Cybersecurity Detection Model using Machine Learning Techniques

Mustafa El-Taie [1], Aaras Y.Kraidi [2,*]

[1] Digital Charging Solutions GmbH, Germany

[2] University of Technology and Applied Science, Shinas, Oman

Emails: Mustafa.iessa@gmail.com; aaras.kraidi@shct.edu.om

Abstract

The use of machine learning methods in cybersecurity is only one of many examples of how this once-emerging innovation has entered the mainstream. Anomaly-based identification of common assaults on vital infrastructures is only one instance of the various applications of malware analysis. Scholars are using machine learning-based identification in numerous cybersecurity solutions since signature-based approaches are inadequate at identifying zero-day threats or even modest modifications of established assaults. In this work, we introduce the machine-learning models-based security framework to detect cyber-attacks. This paper used three machine learning models Logistic Regression, Random Forest, and K-Nearest Neighbor This framework not only reduces the computational difficulty of the framework by minimizing the feature parameters, but it also performs well in terms of accuracy in forecasting unknown scenarios in the tests. Finally, we ran trials using cybersecurity datasets to measure the machine learning model's performance using metrics including precision, recall, and accuracy.

Keywords: Machine Learning; Cybersecurity; Cyberattacks; Logistic Regression; K-Nearest Neighbor; Random Forest

1. Introduction

Cybersecurity and defenses versus cyberattacks have been more in need as of late. The widespread adoption of Internet-of-Things (IoT), the meteoric rise of digital infrastructure, and the plethora of appropriate programs employed for various ends by people and organizations account for the bulk of this trend. Massive networks have suffered irreversible harm and monetary damage as a result of cyber assaults like denial-of-service (DoS) assaults, computer viruses, and unauthorized entry[1], [2].

The two main components of any cyber security solution are the network and the computer. An intrusion detection system (IDS) is more effective at protecting a network of machines from outside threats than other systems, like a firewall or encryption software, that is intended to deal with Internet-based cyberattacks. Therefore, an IDS's primary function is to identify and block potential threats to a computer or network. Firewalls and other common approaches fall short of expectations. By constantly keeping tabs on and analyzing what's going on in a network or computer framework, an IDS can keep an eye out for potential security issues like denial-of-service (DoS) attacks. Unauthorized

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

actions, such as gaining entry to the system, making changes, or destroying data, may all be uncovered, determined, and identified with the use of an IDS. Thus, it is necessary to facilitate system security by identifying different sorts of cyber-attacks and abnormalities in a network and by developing an efficient IDS that plays a crucial part in today's network safety[3], [4].

The core tenet of Machine Learning is the concept of programmatically unguided automatic learning from samples and knowledge. Machine learning methods come in many forms, including supervised learning, unsupervised learning, reinforcement learning, and so on. Computer vision, voice recognition, object identification, and content-based multimedia retrieval are just a few of the many applications of machine learning[5], [6]. It is also employed in creating forecasting systems, such as those used to anticipate the stock market. It may be used in systems that suggest content like websites, news stories, TV shows, and products for purchase[7], [8].

Many machine learning solutions in cybersecurity are receiving broad use as of late, as was to be anticipated. The STUXNET and Sony Zero Day assaults, among others, are examples of zero-day threats that may be detected using machine learning. Yet, cybercriminals and virus developers are proceeding at breakneck speed to undermine protections[9], [10]s. The issue is that traditional machine learning methods were developed for static settings, where it is expected that both the training data and the test data come from an identical distribution. The existence of smart and flexible enemies makes it probable that this working hypothesis will be broken. The safety of a system may be compromised by a malevolent adversary if they can change the input data and exploit certain flaws in learning techniques[11]–[13].

The area of this analysis is to develop machine learning models to predict cyber-attacks. The three machine learning models are applied in the cybersecurity dataset as Logistic Regression, K-Nearest Neighbor, and Random Forest.

