# Detecting Botnet Attacks in IoT Environments: An Optimized Machine Learning Approach

MohammadNoor Injadat, Abdallah Moubayed, and Abdallah Shami
Electrical and Computer Engineering Department, Western University, London, Ontario, Canada
emails: {amoubaye, minjadat, abdallah.shami}@uwo.ca

2020 32nd International Conference on Microelectronics (ICM) | 978-1-7281-9664-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICM50269.2020.9331794

***Abstract*—The increased reliance on the Internet and the corresponding surge in connectivity demand has led to a significant growth in Internet-of-Things (IoT) devices. The continued deployment of IoT devices has in turn led to an increase in network attacks due to the larger number of potential attack surfaces as illustrated by the recent reports that IoT malware attacks increased by 215.7% from 10.3 million in 2017 to 32.7 million in 2018. This illustrates the increased vulnerability and susceptibility of IoT devices and networks. Therefore, there is a need for proper effective and efficient attack detection and mitigation techniques in such environments. Machine learning (ML) has emerged as one potential solution due to the abundance of data generated and available for IoT devices and networks. Hence, they have significant potential to be adopted for intrusion detection for IoT environments. To that end, this paper proposes an optimized ML-based framework consisting of a combination of Bayesian optimization Gaussian Process (BO-GP) algorithm and decision tree (DT) classification model to detect attacks on IoT devices in an effective and efficient manner. The performance of the proposed framework is evaluated using the Bot-IoT-2018 dataset. Experimental results show that the proposed optimized framework has a high detection accuracy, precision, recall, and F-score, highlighting its effectiveness and robustness for the detection of botnet attacks in IoT environments.**

***Keywords*—IoT, Botnet Detection, Bayesian Optimization, Decision Trees**

## I. Introduction

The increased reliance on the Internet and the corresponding surge in connectivity demand has led to a significant growth in Internet-of-Things (IoT) devices. This is supported by the recent projections that the number of connected devices will reach around 28.5 billion devices by 2022 [1], with these IoT devices covering multiple use cases such as healthcare [2], smart cities [3], and intelligent transportation systems [4].

The continued deployment of IoT devices has in turn led to an increase in network attacks due to the larger number of potential attack surfaces. This is substantiated by Forbes' recent report stating that more than 2.9 billion events in 2019, a three-fold increase from the previous year [5]. Moreover, SonicWall reported that IoT malware attacks increased by 215.7% from 10.3 million in 2017 to 32.7 million in 2018 [6]. This illustrates the increased vulnerability and susceptibility of IoT devices and networks. Therefore, there is a crucial need for proper effective and efficient attack detection and mitigation techniques with researchers increasingly investigating and proposing multiple potential mechanisms [7]–[9].

Machine learning (ML) has emerged as one potential solution due to the abundance of data generated and available for IoT devices and networks. ML allows systems to be dynamic and flexible to new inputs as they can "learn" without explicitly being told what to do [10]. Additionally, ML techniques have illustrated their effectiveness and efficiency in various applications and use cases [11]–[16]. Therefore, they have significant potential to be adopted for intrusion detection for IoT environments. Additionally, optimizing the parameters of these ML models is crucial to further enhance their detection performance and effectiveness [17]–[19].

Therefore, this paper proposes an optimized ML-based framework consisting of a combination of Bayesian optimization Gaussian Process (BO-GP) and decision tree (DT) classification model to detect botnet attacks on IoT devices. The goal is to develop a dynamic, effective, and efficient IoT attack detection framework. As such, the main contributions of this work can be summarized as follows:

- *Proposing* a combination of BO-GP and DT classifier to detect botnet attacks on IoT devices.
- *Evaluating* the performance of the proposed model using a recent IoT dataset titled Bot-IoT-2018.

The remainder of this paper is organized as follows: Section II briefly surveys the literature. Section III describes the proposed approach for IoT botnet detection and illustrates its complexity. Section IV describes the dataset investigated. Section V presents the experimental setup and the obtained results. Finally, Section VI concludes the paper.

