this paper, cleaning and conversion of data are the techniques performed to remove outliers and standardize the information so it can be used effectively to build a model.
- Feature selection: provides the procedure for reducing the input variables in increasing a predictive paradigm. It utilizes the features required for testing and training the methods in order to discover a less cumbersome resolution of protection appropriate for the systems of IoT networks. In this paper, 9 features have been selected.
- Data splitting: data are split into various categories by utilizing machine learning methods. They are also needed in both training and testing to evaluate ML approaches for their effectiveness. In our investigation, 80% of Bot-IoT metadata was analyzed for training data and the remaining 20% for testing data.
- Machine-learning approaches and implementations: full execution is realized in Anaconda by implementing the libraries of machine learning like NumPy, scikit-learn, Pandas, and Matplotlib. The algorithms of machine learning used in this Bot-IoT dataset are Naive bayes, logistic regression, gradient boosting, random forest, artificial neural network, and decision tree.

6. RESULTS AND ANALYSIS

While estimating the achievement of ML standards, it is crucial to identify models of achievement for accountability. For the estimation of the decision metrics, the most crucial measures of performance are utilized for precision, accuracy, recall, and F-measure, as explained here:

- Accuracy: the accuracy score indicates how exact the model makes a prediction in the ensemble and is one of the most accurate metric scores. It is as shown in (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

- Precision: This is a real value recognized among all the expected real values. It is presented as indicated in:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

- Recall: is a measure that indicates the quality of our model if all values are positive. This is presented as:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

- F-measure: this could refer to a measure that, by averaging its value, mixes precision and recall. It is provided as shown in (4):

$$F - \text{measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{4}$$

thus:

- True positive (TP): In a situation, the classifier correctly identified the class feature as well as the feature value is positive (in this case, an attack has been detected).
- True negative (TN): same as TP however the feature value is negative (normal traffic).
- False positive (FP): the classifier identifies the record as a malicious attack although it was in fact normal traffic.
- False negative (FN): an attack record is incorrectly categorised as normal traffic.

As discussed in the previous section, for this study work, we selected six well-known machine learning algorithms, which are: random forest, gradient boosting, decision tree, artificial neural network, Naïve bayes, and logistic regression, to accurately identify intrusions in the internet of things network environment. As shown in Figures 4-7, all the applied machine learning algorithms perform well in terms of accuracy, precision, recall, and F-measure. The most accurate in Figure 4 is the random forest classifier which reaches an accuracy of 99.99%, then comes in second place the artificial neural network classifier with a value equal to 99.91%. The last value is given by the Naïve bayes classifier with an accuracy equal to 98.13%.

Regarding the precision in Figure 5, we notice that the random forest classifier and the gradient boosting classifier both achieve the best results with a precision equal to 100%. Then comes the ANN classifier and the decision tree classifier, with a value of 99.96% and 99.95% respectively. The last value of 98.14% is given again by the Naïve Bayes classifier. For the recall in Figure 6, the random forest classifier reaches the best result with a value of 99.99% then the ANN classifier with a recall equal to 99.95%. The last value is given by the Naïve Bayes classifier which is equal to 99.12%. Finally, regarding the F-measure in Figure 7, the random forest classifier always reaches the first place with a value equal to 99.99%, followed by the ANN classifier with an F-measure equal to 99.95%. The Naïve Bayes classifier gives the lowest F-measure value. As shown in Table 1 the four metrics: accuracy, precision, recall, and F-measure in % for the six models.

