Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 187-192, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 187-192, December 2021

Research Article

The Effects of Normalization and Standardization an Internet of Things Attack Detection

Gozde Karatas[1*]

[1] Biruni Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0003-2303-9410)

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1017427)

ATIF/REFERENCE: Karatas, G. (2020). The Effects of Normalization and Standardization an Internet of Things Attack Detection. *European Journal of Science and Technology*, (29), 187-192.

Abstract

It is a known fact that we live in the computer age and that many devices in the world have access to the internet. So how secure are these devices? Is there any guarantee that user information is not accessed from intruder? After the concept of the Internet of Things came into our lives, many things such as seeing the food in our home refrigerator, connecting to the Internet from the car and, and video chatting from our smart watch entered our lives. The number of malicious software is also increasing with these new connections. Researchers are increasingly emphasizing the importance of network security and intensifying their studies.
Data preprocessing is very important when designing a secure system. In this study, the importance of normalization and standardization in data preprocessing is examined to make machine learning approaches more successful for detecting attacks on IoT devices. The study was carried out in Logistic Regression, Decision Tree, and Stochastic Gradient Descent machine learning algorithms using the Bot-IoT dataset. Bot-IoT dataset is a popular dataset that is widely used in security studies on IoT devices. Normalization and standardization processes were applied to Bot-IoT dataset separately, so data preprocessing was performed, then selected machine learning algorithms were trained with these -normalized / standardized- datasets. As a result of the trainings made with machine learning algorithms, the values of Accuracy, Precision, Recall and F1 Score rates were examined. And as a result of the study, it was seen that the standardization increased the accuracy rate up to 99.96% in Logistic Regression.

Keywords: Bot-IoT, IoT intrusion, Machine learning, Intrusion detection.

Normalizasyon ve Standardizasyonun Nesnelerin İnterneti Saldırılarındaki Etkileri

Öz

Bilgisayar çağında yaşadığımız ve dünyadaki birçok cihazın internete erişimi olduğu herkes tarafından bilinen bir gerçektir. Peki bu internete bağlanan cihazlar ne kadar güvenlidir? Davetsiz misafirlerden –saldırgan- kullanıcı bilgilerine erişilmeyeceğine dair herhangi bir garanti verilebilir mi mı? Nesnelerin İnterneti (IoT) kavramının hayatımıza girmesinden sonra evdeki buzdolabında bulunan yiyecekleri görmek, arabanın içinden internete bağlanmak, kullanılan akıllı saatten görüntülü sohbet etmek gibi pek çok şey insan hayatına girmiştir. Bu yeni kavramlar ile birlikte kötü amaçlı yazılımların ve saldırganların da sayısı artmaktadır. Bu konularda çalışma yapan araştırmacılar giderek artan veri sayısına bağlı olarak ağ güvenliğinin önemini vurgulamakta ve çalışmalarını bu alanda yoğunlaştırmaktadır.
Güvenli bir saldırı tespit sistemi tasarlarken veri ön işleme en önemli aşamalardan biridir. Ve IoT cihazlarında bu alanlarda yapılan çalışmalar hem dikkat çekmektedir hem de hız kazanmıştır. Yapılan bu çalışmada, IoT cihazlarına yönelik saldırıları tespit etmede makine öğrenmesi yaklaşımlarını daha başarılı kılmak için veri ön işlemede normalizasyon ve standardizasyonun önemini incelemek hedeflenmiştir. Buna göre çalışma, Bot-IoT veri kümesi kullanılarak Lojistik Regresyon, Karar Ağacı ve Stokastik Gradyan Arttırma makine öğrenme algoritmaları üzerinde gerçekleştirilmiştir. Bot-IoT veri kümesi, IoT cihazlarında güvenliği sağlamak için yapılan çalışmalarda yaygın olarak kullanılan popüler bir veri kümesidir. Seçilen bu veri kümesine veri ön işleme yapılmıştır, bunun için ayrı ayrı normalizasyon ve standardizasyon işlemleri uygulanmış ardından seçilen bu –normalize/standardize edilmiş- veri kümeleri ile belirlenen makine öğrenmesi algoritmaları eğitilmiş ve test edilmiştir. Makine öğrenmesi algoritmaları ile yapılan eğitimler sonucunda Doğruluk, Kesinlik, Geri Çağırma ve F1 Skor sonuçlarının değerleri incelenmiştir. Yapılan çalışma sonucunda ise Lojistik Regresyonda standardizasyonun doğruluk oranını %99.96'ya kadar arttırdığı görülmüştür.