## II. Related Work

The recent surge in computing capabilities has led to an increase in investigating ML techniques and algorithms as an effective solution for network security [20]–[23]. For example, Li *et al.* proposed such models for intelligent transportation systems [20]. More specifically, the authors developed tree-based classification models in an attempt to detect intrusions in autonomous vehicles [20]. In contrast, Injadat *et al.* proposed an optimized ML-based network intrusion detection framework using Bayesian optimization [21]. Their experiments showed that the proposed model had a higher detection accuracy and a lower false alarm rate [21]. In a similar manner, Injadt *et al.* also proposed a multi-stage optimized ML-based intrusion detection framework that reduced the computational complexity while simultaneously improving the detection accuracy [22]. Salo *et al.* also proposed the use of ML classification techniques for intrusion detection [23]. The authors proposed the combined use of ensemble feature selection and clustering-enabled classification models to detect unseen attack patterns [23].

Within the specific context of IoT, multiple researchers proposed ML-based solutions to detection various attacks in such environments [24]–[26]. Teixeira *et al.* investigated five

different ML classification algorithms to detect network attacks in an industrial IoT environment, namely in a water storage tank's control system [24]. In a similar fashion, Almiani *et al.* proposed the use of deep recurrent neural networks to effectively detect network intrusions in IoT environments with the model showing high detection accuracy [25]. In contrast, Anthi *et al.* proposed the use of various supervised ML classification models to detect four different types of IoT intrusion attacks in a smart home environment [26]. The authors show that the proposed models had a high precision and recall values as well as a low classification time, illustrating the real-time deployment potential of these models [26]. However, many of the frameworks dedicated to IoT environments use default parameters rather than optimized parameters for the different ML algorithms proposed as well as avoid addressing the class imbalance issue commonly found in network attack scenarios. As such, there is a need to further improve the detection performance by proposing optimized ML models.

## III. Proposed Approach

### A. Proposed Approach Description

This paper proposes combining the BO-GP algorithm to optimize the parameters of the DT classification model as part of an optimized ML-based framework for effective and efficient detection of botnet attacks on IoT devices. The proposed approach, as shown in Fig.1, consists of two components, namely:

1) Data pre-processing: The goal of this component is to prepare the data into a format that would maximize the performance of the developed ML classification model. As such, this is done by initially normalizing the features using the min-max method. The feature normalization step helps unify the dynamic range of the different features so that no single feature dominates in the model training stage. This is done using the following equation [21]:

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

After feature normalization, Synthetic Minority Oversampling TEchnique (SMOTE) is applied to oversample the minority class with the aim of addressing the class imbalance problem encountered in such environments [27]. It is worth noting that SMOTE is proposed as it can generate new high quality instances of the minority class [28], resulting in an enhanced classification model performance and a reduced training sample size [28].

2) Hyper-parameter Optimization: The goal of this component is to optimize the hyper-parameters of the ML model to ensure that the detection performance is maximized. To that end, BO-GP algorithm is proposed in this work to optimize the hyper-parameters of the DT model. BO-GP is one version of the Bayesian Optimization group of algorithms which belong to the class of probabilistic global optimization models [22]. In our case, this is represented as the set of suitable values for the DT hyper-parameters. The BO-GP algorithm is chosen in this case since it can identify near-optimal hyper-parameter combinations within a few iterations [17].

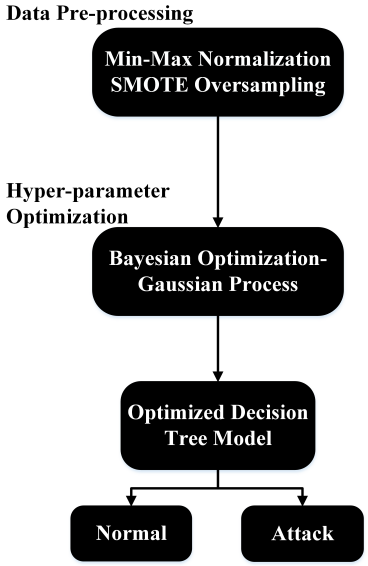


Fig. 1. Proposed Optimized IoT Botnet Detection Framework

Note that the proposed framework does not include a feature selection component. This is attributed to the fact that in many cases, data collected from IoT devices often contains a small number of bits and features due to their limited computing power. This is evident in multiple IoT deployment scenarios such as in smart homes [26] and autonomous vehicles [20].

