Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Investigation of The Effect of Data Normalization on Classification and Feature Selection in Intrusion Detection System

Mesut Polatgil[1]
mesutpolatgil@cumhuriyet.edu.tr
[1]Sivas Cumhuriyet Üniversitesi

Article Information

Submitted : 20 Feb 2022
Reviewed : 1 Mar 2022
Accepted : 15 Apr 2022

Keywords

Intrusion Detection System
machine learning
feature selection
data scaling
data normalization

Abstract

The increasing use of the internet and the developments in the world of informatics have brought security problems together. Hardware and software known as Intrusion Detection System aim to detect attacks from the outside world and protect the system from them. These systems need to be fast and intelligent. Establishing intelligent systems for IDS requires data collection, processing and establishment of models in this area. It is very important to pre-process the collected data and select the necessary attributes. The fact that there are many feature selection methods and data preprocessing steps raises the question of which of these should be used and even which of these combinations of options would be better. Although there are studies on selecting the required features or on different normalization methods, there is no study that applies them together on IDS systems. This study was carried out for this purpose. With the study performed on the Kddcup99 dataset, 4 different normalization and 4 different feature selection methods were evaluated together. In this context, 20 different datasets were created and K nearest neighbor, Artificial Neural Networks and Random Forest algorithms were applied to each of them. When the results obtained are evaluated together with the normalization methods and feature selection methods, it has been shown that ideal features that produce successful results for IDS can be found.

A. Introduction

In We Are Social Digital's 2021 report, it is stated that 59% of the world's population has 4.66 billion internet users, 66% of the world's population has 5.22 billion mobile users. These numbers are increasing day by day [1]. Developments such as widespread use of mobile technologies, easier access to the internet, and cloud-based delivery of many technologies have increased the use of computer networks and internet. It created opportunities for malicious people such as hackers in widely used network and internet technologies. By taking advantage of security vulnerabilities, they started to have significant effects, especially intrusion and damage to systems. To prevent such harmful effects, intrusion detection systems have been developed and important steps have been taken to prevent attacks [2].

Intrusion detection systems (IDS) are systems that detect and prevent actions taken with the aim of damaging a personal computer or computer network. These systems use signature-based and anomaly-based approaches. In the signature-based approach, the IDS has a signature that identifies each attack, revealing its characteristics. IDS detects the information obtained from the previous attack patterns in a database and compares the new attack with this database. The disadvantage of this system is that if the new attack has not occurred and recorded before, the attack may not be detected. In the anomalous-based approach, which is the other approach, attack detection is made through abnormal movements on the network. The most important advantage of this approach is that it can detect attacks that have not been carried out before [3].

Machine learning, which works by finding patterns in data using statistics and computer power, has been successfully applied in different disciplines in recent years [4–8]. In addition to signature-based and abnormal-based approaches, machine learning methods can also be used successfully for IDS. Reference [9] proposed a three-stage machine learning method on the kddcup99 dataset. ExtraTrees Classfier achieved a high success rate of 99.75% in the multi-classification problem in which it included the feature selection method. Reference [10] proposed an attribute selection method using the pigeon inspired optimizer method. In his study on the Kddcup99 dataset, he stated that it performed better than other feature selection methods. Reference [11] proposed a model called multilevel semi-supervised ML (MSML). With the method he tried on the Kddcup99 dataset, he showed that he found more successful results than other IDS systems. Reference [12] suggested a two-stage hybrid methodology. KDDCup99 achieved a high classification success of 99.9% in multiclassification for the Kddcup99 dataset in its study on NSL-KDD and UNSW-NB15 datasets. Reference [13] stated that he achieved a high success rate on the Kddcup99 dataset with the feature selection method based on the CART algorithm. Reference [14] developed a new ensemble learning model and achieved high classification success on KDD, KDD99, and UNSW-NB-15 datasets. Reference [15] has increased the success of multiple classification for IDS by proposing a three-stage hybrid method. It preprocessed with the min-max method, selected the feature with the random forest recursive feature elimination method, and classified it with the SVM and ANFIS methods. Reference [16] has increased the success of the algorithm by preventing the Boruta feature selection method from entering an infinite loop.