Anahtar Kelimeler: Bot-IoT, IoT intrusion, Machine learning, Intrusion detection

[*] Corresponding Author: Biruni Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, ORCID: 0000-0003-2303-9410, gbaydogmus@biruni.edu.tr

1. Introduction

The internet has started to include in every aspect of our lives with the developing technology. Smart phones, smart watches, smart cities, etc. have become an indispensable part of our lives. Such concepts, in which the Internet is connected and data flow is provided are called the Internet of Things (Luo et al., 2020).

In the following years, it is aimed that data flow will be realized even from the thermos where we drink coffee, the chair we sit on, the clothes we wear and many more. Thanks to the increasing artificial intelligence studies in recent years, it has been seen that many things we see in fiction films are now possible. The need for security has started to increase with the increasing data with IoT technology.

Big data flow will also introduce uncontrollable security vulnerabilities. Network security, which is the biggest problem of our age, is now a big problem not only for computers but also for every device connected to the internet (Luo et al., 2020; Abbasi et al. 2021; Shafiq et al., 2020). Researchers have been working to solve this problem for years, and especially with the graphics card technology that has developed in recent years, machine learning and artificial intelligence algorithms allow network security studies to be more successful.

One of the most important things for people in all areas of life has been security. Approaches such as personal security, property security, device security have been in our lives for a very long time. The common name given to the systems used to ensure security in computers is gathered under the title of Intrusion Detection System (IDS). Before explaining IDS, it is necessary to look at the concept of attack. Actions taken by malicious people to harm something or a person are called attacks. In computer science, it can be explained with examples such as making a network dysfunctional by sending too many requests, accessing interpersonal information secretly, changing the information of users. Here, systems that prevent/warn such attacks by noticing in advance are called intrusion detection systems. It is very important and necessary to develop IDS for IoT devices. Figure 1 shows IoT device security.

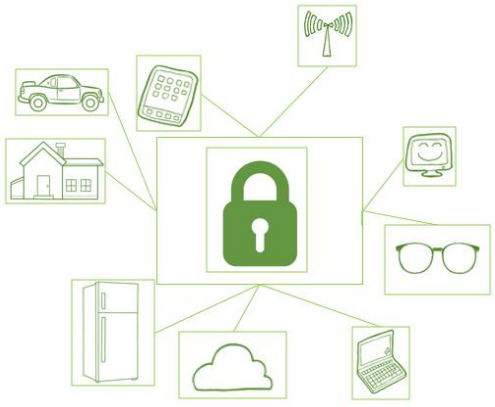


Figure 1. IoT Security with IDS

It seen that normalization is generally used for data preprocessing when the literature is examined. Researchers overlook that standardization can further increase the accuracy of the model. In this study, the effects of normalization and standardization methods on machine learning algorithms used to provide security in IoT devices were examined. The aim of the study is to obtain both faster and more successful algorithms with dataset normalization and standardization. In the study, the performances of Logistic Regression, Decision Tree and Stochastic Gradient Descent algorithms were examined using the Bot-IoT dataset. For this, Min-Max normalization and Standard Scaling methods were applied on Bot-IoT dataset after data preprocessing. Then, the selected machine learning algorithms were examined separately with both Min-Max and Standard Scaling and the results were compared. As a result of the study, it was seen that the standardization greatly increased the performance in IoT attack detection.

The following are mentioned in the rest of the study; Section II gives literature review, Section III describes the proposed methodology, normalization and standardization methods, used dataset and algorithms. Section IV describes the experimental results and finally, Section VI gives the conclusion and future work.

2. Material and Method

In this section, the content of the study is discussed in detail. First, information about similar studies in the literature is given, and then the important concepts of the proposed system are explained. These are Normalization, Standardization and Machine Learning algorithms used. In addition, the Bot-IoT dataset used in the study and the data preprocessing stages are also mentioned.

2.1. Related Work

Luo and his friends proposed an ensemble learning algorithm using deep learning methods for find to ease the problems that IoT systems have (Luo et al., 2020). They created the model using three deep learning algorithms, these algorithms are Long Short Term Memories, Convolutional Neural Networks and MRN. They used HTTP CSIC dataset 2010 dataset for the experiment which is public for researchers. In the end results show that the developed system can findIoT web attacks completely with high accuracy rate.