### B. Complexity of Proposed Approach

The overall complexity of the proposed framework depends on the complexity of each of its components. Assume that the considered dataset is composed of $M$ instances and $N$ features. The complexity of the min-max feature normalization method is $O(N)$. This is because the algorithm determines the minimum and maximum of each feature to normalize the dataset. The complexity of the SMOTE algorithm is $O(M_{min}^2 N)$ where $M_{min}$ is the number of minority class instances [29]. The complexity of the BO-GP method for the hyper-parameter optimization stage is $O(M^3)$ [22]. Finally, the complexity of the DT classification model is $O(M^2 N)$. However, the training time can be significantly reduced to approximately $O(\frac{M^2 N}{threads})$ where $threads$ is the maximum number of participating threads. This is because multi-threading is enabled when using this classifier [20]. Therefore, the overall complexity of the proposed framework is $O(M^3)$, making it computationally efficient.

## IV. Dataset Description

The dataset considered in this work is the Bot-IoT-2018 dataset developed in [30]. The dataset is built by designing a realistic IoT network environment using three main components: network platforms, simulated IoT services, and feature extraction platform. The network platforms consist of different virtual machines (VMs) acting either as normal or attacking VMs. The simulated IoT services where generated using the Node-red tool to mimic the network behavior of IoT devices such as a weather station, a smart fridge, motion activated lights, garage door, and a smart thermostat. Finally, the feature extraction platform used

Argus tool to extract the corresponding data features from the collected pcap files [30]. The resulting dataset contains close to 72 million records and 46 features. For the purpose of this work, the reduced dataset consisting of around 3.6 million records (representing 5% of the data as recommended by the dataset authors in [30]) and the 10 best features is used for brevity. More specifically, the reduced dataset contains 477 **normal** instances and 3,668,045 **attack** instances. This illustrates that the dataset is significantly imbalanced. Fig. 2 plots the first and second principal components of the dataset. It can be clearly seen that the dataset is highly non-linear and suffers from significant class imbalance, an issue commonly encountered in the network security field. This reiterates the need for applying data oversampling to ensure that the trained ML model has enough instances of each class to learn from.

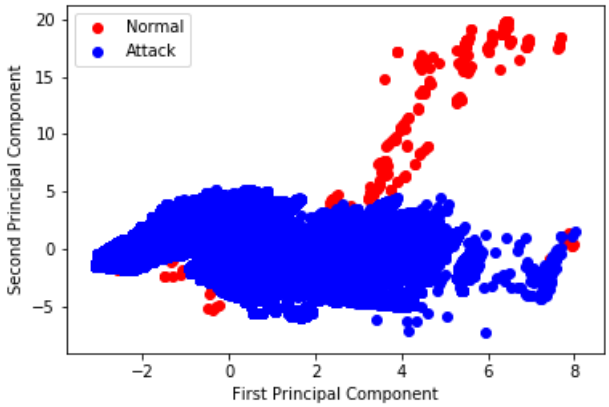


Fig. 2. First and Second Principal Components of Bot-IoT-2018 Dataset

## V. Experiment Results & Discussion

### A. Experiment Setup

The following settings were used to conduct the experiments in this work:

1) Software: Python 3.7.4 running on Anaconda's Jupyter Notebook.
2) Hardware: Intel® Core™ i7-9750H CPU 6 Cores at 2.6 GHZ and 16GB of memory running Windows 10.