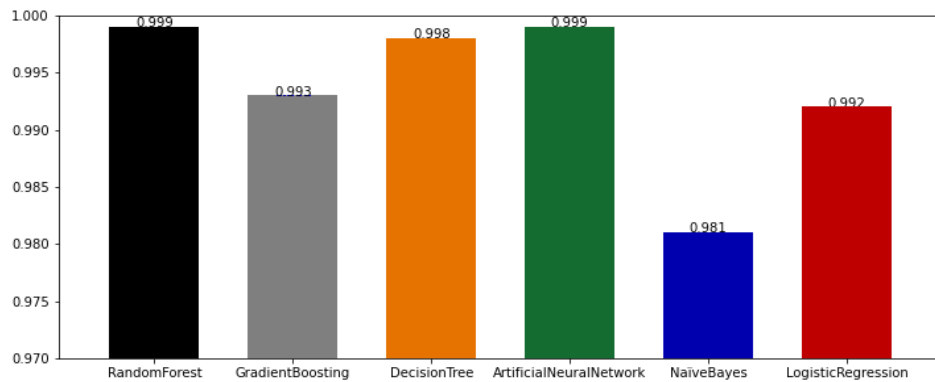


Figure 4. Accuracy

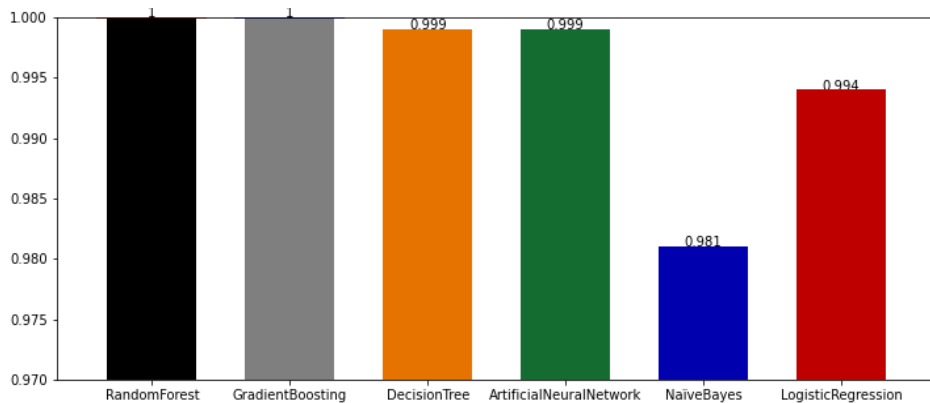


Figure 5. Precision

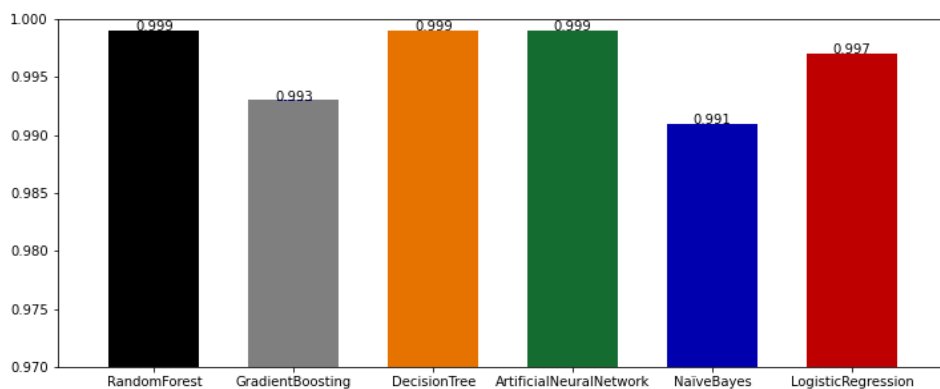


Figure 6. Recall

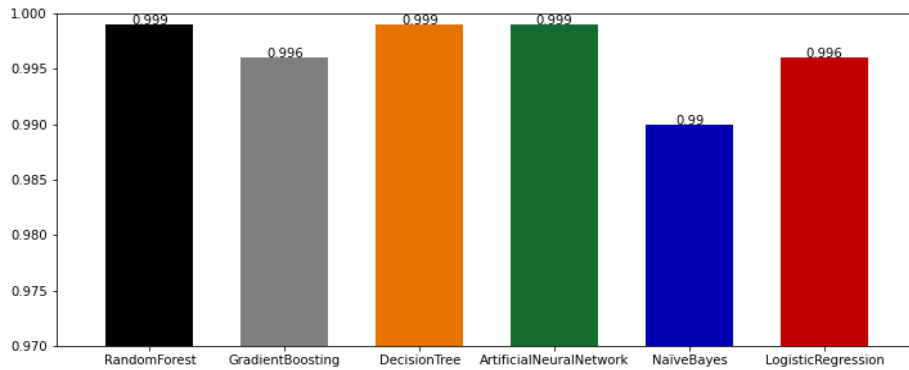


Figure 7. F_measure

Table 1. Applied machine learning algorithms results

Models	Accuracy	Precision	Recall	F-measure
Random forest classifier	99.99%	100%	99.99%	99.99%
Gradient boosting classifier	99.35%	100%	99.35%	99.67%
Decision tree classifier	99.88%	99.95%	99.93%	99.94%
Artificial neural network classifier	99.91%	99.96%	99.95%	99.95%
Naïve bayes classifier	98.13%	98.14%	99.12%	99.06%
Logistic regression	99.21%	99.42%	99.78%	99.60%