2. Related Work

While keeping track of and analyzing regular network or computer system activity, an IDS looks for unusual or suspicious behavior indicative of a cyber assault. Numerous studies in the field of cybersecurity have been undertaken to better detect and avoid cyber assaults and intrusions. The authors in [14] presented a novel hybrid machine-learning framework for detecting numerous cyber intrusions. The weighted vote of adaptive boosting is also utilized to integrate several classifiers, and correlation-based choosing features are used to eliminate superfluous data points. They used the network traffic dataset to evaluate their model. The authors in [15] proposed a framework for cybersecurity and they used many datasets in their work. These datasets were also max-min normalized before being classified using standard machine learning techniques including support vector machine (SVM), K-Nearest neighbor (KNN), and Decision Tree (DT) techniques.

The researchers in [16] reported on the creation of a SCADA system testbed for cyber security studies. The testbed was subjected to complex cyberattacks. Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and K-Nearest Neighbour were the five classic machine learning techniques that were taught to identify the assaults. After the algorithms used for machine learning had been trained, they were produced and released into the network, where they were put through further testing using live network data.

The researchers in [17] reviewed multiple studies that used different approaches and strategies to address IoT security issues and spoke about how machine learning algorithms may be used to help. The scope and possibilities for exploitation of the Internet of Things were immediately brought into focus by this research. Then it started zeroing in on specific use cases, such as malware and IDS, and the recognition of unfamiliar IoT devices, in which machine learning may improve IoT security. A new library for machine learning and cyber security was introduced and explored [18]. Evaluating how helpful the resources are. Spearman correlations are often employed to evaluate sequential variable associations, making them an ideal choice for their research.

3. Material and Methods

This section presented the machine learning models used in this paper. It also, shows the steps of the framework applied in this study. Figure 1 shows the steps of the machine learning models.

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

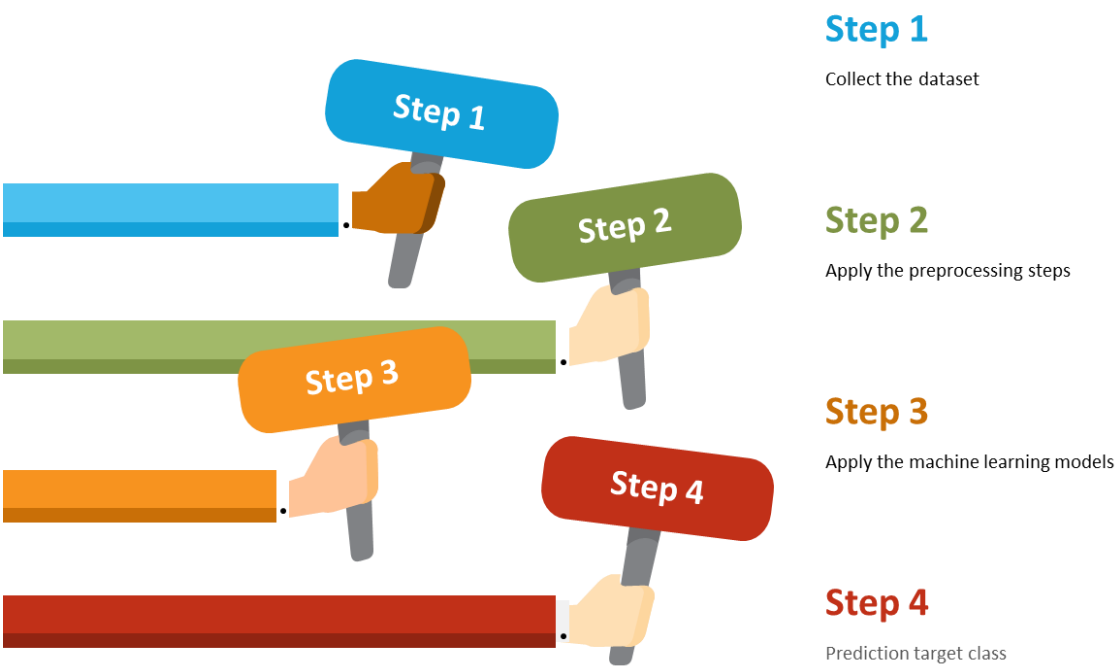

Figure 1: The steps of the proposed model.