### B. Results & Discussion

To evaluate the performance of the proposed optimized DT-based framework, its performance is compared to a default DT classifier and the support vector machine (SVM) model proposed in [30] using the following metrics [21]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

TABLE I
Performance Evaluation of Classifiers

| Algorithm | Testing Accuracy (%) | Precision | Recall | F-score |
|---|---|---|---|---|
| Default DT | 99.82 | 0.53 | 0.91 | 0.56 |
| SVM [30] | 88.37 | 1 | 0.88 | 0.94 |
| Optimized DT-based Framework | **99.99** | **0.99** | **1.00** | **1.00** |

where $TP$ is the number of true positive instances, $TN$ is the number of true negative instances, $FP$ is the number of false positive instances, and $FN$ is the number of false negative instances.

It can be seen in Table I that the proposed optimized DT-based model outperforms the default model as well as the SVM model. More specifically, the proposed framework improved all the performance metrics including the testing accuracy, precision, recall, and F-score. For example, the results show that the default DT model achieves relatively low precision and high recall. This means that the model is incorrectly flagging normal traces as attacks. This is undesirable since this can lead to legitimate users being blocked or blacklisted. This behavior is attributed to the fact that the dataset used is significantly imbalanced with much more attack traces than normal ones. In a similar manner, the SVM model also had a lower recall value. In contrast, the proposed optimized DT-based framework was able to improve both the precision and recall values. This means that the model is effective in correctly identifying and differentiating between normal and attack instances. This is due to the oversampling performed to ensure that the model has enough instances of both classes to learn from. These results highlight the effectiveness and robustness of the proposed framework for botnet detection in IoT environments.

## VI. Conclusion & Future Works

A significant growth in the deployment of Internet-of-Things (IoT) devices has been observed in recent years due to the increased reliance on the Internet and the corresponding surge in connectivity demand. This is supported by the recent projections that the number of connected devices will reach around 28.5 billion devices by 2022. In turn, this has led to an increase in network attacks due to the larger number of potential attack surfaces. Therefore, proper effective and efficient attack detection and mitigation techniques are needed to ensure these devices are well protected.

To that end, this paper proposed an optimized ML-based framework that combined Bayesian optimization Gaussian Process (BO-GP) and decision tree (DT) classification model to detect botnet attacks on IoT devices. The goal was to develop a dynamic, effective, and efficient IoT attack detection framework. Experimental results showed that the proposed optimized DT-based framework improved the accuracy, precision, recall, and F-score. More specifically, it achieved values of 99.99%, 0.99, 1.00, and 1.00 for these four metrics respectively. This illustrated that the proposed framework is both effective and robust in detecting botnet attacks in IoT environments.

This work can be extended in multiple directions. One intuitive direction is to use the complete dataset to ensure that more normal instances are used as part of the data oversampling

process to further enrich the normal traces scenario. Another direction worth exploring is to investigate the time-related features to identify any temporal behaviors or patterns that may be helpful in detecting botnet attacks in IoT environments.

## REFERENCES


[1] Cisco, "Cisco Predicts More IP Traffic in the Next Five Years Than in the History of the Internet," Nov. 2018.

[2] Z. Alansari, S. Soomro, M. R. Belgaum, and S. Shamshirband, "The rise of internet of things (iot) in big healthcare data: review and open research issues," in *Progress in Advanced Computing and Intelligent Engineering*. Springer, 2018, pp. 675–685.

[3] H. Arasteh, V. Hosseinnezhad, V. Loia, A. Tommasetti, O. Troisi, M. Shafie-khah, and P. Siano, "Iot-based smart cities: A survey," in *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*, 2016, pp. 1–6.

[4] I. Al Ridhawi, M. Aloqaily, B. Kantarci, Y. Jararweh, and H. T. Mouftah, "A continuous diversified vehicular cloud service availability framework for smart cities," *Computer Networks*, vol. 145, pp. 207–218, 2018.