7. DISCUSSION

According to the experimental evaluation, the random forest algorithm is clearly effective for the machine learning algorithm selection among several different algorithms of machine learning and accurate detection of anomalies as well as intrusions using the Bot-IoT dataset in the IoT network. Although the obtained results from our suggested methods are hopeful, a thorough analysis provides useful and valuable information for the effective machine learning algorithm selection and the effective attack detection using machine learning algorithms in the internet of things network. In this study of research, it is evident that the random forest classifier is optimal for selecting an efficient algorithm of machine learning with respective accuracy, precision, recall, and F-measure from a set of machine learning algorithms. Analyses show that the random forest algorithm is highly effective and selected as the best machine learning algorithm among six different algorithms of machine learning including gradient boosting, decision tree, artificial neural network, nave bayes, and logistic regression. The selected machine learning performance is noteworthy for anomaly and intrusion detection in the IoT network using machine learning algorithms. As discussed before, in the present study, four different metrics, such as accuracy, precision, recall, and F-measure, were selected for evaluating the results of machine learning algorithms. Nevertheless, all machine learning algorithms applied perform well against all selected machine learning metrics, but it remains random forest, the algorithm that is extremely effective in identifying attacks in the IoT network. The experimental analysis shows that the Naïve Bayes classifier remains the only classifier among the six classifiers presented above that has the lowest performance in all scores. If we compare in a second step our results obtained in Table 1 with the results presented in subsection 2.3, we will notice that our proposed random forest model is the classifier that represents the best results for the four parameters: accuracy, precision, recall, and F-measure compared to all the models presented. Table 2 shows the work results presented in subsection 2.3.

Table 2. Performance results

Models	Accuracy	Precision	Recall	F-measure
Artificial neural network	99.42%	98.69%	99.42%	99.05%
Extreme gradient boosting	99.99%	99.50%	97.50%	98.50%
Feedforward neural networks	99.90%	99.00%	99.00%	99.90%
Deep neural network	94.00%	94.50%	93.50%	94.00%
Random neural network	83.71%	100%	75.00%	-

8. CONCLUSION

This article presents the Bot-IoT dataset that integrates the normal traffic related to IoT and additional network traffic in addition to different kinds of commonly used botnet attacks. This dataset has been evolved using a realistic testbed and labeled, with the features of the label indicating the flow of attacks, the category, and

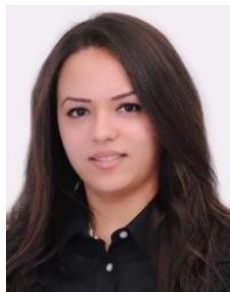
subcategory of attacks in potential purposes of multiclass classification. Other features have been generated to improve the ability of the classifiers trained on this model to predict. The following statistical analysis was used to generate from the original dataset a subset of the 9 best features. Finally, four specific metrics, including accuracy, precision, recall, and F-measure, were utilized to compare the validity of the dataset. The results showed in this paper that the intelligent models for intrusion detection, such as decision tree classifier, artificial neural network classifier, and random forest classifier, improve the performance of the detection methods significantly and achieve a perfect sensitivity on the Bot-IoT dataset. In future research, we anticipate training and evaluating our models on other intrusion detection datasets. We also envisage exploring the possibility of implementing the most accurate classifier into a real-time system for intrusion detection.