3.1 Random Forest

As an established ensemble categorization method, random forest classifiers have found widespread usage in many branches of machine learning and data science. In this technique, "parallel ensembling" is used to simultaneously train several decision tree classification methods on independent subsamples of the data set, with the final result being determined by a vote or an average of the results. By doing so, it improves prediction accuracy and regulates the issue of over-fitting. For this reason, the RF learning framework that incorporates many distinct decision trees often yields superior results over the alternative. It mixes aggregation of bootstraps (bagging) with a random choice of features to construct a succession of decision trees with a managed variability. It works well with both categorized and continuous data and may be used for classification and regression issues[19], [20].

3.2 Logistic Regression

Logistic Regression (LR) is a typical stochastic-based statistical approach employed to address categorization problems in machine learning. To calculate the likelihoods, logistic regression often employs a logistic function, which is also known as the mathematically specified sigmoid. Overfitting is possible with data that is high-dimensional, and it performs best when the data can be linearly partitioned. Over-fitting may be avoided with the use of regularization (L1 and L2) procedures. Logistic Regression has been criticized for its strong reliance on a linear relationship between the dependent and independent variables. It may be employed to address both regression and classification issues, albeit the former is more prevalent[21], [22].

$$f(x) = \frac{1}{1+\exp(-x)}$$

3.3 KNN

Often referred to as a "lazy learning" technique, K-Nearest Neighbors (KNN) is a kind of "instance-based learning" or non-generalizing learning. Instead of concentrating on building a generic internal model, it maintains an n-dimensional storage space for all training data instances. Similarity measurements (such as the Euclidean distance function) are used by KNN to classify fresh data points. The k closest neighbors of each location are used to determine the categorization. The accuracy is data-dependent, however, it is quite tolerant to noisy training data. The most difficult part of using KNN is determining how many neighbors should be taken into account. KNN may be used for both regression and classification[23], [24].

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

4. Results

We applied the three machine learning models to the IDS dataset. We obtained the dataset from Kaggle. The dataset has 42 features. Table 1 shows the sample of the dataset. We made some preprocessing of the dataset such as encoding, and feature selection. For example, in the class target, we have the normal and anomaly class, so we encoded it to 0 or 1. This is a binary classification problem. Then get some descriptive statistics on the dataset as shown in Table 2.

Table 1: The sample of the dataset.

	0	1	2	3	4
Duration	0	0	0	0	0
protocol_type	tcp	udp	tcp	tcp	Tcp
Service	ftp_data	other	private	http	http
Flag	SF	SF	S0	SF	SF
src_bytes	491	146	0	232	199
dst_bytes	0	0	0	8153	420
Land	0	0	0	0	0
wrong_fragment	0	0	0	0	0
Urgent	0	0	0	0	0
Hot	0	0	0	0	0
...	...	...	...	...	...
dst_host_srv_count	25	1	26	255	255
dst_host_same_srv_rate	0.17	0	0.1	1	1
dst_host_diff_srv_rate	0.03	0.6	0.05	0	0
dst_host_same_src_port_rate	0.17	0.88	0	0.03	0
dst_host_srv_diff_host_rate	0	0	0	0.04	0
dst_host_serror_rate	0	0	1	0.03	0
dst_host_srv_serror_rate	0	0	1	0.01	0
dst_host_rerror_rate	0.05	0	0	0	0
dst_host_srv_rerror_rate	0	0	0	0.01	0
Class	normal	normal	anomaly	normal	Normal

Table 2 shows some descriptive statistics on the IDS dataset. The descriptive statistics are count, mean, standard deviation, min, max, 25%,27%, and 50%.