In 2021 researchers proposed a model using feature selection methods with machine learning algorithms for IoT attack detection (Abbasi et al., 2021). They used Logistic Regression to extract features from the dataset and then applied Artificial Neural Networks for classification. The n_Balot dataset used to examine the performance of the proposed model. The developed model was found to be much more successful compared to other existing systems so the use of logistic regression in feature extraction was suggested.

Shafiq and others has done some research to ensure the IoT security of smart cities (Shafiq et al., 2020). Realizing that there are incomplete studies in this area, the researchers proposed a new model using feature selection and machine learning algorithms, these algorithms are Naive Bayes, BayesNet, Decision Tree C4.5, Random Forest and Random Tree. Five machine learning algorithms were examined while developing the study and method Bijective soft set technique was used to see which of these algorithm was successful and examination was done according to this algorithm at the end of the study. Experimental results showed that the developed model is more efficient for the selected machine learning algorithm.

Ferrag et al. aimed to develop a system that will provide IOT security by using rule-based systems and decision trees (Ferrag et al., 2020). REP Tree and Forest PA algorithms were used as decision tree algorithms and JRip algorithm were used as rule-based algorithms. CICIDS2017 and Bot-IoT datasets were used to test the developed algorithm. In the end, results showed that proposed algorithm was found to be better when compared to other methods.

Detecting bot attacks on IoT technologies is very important for a successful system. In 2020 researchers conducted a study to see how machine learning algorithms with feature selection work (Alshamkhany et al., 2020). The features in the dataset that do not affect the classification are removed with the feature selection. For feature selection researchers used Principal Component Analysis approach. After feature selection, the system was tested with certain machine learning algorithms. These algorithms are Naive Bayes, K-Nearest Neighbor, Support Vector Machine, and Decision Tree. In addition, UNSW-NB15 dataset was used for the study. As a result of the study, the best algorithm was found to be Decision Tree which works with the feature selection dataset with 99.89% accuracy rate.

There is a need for successful software to protect systems with the increase in IoT devices (Injadat et al., 2020). Machine learning has come to the fore as a successful solution due to the increasing IoT data in recent years. Injadat et al. have created a hybrid model combining Bayesian optimization Gauss Process and decision tree algorithms to detect attacks against IoT devices. Examination on the model were carried out with the Bot-IoT dataset. As a result of the study, it has been seen that the proposed model has a high detection accuracy for the detection of attacks in IoT environments.

In 2017, Shukla examined the effects of clustering and classification algoithms working together on IoT (Shukla, 2017). For this, he ran the selected dataset separately with K-means and Decision Tree algorithms and examined the performance results. Then he determined a threshold value using the decision tree algorithm and created a centeroids using this threshold value. As a result of the experiments, it was seen that the proposed hybrid algorithm had a higher success rate.

Sugi and Ratna examined the effect of DL and ML algorithms to solve security problems in IoT devices (Sugi and Ratna, 2020) For this, they examined various performance results of K-Nearest Neighbor and Long Short-Term Memory algorithms as a result of training. The metrics they examined at sensitivity, geometric mean, kappa statistic, and detection time. They used the Bot-IoT dataset for improvements.

Haq and Singh calculated the accuracy rate by constructing a hybrid model consisting of a combination of k-means and j48 classification approaches using certain parts of two different datasets whose sum is always equal to the original dataset (Haq and Singh, 2018). Comparative analysis of three techniques such as hybrid, classification, and clustering approach shows that clustering and classification results are stable at one extreme, i.e. lower in classification case and higher in clustering case.

In 2020, researchers created a model of anomalous requests to IoT networks consisting of WSNs (Aysa et al, 2020). The impact of IoT specific features such as node density, power limitations, and insufficient processing power on a botnet has been observed. They used machine learning algorithms such as LSVM, Neural Network, Decision tree, random forest separately and in combination to detect attacks. In the experimental results, they found that the hybrid algorithm created from random forest and decision tree achieved high accuracy to detect attacks.

2.2. Proposed Approach

Information is given about the methods and dataset used while developing the study in this section.

2.2.1. Normalization and Standardization

Normalization and standardization are two important methods used in data preprocessing. Normalization rescales the data between 0 and 1, while Standardization is rescaling the data so that it has the same mean and the same standard deviation. In this study, the Min-Max method, which is one of the most popular methods for normalization, was used. It is applied with Formula 1.