Reference [17] provided a performance increase on the Kddcup99 dataset by applying the correlation-based feature selection method.

In addition, there are studies in the literature examining the effects of feature selection methods [18–21] and normalization methods on machine learning success [22–25]. However, there is no study evaluating feature selection methods and normalization methods together. This study was carried out to examine the effect of normalization methods on feature selection methods and the increase in success for the IDS system.

This study was carried out to examine the effect of normalization processes on feature selection and classification success in machine learning algorithms that can be used for IDS.

This study is unique in terms of examining the classification success for IDS by evaluating feature selection and normalization methods together.

In this study, while evaluating normalization methods and feature selection methods, selection was made based on the Scikit-Learn library.

B. Method

Kddcup99 data set was used in the study. This dataset has 41 features and approximately 494 thousand records in total. The dataset was coded according to whether there was an attack or not, and the text values were converted into numeric values. Of these records, 97278 shows non-attack records, and the remaining 396743 shows the 22 types of attacks mentioned. The imbalance situation in the dataset has been resolved with the "RandomUnderSampler" method in the Scikit-Learn library. Thus, it is ensured that there are classes with and without an equal number of attacks [14]. A total of five different datasets were created by recording the original form of the dataset and its scaled shapes with four different normalizations separately. Chi-square, Forward selection (FS), Backward Selection (BS) and Model-based (Logistic Regression-LR) feature selection methods were applied to each dataset and its success on the intrusion detection system was examined. Among the machine learning algorithms, K nearest neighbor (KNN), Artificial Neural Networks (MLP) and Random Forest (RF) algorithms were applied. The selected normalization methods and machine learning algorithms are preferred considering their availability in the Scikit-learn library [26].

The processes performed in the study are visualized in Figure 1.

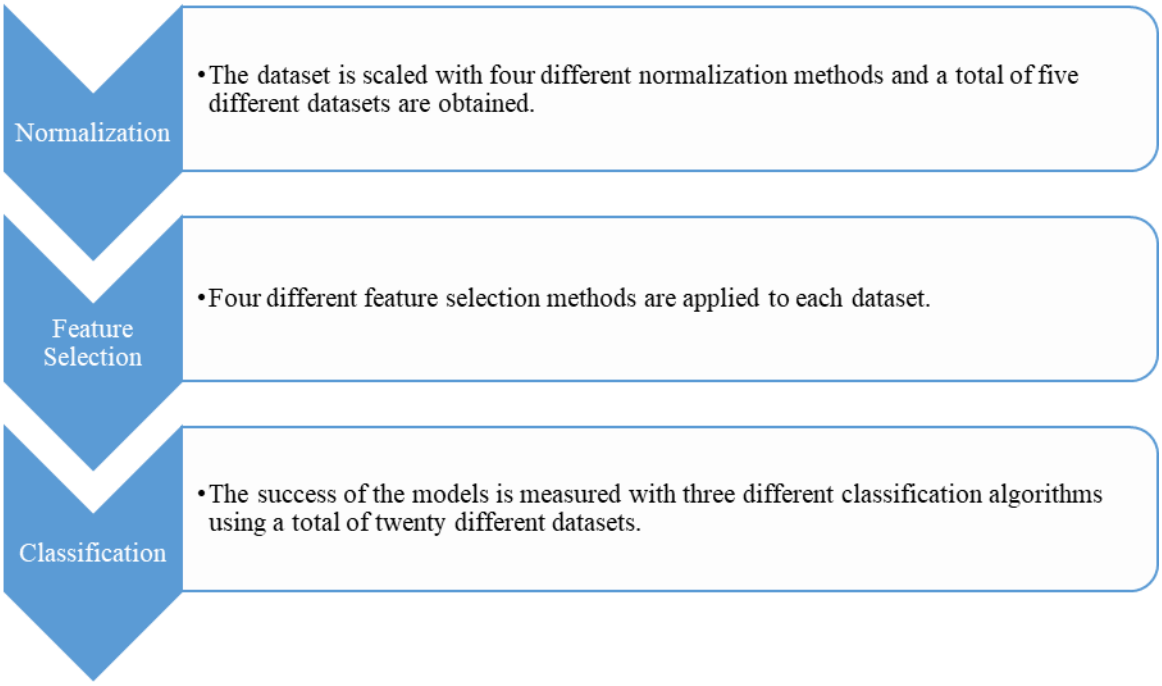

Figure 1. Process of Method.