Table 2: The descriptive statistics on the IDS dataset

	count	mean	std	min	25%	50%	75%	max
Duration	25192	305.0541	2.69E+03	0	0	0	0	42862
src_bytes	25192	24330.63	2.41E+06	0	0	44	279	3.82E+08
dst_bytes	25192	3491.847	8.88E+04	0	0	0	530.25	5151385
Land	25192	0.000079	8.91E-03	0	0	0	0	1
wrong_fragment	25192	0.023738	2.60E-01	0	0	0	0	3
Urgent	25192	0.00004	6.30E-03	0	0	0	0	1
Hot	25192	0.198039	2.15E+00	0	0	0	0	77
num_failed_logins	25192	0.001191	4.54E-02	0	0	0	0	4
logged_in	25192	0.394768	4.89E-01	0	0	0	1	1
num_compromised	25192	0.22785	1.04E+01	0	0	0	0	884
root_shell	25192	0.001548	3.93E-02	0	0	0	0	1
su_attempted	25192	0.00135	4.88E-02	0	0	0	0	2
num_root	25192	0.249841	1.15E+01	0	0	0	0	975
num_file_creations	25192	0.014727	5.30E-01	0	0	0	0	40

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

num_shells	25192	0.000357	1.89E-02	0	0	0	0	1
num_access_files	25192	0.004327	9.85E-02	0	0	0	0	8
num_outbound_cmds	25192	0	0.00E+00	0	0	0	0	0
is_host_login	25192	0	0.00E+00	0	0	0	0	0
is_guest_login	25192	0.00913	9.51E-02	0	0	0	0	1
Count	25192	84.59118	1.15E+02	1	2	14	144	511
srv_count	25192	27.69875	7.25E+01	1	2	8	18	511
serror_rate	25192	0.286338	4.47E-01	0	0	0	1	1
srv_serror_rate	25192	0.283762	4.48E-01	0	0	0	1	1
rerror_rate	25192	0.11863	3.19E-01	0	0	0	0	1
srv_rerror_rate	25192	0.12026	3.22E-01	0	0	0	0	1
same_srv_rate	25192	0.660559	4.40E-01	0	0.09	1	1	1
diff_srv_rate	25192	0.062363	1.79E-01	0	0	0	0.06	1
srv_diff_host_rate	25192	0.095931	2.57E-01	0	0	0	0	1
dst_host_count	25192	182.5321	9.90E+01	0	84	255	255	255
dst_host_srv_count	25192	115.063	1.11E+02	0	10	61	255	255
dst_host_same_srv_rate	25192	0.519791	4.49E-01	0	0.05	0.51	1	1
dst_host_diff_srv_rate	25192	0.082539	1.87E-01	0	0	0.03	0.07	1
dst_host_same_src_port_rate	25192	0.147453	3.08E-01	0	0	0	0.06	1
dst_host_srv_diff_host_rate	25192	0.031844	1.11E-01	0	0	0	0.02	1
dst_host_serror_rate	25192	0.2858	4.45E-01	0	0	0	1	1
dst_host_srv_serror_rate	25192	0.279846	4.46E-01	0	0	0	1	1
dst_host_rerror_rate	25192	0.1178	3.06E-01	0	0	0	0	1
dst_host_srv_rerror_rate	25192	0.118769	3.17E-01	0	0	0	0	1

Figure 2 shows the heatmap in the dataset. The heatmap shows the correlation between variables and target variables. So, the heatmap aims to delete the less correlated variables in the IDS dataset. So, we applied the feature selection process to delete the less weight of the variable. The fewer variables have negative effects on the output of the models. So, we used the largest correlation to give the largest accuracy.

