Reconnaissance Attack Detection via Boosting Machine Learning Classifiers

Conference Paper · October 2023

DOI: 10.1063/5.0174730

6 authors, including:

Omar Almomani
The World Islamic Science and Education University
82 PUBLICATIONS **1,383** CITATIONS
SEE PROFILE

Drmohammed Almaayah
King Faisal University
136 PUBLICATIONS **5,108** CITATIONS
SEE PROFILE

Mohammed Madi
Hasan Kalyoncu University
12 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Adeeb Saaidah
The World Islamic Science and Education University
28 PUBLICATIONS **390** CITATIONS
SEE PROFILE

RESEARCH ARTICLE | OCTOBER 20 2023

Reconnaissance attack detection via boosting machine learning classifiers FREE

Omar Almomani ✉; Mohammed Amin Almaiah; Mohammed MADI; Adeeb Alsaaidah; Malek A. Almomani; Sami Smadi

AIP Conf. Proc. 2979, 060002 (2023)
https://doi.org/10.1063/5.0174730

Reconnaissance Attack Detection via Boosting Machine Learning Classifiers

Omar Almomani[1,a)], Mohammed Amin Almaiah[2,b)], Mohammed MADI[3,c)], Adeeb Alsaaidah[1,d)], Malek A. Almomani[4,e)] and Sami Smadi[1,f)]

[1]*Department of Information System and Networks, The World Islamic Sciences and Education University, Amman, Jordan*
[2]*College of Computer Science and Information Technology, King Faisal University, Al-Ahsa 31982, Saudi Arabia*
[3].*Department of Computer Engineering, Hasan Kalyoncu University,Gaziantep, Turkey*
.[4]*Department of Software Engineering, The World Islamic Sciences and Education University University, Amman, Jordan*

[a)] Corresponding author:Omar.almomani@wise.edu.jo
[b)] malmaiah@kfu.edu.sa
[c)] mohammed.madi@hku.edu.tr
[d)] Adeeb.saaidah@wise.edu.jo
[e)] malek.almomani@wise.edu.jo
[f)] sami.smadi@wise.edu.jo

Abstract. With the advancement of Internet technologies, network security concerns are growing exponentially. One of the most difficult issues of network security is keeping it safe. To detect and identify any malicious behavior the network, many security techniques were deployed. Intrusion Detection Systems (IDS) is one of the most frequent strategies for mitigating the effects of these attacks. Reconnaissance is a common attack in computer networks in which the attacker gathers as much information as possible about the target before conducting an attack. Machine Learning (ML) classifiers are commonly used to distinguish between normal and abnormal network traffic. In this paper, Reconnaissance attacks detection is an exam with the following ML classifiers: Adaptive Boosting (AdaBoost), Gradient Boosting, cat Boosting, and eXtreme Gradient Boosting (XGBoost) to determine the most effective classifier in identifying Reconnaissance attacks. Evaluation metrics used are accuracy, precision, F-measure True Positive. The experiment on the UNSW-NB15 dataset shows that the cat Boosting classifier is superior to the XGBoost, AdaBoost and Gradiant Boosting.

INTRODUCTION

The significance of computer networks has risen as central information systems in modern life during the last few years [1] [2][3]. Computer networks [4] [5] [6] have grown in size, application, and architecture, exposing them to a variety of major risks, including malicious activity, network invaders, and network criminals [7] [8]. Combating these harmful network activities is one of the world's top priorities and most important research areas right now. To secure data from unauthorized access, data protection and analysis are required. Because of the vast amount and rapid speed of data, traditional detection methods are unable to detect attacks quickly [9] [10],[11],[12],[13],[14],[15]. Therefore, several approaches have been proposed to protect data and networks from the attacks such as [16],[17],[18],[19],[1], [20],[21], [22]. The IDS is the most promising method of protecting the network from various complex intrusion activities. The IDS differentiates between intrusive and normal network activities by categorizing data into different .groups

Reconnaissance attacks, access attacks, and denial-of-service attacks are all examples of network attacks [23]. An illegal user's effort to find and map network system devices, services available on those devices, and the vulnerabilities of those systems is described as a reconnaissance attack. It's also referred to as data gathering. The malicious intruder will usually start by pinging the victim machine to see which IP is active and responsive. As a result, the intruder may be able to learn what processes or ports are active on the active IP. The intruder searches the application ports using the active IP information to identify the program kind and edition, as well as the kind and edition of the operating system running on the victim machine.