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

The standardization which standardizes the features is carried out with Formula-2.

$$z = \frac{X - u}{s} \quad (2)$$

where X is sample, u is the mean of the training samples, and s is the standard deviation.

2.2.2. Bot-IoT Dataset

It was seen that one of the most popular IoT dataset is Bot-IoT when the literature was searched for the study (Koroniotis et al., 2019). Therefore, the experimental results were made with the Bot-IoT dataset. Bot-IoT is a publicly available dataset developed in 2018 using real network environments in the Cyber Range Lab. The dataset is built in a designed environment using simulated IoT services, network platforms, and feature extraction platform. The source files of the dataset are available in different formats as pcap, argus and csv files. Captured pcap file has approximately 73 million records and is 70 GB in size, while the csv file is approximately 17 GB in size and contains 46 features. The dataset includes 6 types of tools/attacks: DoS, DDoS, Service Scan, OS, Data exfiltration and Keylogging.Also, for the convenience of researchers, 5% of the dataset is allocated as two different files; containing all features and containing 10 features.

<u>*Data Preprocessing;*</u> In this study 5% of the dataset was used with the aim of both seeing the effects of the study on low-dimensional data and making a preliminary study. "*attack*", "*category*", and "*subcategory*" are very important features of the dataset. "*attack*" feature shows whether the data is normal or attack, "*category*" feature shows the normal or attack type such as DoS, DDoS, Reconnaissance or Theft, "*subcategory*" feature shows the normal and traffic subcategory such as *UDP, TCP, Service-Scan, OS-fingerprint, http, keylogging, Data Exfiltration*.

In the study, the classification process was carried out according to the "*category*" feature. Accordingly, the categorical

data were assigned as in Table 1. Table 1 also shows how many data from which attack has.

Table 1. Dataset label and number of data

Category name	Label	# of Data
DDoS	0	1,926,624
DoS	1	1,650,260
Reconnaissance	2	477
Normal	3	91,082
Theft	4	79

It has been seen that it is not important to use the IP and port information on the classification when the literature is studied. In addition, there is no need for the sequence number of the data. It was decided that the "*attack*" and "*subcategory*" labels should not be used during the training in order not to affect the classification and avoid overfitting. Considering all of these information, 10 features in the dataset were extracted and these features are *flgs, proto, state, subcategory, attack, pkSeqID, saddr, sport, daddr, dport*.

As a result of data preprocessing, the number of features in the dataset decreased to 36. After the transformation of the categorical data and feature extraction, MinMaxScaling and StandardScaling processes were applied to the dataset separately. The data obtained as a result of these were evaluated as two separate datasets and the results were examined accordingly.

2.2.3. Algorithms

Machine Learning algorithms used in the study are given in this section.

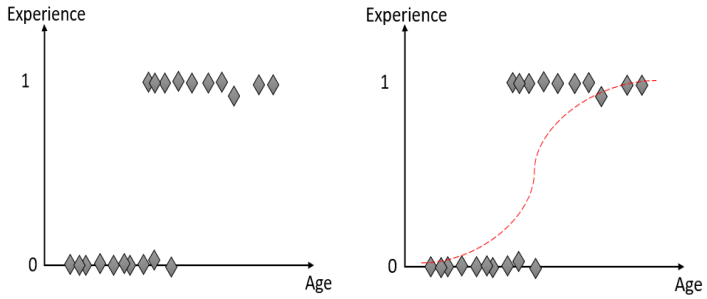
Logistic Regression (LG); Logistic Regression is a method used to group binary or multiclass data when the dependent variable is not continuous. Although it is a regression approach, it is applied to solve classification problems. The goal of logistic regression is to find the most convenient model to describe the relationship between the dependent variable and the independent variable/variables (Wright, 1995; Menard, 2002; Hilbe, 2009; Bayazit et al., 2020). Logistic regression is given by Formula (3).

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_3$ and

$sigmoid = p = \frac{1}{1+e^{-y}}$ with

$$\ln\left(\frac{1}{1-p}\right) = y \tag{3}$$

According to the formula, y is the dependent variable, X's are the independent variable, and β's are the constants. Accordingly, β calculations are made and predictions are made in order to find the relationship between X and Y.