C. Findings

Table 1. Original Data With KNN

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9992	0.9992	0.9992	0.9992	0.9992
chi-square	6	0.9979	0.992	0.9965	0.9979	0.9979
FS	5	0.9928	0.9969	0.9886	0.9927	0.9928
BS	5	0.9884	0.9970	0.9818	0.9893	0.9894
LR	5	0.9934	0.9975	0.9893	0.9934	0.9934

Table 2. Original Data With MLP

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9953	0.9994	0.9912	0.9953	0.9953
chi-square	6	0.9948	0.9971	0.9925	0.9948	0.9948
FS	5	0.9921	0.9961	0.9881	0.9921	0.9921
BS	5	0.9911	0.9988	0.9832	0.9910	0.9910
LR	5	0.9932	0.9984	0.9878	0.9931	0.9931

Table 3. Original Data With RF

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.991	0.9997	0.9822	0.9909	0.9910
Chi-square	6	0.9890	1	0.9781	0.9889	0.9890
FS	5	0.9860	0.9932	0.9790	0.9860	0.9861
BS	5	0.9862	0.9931	0.9790	0.9860	0.9862
LR	5	0.9867	0.9993	0.9743	0.9866	0.9868

When Table 1-3 data is examined, the highest success rate in the analysis made with the original datasets was found with the KNN algorithm with a rate of 9992%. Although there was a serious decrease in the number of features, it was observed that there was no significant decrease in classification success. In feature selection algorithms, the closest value to the original dataset was found by the chi-square feature selection method.

Table 4. Original Data Features

Method	**Features**
Chi-square	'duration', 'src_bytes', 'dst_bytes', 'count', 'srv_count', 'dst_host_count'
FS	'wrong_fragment', 'hot', 'count', 'same_srv_rate', 'diff_srv_rate'
BS	‘wrong_fragment‘,‘count‘,‘serror_rate‘,‘same_srv_rate‘,‘diff_srv_rate‘
LR	'service', 'flag', 'count', 'srv_count', 'dst_host_srv_count'

When the features selected in Table 4 are examined, it is seen that each method chooses different features.

The results obtained when the min-max normalization method is applied to the dataset and then feature selection and classification is performed are given in Table 5-8.

Table 5. Min-Max Data With KNN

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9989	0.9993	0.9984	0.9989	0.9989
Chi-square	6	0.9933	0.9987	0.9880	0.9933	0.9933
FS	5	0.9930	0.9977	0.9883	0.9930	0.9930
BS	5	0.9921	0.9991	0.9850	0.9920	0.9920
LR	15	0.9986	0.9989	0.9983	0.9986	0.9986

Table 6. Minmax Data With MLP

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9987	0.9993	0.9981	0.9987	0.9987
Chi-square	6	0.9894	0.9934	0.9856	0.9894	0.9894
FS	5	0.9897	0.9977	0.9817	0.9896	0.9897
BS	5	0.9907	0.9973	0.9841	0.9907	0.9907
LR	15	0.9985	0.9985	0.9984	0.9985	0.9985

Table 7.Minmax Data With RF

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9918	1	0.9837	0.9918	0.9918
Chi-square	6	0.9847	0.9926	0.9767	0.9846	0.9847
FS	5	0.9846	0.9916	0.9776	0.9846	0.9847
BS	5	0.9863	0.9922	0.9803	0.9862	0.9863
LR	15	0.9904	0.9995	0.9814	0.9904	0.9905

When Table 5-7 is examined, the highest success was obtained with KNN with the data using the min-max normalization method. Although the number of features decreased drastically, the classification performance remained almost the same.