Proceedings of the 4th International Computer Sciences and Informatics Conference (ICSIC 2022)
AIP Conf. Proc. 2979, 060002-1–060002-7; https://doi.org/10.1063/5.0174730
Published by AIP Publishing. 978-0-7354-4703-5/$30.00

IDS has proposed by Anderson J P in 1981 [24]. IDS detects network traffic features and automatically sends a warning when it detects an attack. According to detection methods, IDS is classified as signature-based IDS (SIDS) or anomaly-based IDS (AIDS) [25][26]. SIDS determines whether network traffic is an attack by retrieving records from the signature database of known attacks. AIDS makes a decision. Learning the features of normal network traffic allows one to distinguish between current network traffic and normal network traffic. AIDS can be developed based on statistical knowledge or ML. IDS based on machine learning can detect both known and unknown attacks more effectively [27].

Attacks can be effectively managed and classified using ML classifiers. ML classifiers are the enhancement of a computer's learning process that is based on its own experiences rather than being programmed. This paper investigates the performance of AdaBoost, Gradient Boosting, cat Boosting, and XGBoost ML classifiers, for classifying intrusion. To compare the performance of the ML classifiers, metrics such as accuracy, precision, F-measure, and True Positive, were used.

The following is the rest of this paper: Section 2 goes over the background of AdaBoost, Gradient Boosting, cat Boosting, and XGBoost ML classifiers and IDS, Section 3 goes over the proposed model, evaluation metrics, and dataset used for performance comparison analysis, Section 4 goes over the experimental results, and Section 5 concludes the paper.

BACKGROUND

For tracking and analyzing network traffic for abnormalities, ML classifiers have been used.. There are three kinds of ML anomaly detection approaches supervised, unsupervised, and semi-supervised [28].

Machine Learning classifiers

Several ML classifiers that are being tested for Reconnaissance attack detection, some of these classifiers are:

Gradient Boosting

Gradient boosting [29] new models are generated progressively from a bad model ensemble in order to reduce the loss function with each new model The gradient descent method is used to find this loss function. When the loss function is utilized, each new model fits the observations better, improving total accuracy. Boosting, on the other hand, must be stopped at some point, or otherwise the model will look to be overfit. A prediction accuracy level or a maximum number of models produced might be utilized as a stop condition.

Adaptive Boosting (AdaBoost)

The Adaptive Boosting (Adaboost) Classifier [30] is an ensemble classifier whose main task is to fit a sequence of weak learners to over again changed types of the data, such as models that are marginally better than random guessing. To make the final prediction, all of their predictions are combined using a weighted majority vote or sum.

Cat-Boost

Another machine learning algorithm that predicts class labels accurately is the Cat-Boost classifier. Cat-Boost is an implementation of Gradient Boosting Decision Trees that uses binary decision trees as base predictors. [31],[32].

eXtreme Gradient Boosting (XG Boost)

XGBoost Algorithm produced by P. Deven and N. Khare [22], XGBoost is one of the boosted tree algorithms [16], it follows the concept of gradient boosting [29]. Gradient boosting is a type of supervised machine learning that aims to accurately predict an objective variable by combining the predictions of many poorer models [20]. XGBoost is yet another tree model, a common data mining tool that is fast and efficient. The XGBoost model will compute ten times faster than the Random Forest model. The XGBoost model was built using the additive tree strategy, in which a new tree is added in each stage to supplement the trees that have already been built. As more trees are built, the accuracy improves overall. The conclusion is the weighted sum of each tree's predictions or the best linear combination of all decision-making bodies