(a) LR-Data example *(b) LR-Data example*

Figure 2. Logistic Regression

Logistic regression can be applied to datasets with the distribution given in Figure 2 (a). According to this figure, there is a relationship between age and experience. The data found in the figure changes as Figure 2 (b) when the formula (1) is applied. And accordingly, it is seen that experience increases with age.

Decision Tree (DT); Decision Tree is a supervised learning algorithm that solves classification problems using entropy. Implements a top-down method with a predefined variable which is called root. It operates to divide the dataset into smaller clusters and these smaller clusters are branches of the tree (Myles et al., 2004; Nowozin et al., 2011). According to this it calculates entropy and information gain each time to determine small clusters and creates branches of the tree.

$$Entropy\ (E) = -\sum p(X) log p(X) \tag{4}$$

$$Information\ Gain = E - \sum p(X) E(V) \tag{5}$$

Formula (4) and Formula (5) show how entropy and information gain are calculated where p is the probability of calculation of X with X being the input data (Karatas et al., 2020). Entropy is calculated to measure the uncertainty associated with the data and information gain is calculated to determine the best split. The feature with the highest information gain before starting the partition is assigned as root.

Stochastic Gradient Descent (SGD); is a machine learning approach that is applied by random selection of the result that will minimize the cost function in optimization problems from the dataset and based on the gradient decrease around the selected data. The algorithm updates both label and the data it is connected to each time. Unlike similar gradient-based algorithms, it makes predictions by drawing zigzags, and is more likely to get stuck in the local minimum (Bottou, 2012; Bottou, 2010; Johnson and Zhang, 2013; Kocyigit et al., 2020).

3. Results and Discussion

In this section, the experiment setup and the results of the study are mentioned. The following settings used to handle the examinations in this work: Experimental results were applied using the Scikit-Learn library with Python programming language on PyCharm Compiler. The study was carried out on a computer indicated by the Table 2.

Table 2. Working Environment

Hardware	Features
CPU	Intel(R) Core(Tm) I7-8700 Cpu @3192Mhz, 6 Cores
Op. Sys.	64 bit, Windows 10
Grap.card	NVIDIA GeForce® GTX 1080 Ti Founders Edition 11G
L1/L2/L3 Cache	384 KB/1.5 MB/12.0 MB
RAM	16.00 GB

The experimental results were examined on the Bot-IoT dataset. For this, the dataset is divided into two as 25% test and 75% training. In order to examine the performance of the study, accuracy, precision, recall, and f1-score values were also calculated and interpreted accordingly. The formulas used to calculate these values are given in (6)-(10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 - Score = \beta * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

$$Error\ Rate = 100 - Accuracy \quad (10)$$

where FN is false negative, TP is true positive, TN is true negative, FP is false positive, and β is a balancing factor. Most accepted use for β is 1 which is the average of Precision and Recall. The results of the three selected algorithms run without any preprocessing are shown in Table 3.

Table 3. Normal

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	52.59	27.66	52.59	36.25
DT	99.99	99.99	99.99	99.99
SGD	44.91	27.66	52.59	36.25

The results obtained after normalization to the three selected algorithms are shown in Table 4.

Table 4. With Normalization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	91.10	91.33	91.10	91.11
DT	100.00	100.00	100.00	100.00
SGD	83.23	83.51	83.22	83.25

The results obtained after the standardization of the three selected algorithms are shown in Table 5.

Table 5. With Standardization

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	99.96	99.96	99.96	99.96
DT	100.00	100.00	100.00	100.00
SGD	98.67	97.98	97.95	97.95

It was seen that the most successful algorithm was DT, and this success did not change whether any preprocessing was done or not when the results were examined. It is seen that the results of the LR and SGD algorithms are quite low before performing a preprocessing. It is seen that the error rate of LR decreased from 47.41% to 8.9%, and the error rate of SGD decreased from 55.09% to 16.77% after normalization. In addition, it is seen that the error rate of LR decreased to 0.04% and the error rate of SGD decreased to 1.33% after standalization. The same performance increase is also seen in other performance metrics. In addition, it was observed that the processing speed of all algorithms increased by 5% after normalization or standardization. It is seen that the standardization increases the performance in detecting attacks on IoT devices when we look at the results. Table 6 shows the accuracy rates of the algorithms in all operating conditions. Table 6 shows the accuracy rates of all results.