Table 8.Minmax Data Features

Method	Features

Chi-square	'protocol_type', 'logged_in', 'count', 'srv_count', 'dst_host_same_src_port_rate', 'dst_host_srv_serror_rate'
FS	'wrong_fragment', 'hot', 'count', 'same_srv_rate', 'diff_srv_rate'
BS	‘wrong_fragment‘,‘count‘,‘serror_rate‘,‘same_srv_rate‘,‘diff_srv_rate‘
LR	'protocol_type', 'flag', 'src_bytes', 'wrong_fragment', 'hot', 'is_guest_login', 'count', 'srv_serror_rate', 'srv_rerror_rate', 'same_srv_rate', 'diff_srv_rate', 'dst_host_count', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'dst_host_srv_serror_rate'

When Table 8 is examined, it has been determined that different feature selection methods select different features.

The results obtained when the Standard normalization method is applied to the dataset and then feature selection and classification is performed are given in Table 9-12.

Table 9. Standart Scaler Data With KNN

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9991	0.9992	0.9990	0.9991	0.9991
Chi-square	6					
FS	5	0.9938	0.9975	0.9898	0.9937	0.9937
BS	5	0.9909	0.9989	0.9829	0.9908	0.9909
LR	15	0.9987	0.9992	0.9982	0.9987	0.9987

Table 10. Standart Scaler Data With MLP

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9987	0.9987	0.9988	0.9987	0.9987
Chi-square	6					
FS	5	0.9922	0.9948	0.9893	0.9921	0.9921
BS	5	0.9910	0.9990	0.9830	0.9909	0.9910
LR	15	0.9982	0.9989	0.9975	0.9982	0.9982

Table 11. Standart Scaler Data With RF

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9907	0.9998	0.9815	0.9906	0.9907
Chi-square	6					
FS	5	0.9853	0.9928	0.9775	0.9851	0.9853
BS	5	0.9865	0.9937	0.9793	0.9864	0.9865
LR	15	0.9906	0.9972	0.9839	0.9905	0.9906

When Table 9-11 is examined, it is seen that the highest classification success is obtained with KNN and although there is a serious decrease in the number of features, the classification success has not changed much.

Table 12. Standart Scaler Data Features

Method	Features
Chi-square	
FS	'wrong_fragment', 'hot', 'count', 'same_srv_rate', 'diff_srv_rate'
BS	‘wrong_fragment‘,‘count‘,‘serror_rate‘,‘same_srv_rate‘,‘diff_srv_rate‘
LR	'protocol_type', 'wrong_fragment', 'hot', 'logged_in', 'lnum_compromised', 'lnum_root', 'is_guest_login', 'count', 'srv_serror_rate', 'srv_rerror_rate', 'same_srv_rate', 'dst_host_count', 'dst_host_srv_count', 'dst_host_same_src_port_rate', 'dst_host_srv_serror_rate'

According to Table 12, different feature selection methods included different features in the model.

The results obtained when the Robust normalization method is applied to the dataset and then feature selection and classification is performed are given in Table 13-15.

Table 13. Robust Scaler Data With KNN

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9987	0.9988	0.9985	0.9987	0.9987
Chi-square	6					
FS	5	0.9937	0.9978	0.9895	0.9936	0.9937
BS	5	0.9914	0.9988	0.9841	0.9914	0.9914
LR	8	0.9964	0.9982	0.9947	0.9964	0.9964

Table 14. Robust Scaler Data With MLP

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9989	0.9993	0.9985	0.9989	0.9989
Chi-square	6					
FS	5	0.9936	0.9990	0.9880	0.9935	0.9935
BS	5	0.9914	0.9989	0.9840	0.9914	0.9914
LR	8	0.9951	0.9987	0.9916	0.9951	0.9951

Table 15. Robust Scaler Data With RF

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9911	0.9994	0.9822	0.9909	0.9910
Chi-square	6					
FS	5	0.9848	0.9923	0.9770	0.9846	0.9848
BS	5	0.9861	0.9934	0.9788	0.9861	0.9861
LR	8	0.9881	0.9992	0.9769	0.9880	0.9881

When Table 13-15 is examined, it is seen that the highest classification success is obtained with MLP, unlike other methods, and although there is a significant decrease in the number of features, the classification success has not changed much.