Intrusion Detection System

Intrusion is described as harmful behavior that aims to disrupt the security policy of a network by endangering the confidentiality, credibility, or availability of any network device. There are three different forms of IDS. HIDS (host-based intrusion detection systems), NIDS (network-based intrusion detection systems), and hybrid-based intrusion detection systems (HIDS) [25]. The HIDS's goal is to monitor the computer system's internal activities. The goal of the NIDS is to dynamically track network traffic logs in real-time. It intends to do so to detect any potential network intrusion. The detection mechanisms based on an IDS are classified into three types. [25][33]: Intrusion detection systems include tools like misuse detection, anomaly detection, and hybrid IDS. A set of predetermined signatures or criteria used to detect known threats is called as misuse detection. The technology simulates the detection of unknown attackers. It accomplishes this by determining whether the device's current state is normal. Both known and unexpected intrusions are detected by hybrid intrusion detection systems.

PERFORMANCE EVALUATION

Evaluation Model

The objective of the proposed model is to compare performance. of AdaBoost, Gradient Boosting, cat Boosting, and XGBoost ML classifiers for Reconnaissance attacks detection. The following are the steps for constructing the models:

1. *UNSW-NB15* Dataset pre-processing: Null values are removed, and categorical data are converted to numerical form.
2. constructing a classifier model based on training data for AdaBoost, Gradient Boosting, cat Boosting, and XGBoost
3. Based on the testing dataset, build a test classifier model.
4. compute accuracy, precision, F-measure, and true positive
5. Discover the perfect classifier for detecting reconnaissance attacks.

Performance Evaluation Metrics

The following metrics are used to evaluate the proposed model's efficiency level. These metrics are precision, accuracy, F-measure, and True Positive. [34]. The confusion matrix is displayed in Table.1

Table 1. Confusion matrix

		Predicted	
		Normal	Attack
Actual	Normal	TP	FN
	Attack	FP	TN

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$
$$Precision=TP/(TP+FP)$$
$$F\text{-}Measure=(2*Precision*Sensitivity)/(Precision+Sensitivity)$$
$$Sensitivity=TP/(TP+FN)$$
$$True\ Positive=TP/(TP+FN)$$

Dataset

The tcpdump program is used to capture 100 GB of raw network traffic using the IXIA PerfectStorm product (pcap files). To make packet analysis easier, each pcap file is 1000 MB in size. Argus and Bro-IDS approaches were used to create 49 features with the class label, and 12 procedures were completed. There are two sections to this dataset: training and testing. There are 175,341 records in the training set and 82,332 records in the testing set, which can be attack or normal. The nine types of attacks performed against the UNSW-NB15 dataset are analysis, backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, and Worms. As indicated in Table 2, the relevant attacks in the UNSW-NB15 data set are divided into nine kinds.[35].

Table 2. UNSW-NB15 Set attacks.

	Testing Set			Training set		
Category	**Instance Size**	**Ratio in Testing Set (%)**	**Ratio in Attacks**	**Instance Size**	**Ratio in Training Set (%)**	**Ratio in Attacks**
Normal	37,000	44,94	-	56,000	31,94	-
Analysis	677	0,82	1,49	2000	1,14	1,68
Backdoor	583	0,71	1,29	1746	1,00	1,46
DoS	4,089	4,97	9,02	12,264	6,99	10,28
Exploits	11,132	13,52	24,56	33,393	19,04	27,98
Fuzzers	6,062	7,36	13,37	18,184	10,37	15,24
Generic	18,871	22,92	41,63	40,000	22,81	33,52
Reconnaissance	**3,496**	**4,25**	**7,71**	**10,491**	**5,98**	**8,79**
Shellcode	378	0,46	0,83	1,133	0,65	0,95
Worms	44	0,05	0,10	130	0,07	0,11
Total		82,332			175,341	

Table. 3 provides a list of features found in the UNSW-NB15 dataset.