REFERENCES

- [1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, 2015, doi: 10.1109/COMST.2015.2444095.
- [2] D. Goodin, "Record-breaking DDoS reportedly delivered by >145k hacked cameras," in *ArsTechnica*, 2016, p. 1.
- [3] C. Wei, G. Xie, and Z. Diao, "A lightweight deep learning framework for botnet detecting at the IoT edge," *Computers and Security*, vol. 129, p. 103195, Jun. 2023, doi: 10.1016/j.cose.2023.103195.
- [4] G. A. P. Rodrigues *et al.*, "Cybersecurity and network forensics: Analysis of malicious traffic towards a honeynet with deep packet inspection," *Applied Sciences (Switzerland)*, vol. 7, no. 10, p. 1082, Oct. 2017, doi: 10.3390/app7101082.
- [5] C. Liu, C. Yang, X. Zhang, and J. Chen, "External integrity verification for outsourced big data in cloud and IoT: A big picture," *Future Generation Computer Systems*, vol. 49, pp. 58–67, Aug. 2015, doi: 10.1016/j.future.2014.08.007.
- [6] C. Grajeda, F. Breitingner, and I. Baggili, "Availability of datasets for digital forensics – And what is missing," *Digital Investigation*, vol. 22, pp. S94–S105, Aug. 2017, doi: 10.1016/j.diin.2017.06.004.
- [7] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018, vol. 2018-January, pp. 108–116, doi: 10.5220/0006639801080116.
- [8] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings*, Nov. 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.
- [9] K. Asn, "That 'internet of things' thing," *RFid Journal*, p. 4986, 2009.
- [10] M. U. Ahmed, M. Björkman, A. Čaušević, H. Fotouhi, and M. Lindén, "An overview on the internet of things for health monitoring systems," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICTS*, vol. 169, 2016, pp. 429–436.
- [11] T. Truong, A. Dinh, and K. Wahid, "An IoT environmental data collection system for fungal detection in crop fields," in *Canadian Conference on Electrical and Computer Engineering*, Apr. 2017, pp. 1–4, doi: 10.1109/CCECE.2017.7946787.
- [12] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of things (IoT): a literature review," *Journal of Computer and Communications*, vol. 03, no. 05, pp. 164–173, 2015, doi: 10.4236/jcc.2015.35021.
- [13] D. Andriessse, C. Rossow, B. Stone-Gross, D. Plohmann, and H. Bos, "Highly resilient peer-to-peer botnets are here: An analysis of Gameover Zeus," in *Proceedings of the 2013 8th International Conference on Malicious and Unwanted Software: "The Americas"*, *MALWARE 2013*, Oct. 2013, pp. 116–123, doi: 10.1109/MALWARE.2013.6703693.
- [14] S. S. C. Silva, R. M. P. Silva, R. C. G. Pinto, and R. M. Salles, "Botnets: a survey," *Computer Networks*, vol. 57, no. 2, pp. 378–403, Feb. 2013, doi: 10.1016/j.comnet.2012.07.021.
- [15] P. Amini, M. A. Araghizadeh, and R. Azmi, "A survey on Botnet: classification, detection and defense," in *Proceedings - 2015 International Electronics Symposium: Emerging Technology in Electronic and Information, IES 2015*, Sep. 2016, pp. 233–238.
- [16] A. Djenna, D. E. Saidouni, and W. Abada, "A pragmatic cybersecurity strategies for combating IoT-cyberattacks," in *2020 International Symposium on Networks, Computers and Communications, ISNCC 2020*, Oct. 2020, pp. 1–6, doi: 10.1109/ISNCC49221.2020.9297251.
- [17] M. A. Lawal, R. A. Shaikh, and S. R. Hassan, "An anomaly mitigation framework for iot using fog computing," *Electronics (Switzerland)*, vol. 9, no. 10, pp. 1–24, Sep. 2020, doi: 10.3390/electronics9101565.
- [18] M. Ge, X. Fu, N. Syed, Z. Baig, G. Teo, and A. Robles-Kelly, "Deep learning-based intrusion detection for IoT networks," in *Proceedings of IEEE Pacific Rim International Symposium on Dependable Computing, PRDC*, Dec. 2019, vol. 2019-December, pp. 256–265, doi: 10.1109/PRDC47002.2019.00056.
- [19] P. Jithu, J. Shareena, A. Ramdas, and A. P. Haripriya, "Intrusion detection system for IoT Botnet attacks using deep learning," *SN Computer Science*, vol. 2, no. 3, p. 205, May 2021, doi: 10.1007/s42979-021-00516-9.
- [20] K. Filus, J. Domańska, and E. Gelenbe, "Random neural network for lightweight attack detection in the IoT," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12527 LNCS, 2021, pp. 79–91.
- [21] N. Moustafa, "The Bot-IoT dataset," *IEEE Dataport*, 2019, doi: 10.21227/r7v2-x988.
- [22] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, Nov. 2019, doi: 10.1016/j.future.2019.05.041.
- [23] J. Han, M. Kamber, F. Berzal, and N. Marin, "Data mining: concepts and techniques," *SIGMOD Record*, vol. 31, no. 2, pp. 66–68, Jun. 2002, doi: 10.1145/565117.565130.
- [24] S. Peter, F. Diego, F. A. Hamprecht, and B. Nadler, "Cost efficient gradient boosting," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 1552–1562, 2017, doi: 10.5555/3294771.3294919.
- [25] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jast20165.
- [26] R. Dastres and M. Soori, "Artificial neural network systems," *International Journal of Imaging and Robotics (IJIR)*, vol. 2021, no. 2, pp. 13–25, 2021.
- [27] Y. C. Zhang and L. Sakhanenko, "The naive Bayes classifier for functional data," *Statistics and Probability Letters*, vol. 152, pp. 137–146, Sep. 2019, doi: 10.1016/j.spl.2019.04.017.