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

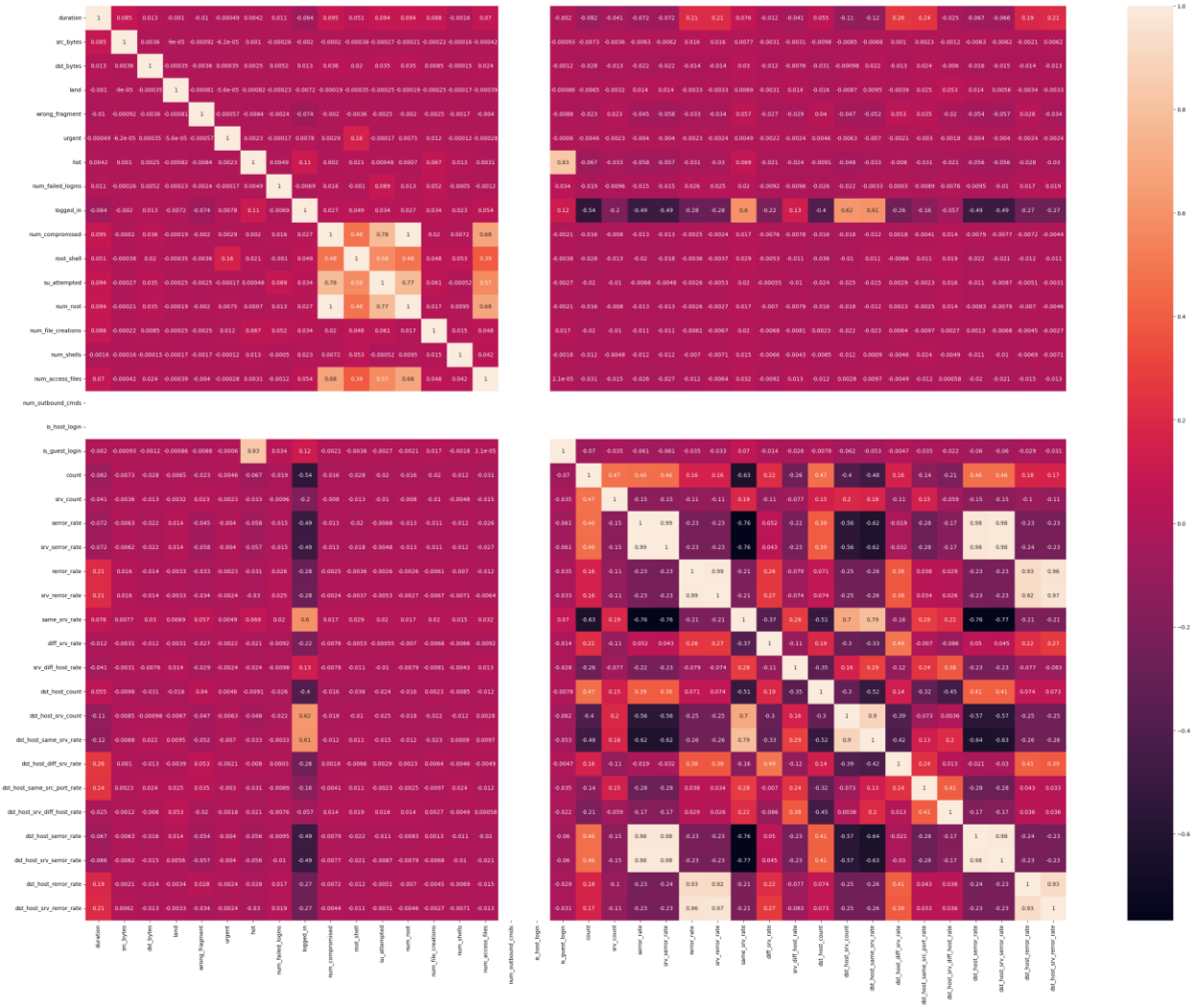

Figure 2: The heatmap of the IDS dataset.

Then we applied the three machine learning models to the IDS dataset. First Figure 2 shows the number of cases of the target class. The goal of this study, detect whether the case is normal or an anomaly. From Figure 3. The size of data in the normal is largest than the size of the anomaly.