Table 3.UNSW-NB15 Dataset features

Features No	Features Name	Features No	Features Name	Features No	Features Name	Features No	Features Name
1	id	12	dttl	23	dtcpb	34	ct_dst_ltm
2	dur	13	sload	24	dwin	35	ct_src_dport_ltm
3	proto	14	dload	25	tcprtt	36	ct_dst_sport_ltm
4	service	15	sloss	26	synack	37	ct_dst_src_ltm
5	state	16	dloss	27	ackdat	38	is_ftp_login
6	spkts	17	sinpkt	28	smean	39	ct_ftp_cmd
7	dpkts	18	dinpkt	29	dmean	40	ct_flw_http_mthd
8	sbytes	19	sjit	30	trans_depth	41	ct_src_ltm
9	dbytes	20	djit	31	response_body_len	42	ct_srv_dst
10	rate	21	swin	32	ct_srv_src	43	is_sm_ips_ports
11	sttl	22	stcpb	33	ct_state_ttl	44	attack_cat
						45	label

EXPERIMENTAL RESULTS

The selected ML classifiers for detecting Reconnaissance attacks on the UNSW-NB15 dataset are AdaBoost, Gradient Boosting, cat Boosting, and XGBoost. The experiments are carried out using Python on a 3.40 GHz i7 CPU with 6.0 GB RAM. The output of four classifiers is then evaluated using a variety of evaluation matrices, including precision, accuracy, F-measure, and True positive. The obtained results are shown in Table.4.

Table 4. Result

	Accuracy	Precision	F-measure	True Positive
Gradient Boost	98.04	97.33	93.57	90.08
Ada Boost	99.68	99.00	99.00	99.00
Cat Boost	99.80	99.20	99.36	99.53
XG Boost	99.01	98.67	96.81	95.02

In terms of accuracy, Precision, F-measure, and True positive, the results in Table 4 show that the cat boost classifier performs admirably over the other classifiers. Figure 1 depicts the accuracy of various Boosting ML classifiers. Figure 2 depicts the precision of selected Boosting ML classifiers, Figure 3 depicts the F-measure, and Figure 4 depicts the true positive.