Table 16. Robust Scaler Data Features

Method	Features
Chi-square	
FS	'wrong_fragment', 'hot', 'count', 'same_srv_rate', 'diff_srv_rate'
BS	‘wrong_fragment‘,‘count‘,‘serror_rate‘,‘same_srv_rate‘,‘diff_srv_rate‘
LR	'protocol_type', 'service', 'flag', 'logged_in', 'count', 'srv_count', 'dst_host_count', 'dst_host_same_src_port_rate'

According to Table 16, different feature selection methods included different features in the model.

The results obtained when Maxabs normalization method is applied to the dataset and then feature selection and classification is performed are given in Table 17-19.

Table 17. Maxabs Scaler Data With KNN

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9992	0.9995	0.9990	0.9992	0.9992
Chi-square	6	0,9987	0.9991	0.9983	0.9987	0.9987
FS	5	0.9932	0.9968	0.9895	0.9931	0.9932
BS	5	0.9911	0.9990	0.9832	0.9910	0.9911
LR	15	0.9988	0.9992	0.9983	0.9988	0.9988

Table 18. Maxabs Scaler Data With MLP

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9949	0.9993	0.9903	0.9948	0.9948
Chi-square	6	0.9936	0.9988	0.9884	0.9936	0.9936
FS	5	0.9931	0.9973	0.9890	0.9931	0.9931
BS	5	0.9911	0.9991	0.9832	0.9911	0.9911
LR	15	0.9976	0.9991	0.9961	0.9976	0.9976

Table 19. Maxabs Scaler Data With RF

Method	FC	ACC	PRE	RECALL	F1	AUC
--	41	0.9907	0.9994	0.9814	0.9906	0.9907
Chi-square	6	0.9890	0.9999	0.9780	0.9888	0.9890
FS	5	0.9860	0.9924	0.9795	0.9859	0.9860
BS	5	0.9856	0.9928	0.9785	0.9856	0.9856
LR	15	0.9906	0.9996	0.9815	0.9905	0.9906

When Table 17-19 is examined, it is seen that the highest classification success is obtained with KNN and although there is a serious decrease in the number of features, the classification success has not changed much.

Table 20. Maxabs Scaler Data Features

Method	Features
Chi-square	'duration', 'src_bytes', 'dst_bytes', 'count', 'srv_count', 'dst_host_count'
FS	'wrong_fragment', 'hot', 'count', 'same_srv_rate', 'diff_srv_rate'
BS	‘wrong_fragment‘,‘count‘,‘serror_rate‘,‘same_srv_rate‘,‘diff_srv_rate‘
LR	'protocol_type', 'flag', 'src_bytes', 'wrong_fragment', 'hot', 'is_guest_login', 'count', 'srv_serror_rate', 'srv_rerror_rate', 'same_srv_rate', 'diff_srv_rate', 'dst_host_count', 'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate', 'dst_host_srv_serror_rate'

D. Discussion

Data preprocessing steps and feature selection methods are important processing steps on data in the field of machine learning. The fact that there are different normlization methods and feature selection methods requires knowing which of these methods should be used. Although many studies have investigated the effect of different feature selection and normalization methods on classification success, it is unclear in the literature how the results will change when these two stages are evaluated together. This study was carried out with the aim of answering this question. In order to investigate this situation, studies have been carried out on intrusion detection systems, which is an important field of study in computer science. Because it is very important to detect the attacks that can be made on the systems at a high rate and quickly if there are no similar attacks before. In this context, five different datasets were created using different normalization and feature selection methods on the Kddcup99 dataset, and

success criteria were calculated by applying KNN, MLP and RF algorithms to each dataset.