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

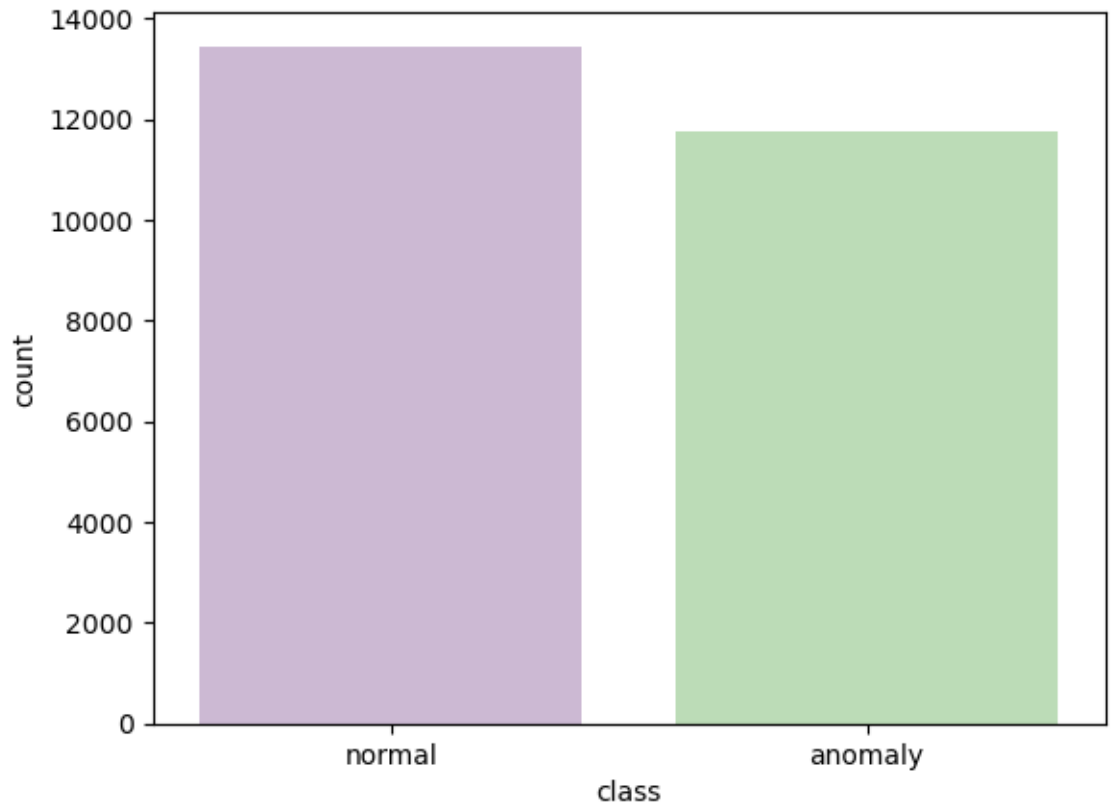

Figure 3: The size of the normal and anomaly class target.

Precision, recall, f1 score, and overall accuracy are the metrics we use to assess the performance of our model cyber security system are defined as:

$$Acc = \frac{A+B}{A+B+C+D} \quad (1)$$

$$F1\ score = 2 * \frac{precision*Recall}{Precision+Recall} \quad (1)$$

$$Recall = \frac{A}{A+D} \quad (1)$$

$$Precision = \frac{A}{A+C} \quad (1)$$

Where A, B, C, D refers to True Positive, True Negative, False Positive, and False Negative

The three models LR, RF, and KNN are applied in the IDS dataset. We compute the accuracy, precision, precision, and f score. Table 3 shows the output of three models. From Table 3, The Random forest is the largest accuracy followed by KNN, then Logistic Regression. Also, the Random forest is largest in precision followed by KNN, then Logistic Regression.

Table 3: The results of three models.