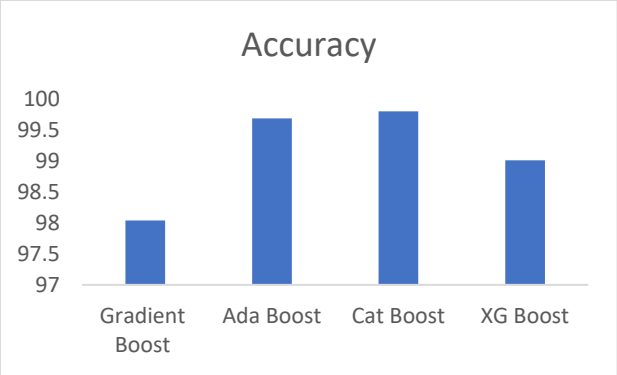

FIGURE 1. Accuracy

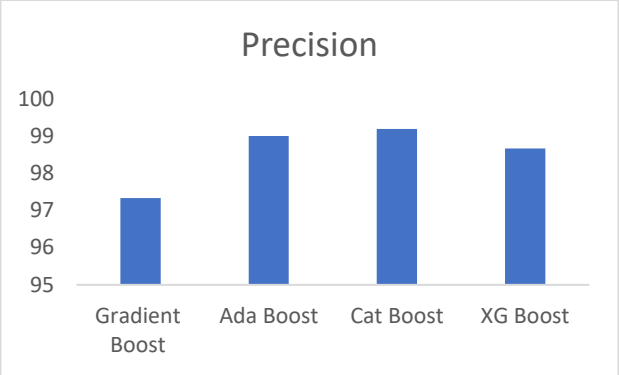

FIGURE 2. Precision

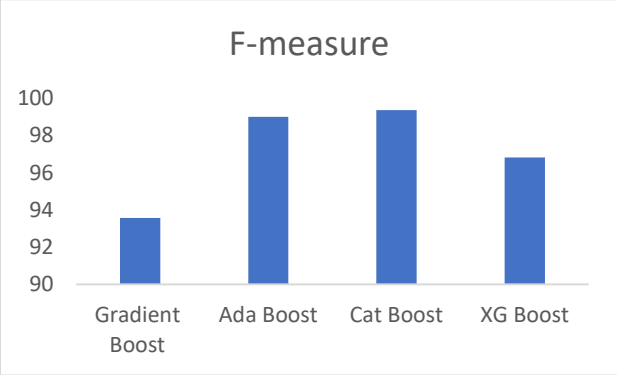

FIGURE 3. F-measure

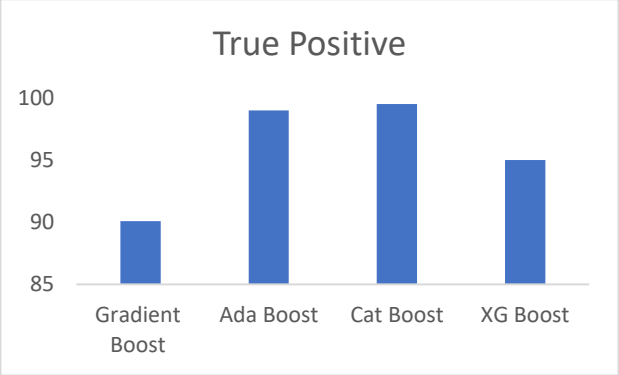

FIGURE 4. True positive

CONCLUSIONS

Reconnaissance attacks have been detected using AdaBoost, Gradient Boosting, cat Boosting, and XGBoost classifiers. These classifiers were evaluated using the UNSW-NB15 dataset. To compare the performance of the classifiers, accuracy, precision, F-measure, and true positive are used. Experiment results show that the cat boost classifier is the best for detecting Reconnaissance attacks in terms of Accuracy, Precision, F-measure, and true positive.

REFERENCES

[1] O. M. D. Al-Momani, "Dynamic redundancy forward error correction mechanism for the enhancement of internet-based video streaming," Universiti Utara Malaysia, 2010.

[2] N. Alsharman, A. Saaidah, O. Almomani, I. Jawarneh, and L. Al-Qaisi, "Pattern Mathematical Model for Fingerprint Security Using Bifurcation Minutiae Extraction and Neural Network Feature Selection," *Secur. Commun. Networks*, vol. 2022, 2022.

[3] M. N. Khan *et al.*, "Improving energy efficiency with content-based adaptive and dynamic scheduling in wireless sensor networks," *IEEE Access*, vol. 8, pp. 176495–176520, 2020.

[4] A. Khalifeh, K. Rajendiran, K. A. Darabkh, A. M. Khasawneh, O. Almomani, and Z. Zinonos, "On the potential of fuzzy logic for solving the challenges of cooperative multi-robotic wireless sensor networks," *Electron.*, vol. 8, no. 12, 2019, doi: 10.3390/electronics8121513.

[5] F. Al Balas, O. Almomani, R. M. A. Jazoh, Y. M. Khamayseh, and A. Saaidah, "An enhanced end to end route discovery in AODV using multi-objectives genetic algorithm," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019, pp. 209–214.

[6] M. H. Qasem, A. Hudaib, N. Obeid, M. A. Almaiah, O. Almomani, and A. Al-Khasawneh, "Multi-agent Systems for Distributed Data Mining Techniques: An Overview," *Big Data Intell. Smart Appl.*, pp. 57–92,