As a result of the study, chi-square, FS and BS methods selected the same number of features in both the original dataset and the normalized dataset, but by choosing different features, they achieved similar classification success. Despite choosing different features, the similar classification successes show that these methods can detect features well. In the literature, there are studies in which different learning achievements were obtained by using different attributes. Reference [10] achieved a success rate of 94.7% with 10 features with the pigeon inspired optimizer feature selection method. Reference [11] reached a success rate of 96.6% for a multi-class problem situation with a model called multilevel semi-supervised ML (MSML) that he proposed. Reference [12] achieved a success rate of 99.9% for the Kddcup99 dataset in his study where he proposed two-stage hybrid methodology. Reference [13] achieved an overall success rate of 96.9% on the Kddcup99 dataset with the feature selection method based on the CART algorithm.

In addition, in the model-based feature selection method, it is seen that different numbers and different features are selected in different normalization methods. 5 features were selected in the original dataset, 8 features in the robust normalization dataset, and 15 features in the standard, min-max and robust normalization methods. It has been determined that these features show almost the same classification success when compared with the original dataset. In other words, even if the model uses 5, 8 or 15 features instead of 41 features, it can make almost the same classification. This will ensure that artificial intelligence models to be established for intrusion detection systems will be advantageous in terms of speed and performance.

E. Conclusion

In this study, the effects of different feature selection and normalization methods for intrusion detection systems on classification were examined together. The results show that instead of using too many features for IDS, similar classification success can be achieved with fewer features and attacks can be detected quickly and effectively.

F. Reference

[1] S. Kemp, Digital 2021 October Global Statshot Report — DataReportal – Global Digital Insights, (2021). https://datareportal.com/reports/digital-2021-october-global-statshot (accessed December 14, 2021).

[2] R.A. Kemmerer, G. Vigna, Intrusion Detection: A Brief History and Overview (Supplement to Computer Magazine), Computer. 35 (2002) 27–30. https://doi.org/10.1109/MC.2002.10036.

[3] H.J. Liao, C.H. Richard Lin, Y.C. Lin, K.Y. Tung, Intrusion detection system: A comprehensive review, Journal of Network and Computer Applications. 36 (2013) 16–24. https://doi.org/10.1016/J.JNCA.2012.09.004.

[4] A.I. Kadhim, Survey on supervised machine learning techniques for automatic text classification, Artificial Intelligence Review. 52 (2019) 273–292. https://doi.org/10.1007/S10462-018-09677-1.

[5] Y. Peng, A novel ensemble machine learning for robust microarray data classification, Computers in Biology and Medicine. 36 (2006) 553–573. https://doi.org/10.1016/J.COMPBIOMED.2005.04.001.

[6] N. Namdev, S. Agrawal, S. Silkari, Recent advancement in machine learning based internet traffic classification, Procedia Computer Science. 60 (2015) 784–791. https://doi.org/10.1016/J.PROCS.2015.08.238.

[7] A.E. Maxwell, T.A. Warner, F. Fang, Implementation of machine-learning classification in remote sensing: An applied review, International Journal of Remote Sensing. 39 (2018) 2784–2817. https://doi.org/10.1080/01431161.2018.1433343.

[8] K. Miettinen, M. Juhola, Classification of otoneurological cases according to bayesian probabilistic models, Journal of Medical Systems. 34 (2010) 119–130. https://doi.org/10.1007/s10916-008-9223-z.

[9] J. Sharma, C. Giri, O.C. Granmo, M. Goodwin, Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation, Eurasip Journal on Information Security. 2019 (2019). https://doi.org/10.1186/S13635-019-0098-Y.

[10] H. Alazzam, A. Sharieh, K.E. Sabri, A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer, Expert Systems with Applications. 148 (2020). https://doi.org/10.1016/J.ESWA.2020.113249.

[11] H. Yao, D. Fu, P. Zhang, M. Li, Y. Liu, MSML: A novel multilevel semi-supervised machine learning framework for intrusion detection system, IEEE Internet of Things Journal. 6 (2019) 1949–1959. https://doi.org/10.1109/JIOT.2018.2873125.