- [28] R. D. Ravipati and M. Abualkibash, "Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets - a review paper," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3428211.
- [29] M. Maalouf, "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011, doi: 10.1504/IJDATS.2011.041335.

BIOGRAPHIES OF AUTHORS






Imane Kerrakchou    graduated as an engineer in electronic systems, computer science and networks from the National School of Applied Sciences at Mohammed First University, Oujda, Morocco in 2017. She is currently a Ph.D. student in Computer Science in the Department of Electronics, Computer Science and Telecommunications at the National School of Applied Sciences, Mohammed First University, Oujda, Morocco. Her research areas are IoT security, attack modeling, and machine learning models of cybersecurity. She can be contacted at email: i.kerrakchou@ump.ac.ma.






Adil Abou El Hassan    received the B.Sc. degree in electrical engineering from Hassan II University (ENSET), Mohammedia, Morocco, in 1993 and the M.Sc. degree in electrical engineering from Mohammed V University (ENSET), Rabat, Morocco, in 2019. He is currently pursuing the Ph.D. degree at Mohammed First University (National School of Applied Sciences), Oujda, Morocco. His research interests include the internet of things (IoT), 5G wireless communication and networks, LPWAN technologies and resource management of wireless communications. He can be contacted at email: a.abouelhassan@ump.ac.ma.






Sara Chadli    graduated as a network and telecommunication engineer from National School of Applied Sciences, Oujda, Morocco, in 2012. She received her Ph.D. in telecommunication and electronics engineering from Mohammed first University in 2016. she is currently an assistance professor in the Department of physiques of Mohammed first University. Her research interests include includes mobile Ad Hoc networks (MANET), network security, modeling and control of advanced electrical power systems, design, application of power electronics and telecommunications engineering. She can be contacted at email: s.chadli@ump.ac.ma.



Mohamed Emharraf    is a professor of robotics at National School of Applied Sciences, Mohamed First University, Oujda, Morocco. He received his Ph.D. in 2017 from CEDOC-EMPO. His research interests include Indoor robot control, IoT security, Smart agricultural, computer engineering, human-computer, interaction, and artificial intelligence. He has published more than 30 papers in peer-reviewed journals and conference proceedings, and has served as a reviewer for several scientific journals and as a program committee member. He can be contacted at email: m.emharraf@ump.ac.ma.



Mohammed Saber    is currently an associate professor in the Department of Electronics, Computer Science and Telecommunications at National School of Applied Sciences at Mohammed First University, Oujda, Morocco (2013). He received a Ph.D. in Computer Science at Faculty of Sciences, Oujda, Morocco, in July 2012, an engineer degree in Network and Telecommunication at National School of Applied Sciences, in July 2004, and Licence degree in Electronics at Faculty of Sciences, in July 2002, all from Mohammed First University, Oujda. He is currently director of Smart Information, Communication & Technologies Laboratory (SmartICT Lab). His interests include network security (intrusion detection system, evaluation of security components, security IoT), AI, robotics, and embedded systems. He can be contacted at email: m.saber@ump.ac.ma.