Table 6. All results

	Normal Accuracy (%)	Normalization Accuracy (%)	Standardization Accuracy (%)
LR	52.59	91.10	99.96
DT	99.99	100.00	100.00
SGD	44.91	83.23	98.67

4. Conclusions and Recommendations

Developing a secure system for IoT devices is one of the biggest problems of technology. This problem continues to grow as the number of devices connected to the Internet increases. Hardware developed in recent years is efficient both for dealing with big data and using machine learning approaches. Considering these situations, a study was conducted to examine the importance of data preprocessing in providing network security in IoT devices. In this study, popular machine learning approaches such as Logistic Regression, Decision Tree, and Stochastic Gradient Descent algorithms were applied to the Bot-IoT dataset. It has been seen that the results obtained after only numerical transformation in the dataset are low in some algorithms. LR reached 52.59%, DT reached 99.99% and SGD reached 44.91% accuracy rates. The main purpose of the study is to observe the effect of normalization and standardization on increasing the performance rate of the model. Therefore, normalization and standardization processes were performed separately on the Bot-IoT dataset, and then the results were

examined with machine learning algorithms. It is seen that the error rate of LR decreased to 0.04% and the error rate of SGD decreased to 1.33% after standalization. Therefore, the importance of standardization in similar approaches has been showed.

The used dataset is quite large and only 5% of it was examined in this study. In future studies, the entire dataset will be examined with other popular machine learning algorithms and deep learning algorithms, and the error detection rate will be tried to be reduced.

References

Abbasi, F., Naderan, M., & Alavi, S. E. (2021, May). Anomaly detection in Internet of Things using feature selection and classification based on Logistic Regression and Artificial Neural Network on N-BaIoT dataset. In *2021 5th International Conference on Internet of Things and Applications (IoT)* (pp. 1-7). IEEE.

Alshamkhany, M., Alshamkhany, W., Mansour, M., Khan, M., Dhou, S., & Aloul, F. (2020, November). Botnet Attack Detection using Machine Learning. In *2020 14th International Conference on Innovations in Information Technology (IIT)* (pp. 203-208). IEEE.

Aysa, M. H., Ibrahim, A. A., & Mohammed, A. H. (2020, October). IoT ddos attack detection using machine learning. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-7). IEEE.

Bayazit, E. C., Sahingoz, O. K., & Dogan, B. (2020, June). Malware detection in Android systems with traditional machine learning models: a survey. In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-8). IEEE.

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade* (pp. 421-436). Springer, Berlin, Heidelberg.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177-186). Physica-Verlag HD.

Injadat, M., Moubayed, A., & Shami, A. (2020, December). Detecting botnet attacks in IoT environments: an optimized machine learning approach. In *2020 32nd International Conference on Microelectronics (ICM)* (pp. 1-4). IEEE.

Ferrag, M. A., Maglaras, L., Ahmim, A., Derdour, M., & Janicke, H. (2020). Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks. *Future internet*, *12*(3), 44.

Haq, S., & Singh, Y. (2018, December). Botnet detection using machine learning. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 240-245). IEEE.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.

Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, *26*, 315-323.

Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE Access*, *8*, 32150-32162.

Kocyigit, E., Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, December). Real-Time Content-Based Cyber Threat Detection with Machine Learning. In *International Conference on Intelligent Systems Design and Applications* (pp. 1394-1403). Springer, Cham.

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, *100*, 779-796.

Luo, C., Tan, Z., Min, G., Gan, J., Shi, W., & Tian, Z. (2020). A novel web attack detection system for internet of things via ensemble classification. *IEEE Transactions on Industrial Informatics*, *17*(8), 5810-5818.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275-285.

Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., & Kohli, P. (2011, November). Decision tree fields. In *2011 International Conference on Computer Vision* (pp. 1668-1675). IEEE.

Shafiq, M., Tian, Z., Sun, Y., Du, X., & Guizani, M. (2020). Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city. *Future Generation Computer Systems*, *107*, 433-442.

Shukla, P. (2017, September). ML-IDS: A machine learning approach to detect wormhole attacks in Internet of Things. In *2017 Intelligent Systems Conference (IntelliSys)* (pp. 234-240). IEEE.

Sugi, S. S. S., & Ratna, S. R. (2020, December). Investigation of machine learning techniques in intrusion detection system for IoT network. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1164-1167). IEEE.

Wright, R. E. (1995). Logistic regression.