2022.
[7] M. A. Almaiah, A. Al-Zahrani, O. Almomani, and A. K. Alhwaitat, "Classification of cyber security threats on mobile devices and applications," in *Artificial Intelligence and Blockchain for Future Cybersecurity Applications*, Springer, 2021, pp. 107–123.
[8] A. ALMAIAH and O. ALMOMANI, "An Investigator Digital Forensics Frequencies Particle Swarm Optimization For Detection And Classification Of Apt Attack In Fog Computing Environment (IDF-FPSO)," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 07, 2020.
[9] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, 2019.
[10] M. A. Almaiah, Z. Dawahdeh, O. Almomani, A. Alsaaidah, Ahmad Al-Khasawneh, and S. Khawatreh, "A new hybrid text encryption approach over mobile ad hoc network," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, 2020, doi: 10.11591/IJECE.V10I6.PP6461-6471.
[11] S. SMADI, M. ALAUTHMAN, O. ALMOMANI, A. SAAIDAH, and F. ALZOBI, "Application Layer Denial of Services Attack Detection Based on StackNet," *Int. J.*, vol. 3929, no. 3936, pp. 2278–3091, 2020.
[12] A. ALMAIAH and O. ALMOMANI, "AN INVESTIGATION OF DIGITAL FORENSICS FOR SHAMOON ATTACK BEHAVIOUR IN FOG COMPUTING AND THREAT INTELLIGENCE FOR INCIDENT RESPONSE," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 07, 2020.
[13] A. ALMAIAH and O. ALMOMANI, "AN INVESTIGATOR DIGITAL FORENSICS FREQUENCIES PARTICLE SWARM OPTIMIZATION FOR DECTECTION AND CLASSIFICATION OF APT ATTACK IN FOG COMPUTING ENVIROMENT (IDF-FPSO)," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 07, 2020.
[14] A. K. Al Hwaitat *et al.*, "Improved Security Particle Swarm Optimization (PSO) Algorithm to Detect Radio Jamming Attacks in Mobile Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, 2020, doi: 10.14569/IJACSA.2020.0110480.
[15] M. Adil, M. A. Almaiah, A. Omar Alsayed, and O. Almomani, "An Anonymous Channel Categorization Scheme of Edge Nodes to Detect Jamming Attacks in Wireless Sensor Networks," *Sensors*, vol. 20, no. 8, p. 2311, 2020.
[16] F. Albalas, M. Al-Soud, O. Almomani, and A. Almomani, "Security-aware CoAP application layer protocol for the internet of things using elliptic-curve cryptography," *Int. Arab J. Inf. Technol.*, vol. 15, no. 3A Special Issue, 2018.
[17] A. Almomani, M. Alauthman, A. Omar, and A. Firas, "A Proposed Framework for Botnet Spam-email Filtering Using Neucube," in *The International Arab Conference on Information Technology, Yassmine Hammamet, Tunisia*, 2017.
[18] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," *Decis. Support Syst.*, vol. 107, pp. 88–102, 2018.
[19] J. Ababneh and O. Almomani, "Survey of error correction mechanisms for video streaming over the internet," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 3, 2014.
[20] A. Saaidah, O. Almomani, L. Al-Qaisi, and M. K. Madi, "An efficient design of RPL objective function for routing in internet of things using fuzzy logic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, 2019, doi: 10.14569/ijacsa.2019.0100824.
[21] M. A. Almaiah, F. Hajjej, A. Ali, M. F. Pasha, and O. Almomani, "A Novel Hybrid Trustworthy Decentralized Authentication and Data Preservation Model for Digital Healthcare IoT Based CPS," *Sensors*, vol. 22, no. 4, p. 1448, 2022.
[22] A. H. Mohammad, T. Alwada'n, O. Almomani, S. Smadi, and N. ElOmari, "Bio-inspired Hybrid Feature Selection Model for Intrusion Detection," *Comput. Mater. Contin.*, vol. 73, no. 1, pp. 133–150, 2022.
[23] M. Srivastava, "An Introduction to Network Security Attacks," in *Inventive Systems and Control*, Springer, 2021, pp. 505–515.
[24] J. P. Anderson, "Computer security threat monitoring and surveillance," *Tech. Report, James P. Anderson Co.*, 1980.
[25] O. Almomani, "A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms," *Symmetry (Basel).*, vol. 12, no. 6, p. 1046, 2020.
[26] O. Almomani, "A Hybrid Model Using Bio-Inspired Metaheuristic Algorithms for Network Intrusion Detection System," *Comput. Mater. \& Contin.*, vol. 68, no. 1, pp. 409–429, 2021, doi: 10.32604/cmc.2021.016113.
[27] G. Kocher and G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," *Soft Comput.*, vol. 25, no. 15, pp. 9731–9763, 2021.

[28] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and unsupervised learning for data science*. Springer, 2019.
[29] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, p. 21, 2013.
[30] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[31] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, pp. 1189–1232, 2001.
[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
[33] M. A. Ayd\in, A. H. Zaim, and K. G. Ceylan, "A hybrid intrusion detection system design for computer network security," *Comput. \& Electr. Eng.*, vol. 35, no. 3, pp. 517–526, 2009.
[34] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 2015, pp. 1–8.
[35] 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, pp. 1–6.