	Accuracy	Precision	Recall	F1 Score
Logistic Regression	94.9%	94.1%	94.9%	95.2%
K-Nearest Neighbor	98.8%	98.5%	98.8%	98.9%

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

Random Forest	**99.81%**	**99.80.%**	**99.81%**	**99.82%**

5. Conclusion

Malware detection, intrusion detection, and security issues like those in power systems, industrial control systems, and so on all benefit from the use of machine learning. Solving them requires training and classifying massive amounts of data efficiently and effectively. A key developing worry is the availability of hostile attackers who may circumvent such technologies by manipulating the classifiers. Three machine learning models are applied in the IDS dataset. We used the Logistic Regression, Random Forest, and KNN models. The pressing process is conducted in this research such as encoding the dataset into 0 and 1, and the feature selection. The feature selection is applied to a dataset to use the largest importance variable in the training process. After that, the random forest has the largest accuracy, followed by the KNN, then the Logistic Regression.

References

[1] A. Handa, A. Sharma, and S. K. Shukla, "Machine learning in cybersecurity: A review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1306, 2019.

[2] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[3] J. B. Fraley and J. Cannady, "The promise of machine learning in cybersecurity," in *SoutheastCon 2017*, IEEE, 2017, pp. 1–6.

[4] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J. Big data*, vol. 7, pp. 1–29, 2020.

[5] P. Dasgupta and J. Collins, "A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks," *AI Mag.*, vol. 40, no. 2, pp. 31–43, 2019.

[6] Y. Miao, C. Chen, L. Pan, Q.-L. Han, J. Zhang, and Y. Xiang, "Machine learning–based cyber attacks targeting on controlled information: A survey," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–36, 2021.

[7] I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. I. Khan, "Intrudtree: a machine learning based cyber security intrusion detection model," *Symmetry (Basel).*, vol. 12, no. 5, p. 754, 2020.

[8] V. Ford and A. Siraj, "Applications of machine learning in cyber security," in *Proceedings of the 27th international conference on computer applications in industry and engineering*, IEEE Xplore Kota Kinabalu, Malaysia, 2014.

[9] R. Prasad, V. Rohokale, R. Prasad, and V. Rohokale, "Artificial intelligence and machine learning in cyber security," *Cyber Secur. lifeline Inf. Commun. Technol.*, pp. 231–247, 2020.

[10] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security," in *2018 10th international conference on cyber Conflict (CyCon)*, IEEE, 2018, pp. 371–390.

[11] K. Shaukat *et al.*, "Performance comparison and current challenges of using machine learning techniques in cybersecurity," *Energies*, vol. 13, no. 10, p. 2509, 2020.

[12] R. Das and T. H. Morris, "Machine learning and cyber security," in *2017 international conference on computer, electrical & communication engineering (ICCECE)*, IEEE, 2017, pp. 1–7.

[13] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan, "Machine learning and deep learning techniques for cybersecurity: a review," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, Springer, 2020, pp. 50–57.

[14] P. Sornsuwit and S. Jaiyen, "A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting," *Appl. Artif. Intell.*, vol. 33, no. 5, pp. 462–482, 2019.

[15] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection:

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023

Datasets and comparative study," *Comput. Networks*, vol. 188, p. 107840, 2021.

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

[17] S. Strecker, W. Van Haaften, and R. Dave, "An analysis of IoT cyber security driven by machine learning," in *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2021*, Springer, 2021, pp. 725–753.

[18] R. A. Calix, S. B. Singh, T. Chen, D. Zhang, and M. Tu, "Cyber security tool kit (CyberSecTK): A Python library for machine learning and cyber security," *Information*, vol. 11, no. 2, p. 100, 2020.

[19] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019.

[20] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, p. e5518, 2018.

[21] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019.

[22] D. Tien Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, pp. 361–378, 2016.

[23] R. Goyal, P. Chandra, and Y. Singh, "Suitability of KNN regression in the development of interaction based software fault prediction models," *Ieri Procedia*, vol. 6, pp. 15–21, 2014.

[24] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.

Doi : https://doi.org/10.54216/JCIM.120104
Received: December 18, 2022 Revised: February 19, 2023 Accepted: April 25, 2023