[5] Z. Doffman, "Cyberattacks on iot devices surge 300% in 2019,'measured in billions,'report claims," 2019.

[6] C. Crane, "20 surprising iot statistics you don't already know," 2019.

[7] A. Moubayed, A. Refaey, and A. Shami, "Software-defined perimeter (sdp): State of the art secure solution for modern networks," *IEEE Network*, vol. 33, no. 5, pp. 226–233, Sep.- Oct. 2019.

[8] P. Kumar, A. Moubayed, A. Refaey, A. Shami, and J. Koilpillai, "Performance analysis of sdp for secure internal enterprises," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2019, pp. 1–6.

[9] H. Hindy, D. Brosset, E. Bayne, A. K. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A taxonomy of network threats and the effect of current datasets on intrusion detection systems," *IEEE Access*, vol. 8, pp. 104 650–104 675, 2020.

[10] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-learning: Challenges and research opportunities using machine learning data analytics," *IEEE Access*, vol. 6, pp. 39 117–39 138, 2018.

[11] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-learning environment: Clustering using k-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, 2020.

[12] ——, "Relationship between student engagement and performance in e-learning environment using association rules," in *2018 IEEE World Engineering Education Conference (EDUNINE)*, 2018, pp. 1–6.

[13] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-based Systems*, vol. 200, p. 105992, Jul. 2020.

[14] ——, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, pp. 1–23, Jul. 2020.

[15] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "DNS Typo-Squatting Domain Detection: A Data Analytics & Machine Learning Based Approach," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–7.

[16] A. Moubayed, E. Aqeeli, and A. Shami, "Ensemble-based feature selection and classification model for dns typo-squatting detection," in *2020 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, Aug. 2020.

[17] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231220311693

[18] A. Moubayed, "Optimization Modeling and Machine Learning Techniques Towards Smarter Systems and Processes," Ph.D. dissertation, University of Western Ontario, Aug. 2018.

[19] M. Injadat, "Optimized Machine Learning Models Towards Intelligent Systems," Ph.D. dissertation, University of Western Ontario, Aug. 2018.

[20] L. Yang, A. Moubayed, I. Hamieh, and A. Shami, "Tree-based intelligent intrusion detection system in internet of vehicles," in *2019 IEEE Global Communications Conference (GLOBECOM)*, Dec 2019, pp. 1–6.

[21] M. Injadat, F. Salo, A. B. Nassif, A. Essex, and A. Shami, "Bayesian optimization with machine learning algorithms towards anomaly detection," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.

[22] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-stage optimized machine learning framework for network intrusion detection," *IEEE Transactions on Network and Service Management*, pp. 1–1, Aug. 2020.

[23] F. Salo, M. Injadat, A. Moubayed, A. B. Nassif, and A. Essex, "Clustering enabled classification using ensemble feature selection for intrusion detection," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 276–281.

[24] M. A. Teixeira, T. Salman, M. Zolanvari, R. Jain, N. Meskin, and M. Samaka, "Scada system testbed for cybersecurity research using machine learning approach," *Future Internet*, vol. 10, no. 8, p. 76, 2018.

[25] M. Almiani, A. AbuGhazleh, A. Al-Rahayfeh, S. Atiewi, and A. Razaque, "Deep recurrent neural network for iot intrusion detection system," *Simulation Modelling Practice and Theory*, vol. 101, p. 102031, 2020.

[26] E. Anthi, L. Williams, M. Słowińska, G. Theodorakopoulos, and P. Burnap, "A supervised intrusion detection system for smart home iot devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 9042–9053, 2019.

[27] Z. Chen, Q. Yan, H. Han, S. Wang, L. Peng, L. Wang, and B. Yang, "Machine learning based mobile malware detection using highly imbalanced network traffic," *Information Sciences*, vol. 433, pp. 346–364, 2018.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[29] F. Hu and H. Li, "A novel boundary oversampling algorithm based on neighborhood rough set model: Nrsboundary-smote," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[30] 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, 2019.