[12] K. Narayana Rao, K. Venkata Rao, P.R. Prasad, A hybrid Intrusion Detection System based on Sparse autoencoder and Deep Neural Network, Computer Communications. 180 (2021) 77–88. https://doi.org/10.1016/J.COMCOM.2021.08.026.

[13] N. Kumar, U. Kumar, Diverse Analysis of Data Mining and Machine Learning Algorithms to Secure Computer Network, Wireless Personal Communications. (2021). https://doi.org/10.1007/S11277-021-09393-0.

[14] T. Acharya, I. Khatri, A. Annamalai, M.F. Chouikha, Efficacy of Heterogeneous Ensemble Assisted Machine Learning Model for Binary and Multi-Class Network Intrusion Detection, 2021 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2021 - Proceedings. (2021) 408–413. https://doi.org/10.1109/I2CACIS52118.2021.9495864.

[15] M. Mehmood, T. Javed, J. Nebhen, S. Abbas, R. Abid, G.R. Bojja, M. Rizwan, A hybrid approach for network intrusion detection, Computers, Materials and Continua. 70 (2021) 91–107. https://doi.org/10.32604/CMC.2022.019127.

[16] A.N. Iman, T. Ahmad, Improving Intrusion Detection System by Estimating Parameters of Random Forest in Boruta, Proceeding - ICoSTA 2020: 2020 International Conference on Smart Technology and Applications: Empowering Industrial IoT by Implementing Green Technology for Sustainable Development. (2020). https://doi.org/10.1109/ICOSTA48221.2020.1570609975.

[17] P. Tahiri, S. Sonia, P. Jain, G. Gupta, W. Salehi, S. Tajjour, An Estimation of Machine Learning Approaches for Intrusion Detection System, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021. (2021) 343–348. https://doi.org/10.1109/ICACITE51222.2021.9404643.

[18] A. Binbusayyis, T. Vaiyapuri, Identifying and Benchmarking Key Features for Cyber Intrusion Detection: An Ensemble Approach, IEEE Access. 7 (2019) 106495–106513. https://doi.org/10.1109/ACCESS.2019.2929487.

[19] P.K. Keserwani, M.C. Govil, E.S. Pilli, P. Govil, A smart anomaly-based intrusion detection system for the Internet of Things (IoT) network using GWO–PSO–RF model, Journal of Reliable Intelligent Environments. 7 (2021) 3–21. https://doi.org/10.1007/S40860-020-00126-X.

[20] M. Aljanabi, M. Ismail, Improved intrusion detection algorithm based on TLBO and GA algorithms, International Arab Journal of Information Technology. 18 (2021) 170–179. https://doi.org/10.34028/IAJIT/18/2/5.

[21] Y. Xue, W. Jia, X. Zhao, W. Pang, An Evolutionary Computation Based Feature Selection Method for Intrusion Detection, Security and Communication Networks. 2018 (2018). https://doi.org/10.1155/2018/2492956.

[22] J. Hyma, P. Reddy, A. Damodaram, Performance analysis of Heterogeneous Data Normalization with a New Privacy Metric, Dat a Mining Journal of Comput Er Science IJCSIS Journal of Comput Er Science IJCSIS Sept Ember. (2018). https://sites.google.com/site/ijcsis/ (accessed October 9, 2021).

[23] A.L. Nogueira, · Casimiro, S. Munita, Quantitative methods of standardization in cluster analysis: finding groups in data, Journal of Radioanalytical and Nuclear Chemistry. 325 (2020). https://doi.org/10.1007/s10967-020-07186-6.

[24] M. Faisal, E.M. Zamzami, Sutarman, Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance, Journal of Physics: Conference Series. 1566 (2020) 012112. https://doi.org/10.1088/1742-6596/1566/1/012112.

[25] X.H. Cao, I. Stojkovic, Z. Obradovic, A robust data scaling algorithm to improve classification accuracies in biomedical data, BMC Bioinformatics 2016 17:1. 17 (2016) 1–10. https://doi.org/10.1186/S12859-016-1236-X.

[26] sklearn.preprocessing.RobustScaler — scikit-learn 0.24.2 documentation, (2020). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler (accessed May 27, 2021).