Journal of Information Systems Engineering and Management
2025, 10(15s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Analysis of feature selection algorithms for IDS using machine learning classification.

Mr. Kiran S Pawar[1], Dr. Babasaheb J Mohite[2]

[1] *Research Scholar, Zeal Institute of Bussiness Administration, Computer Application and Research (Zibacar) Pune, Savitribai Phule Pune University, Pune, India.*

[2] *Associate Professor, Zeal Institute of Bussiness Administration, Computer Application and Research (Zibacar) Pune, Savitribai Phule Pune University, Pune, India.*

ARTICLE INFO

Received: 01 Dec 2024

Revised: 26 Jan 2025

Accepted: 05 Feb 2025

ABSTRACT

The network and digital devices consumers are increasing rapidly, simultaneously the inevitability of its security and intruder detection. An intelligent intrusion detection system is needed to detect novel vulnerable attacks. The existing system practices the NS KDD, KDD Cup99, UNSW-NB15, and CICIDS2017 datasets that have old network traffic. The latest captured dataset UKM-IDS20 involves novel ARP Poisoning, DoS, Exploits, and Port Scan attacks. This article analyzed the filter-based feature selection(FS) method with rule based and tree based machine learning classifiers using multiclass classification: The Gain Ratio (GR), Chi-square, Info. Gain(IG), symmetric uncertainty(SU), and correlation(CR) filter based methods choose the vital feature from the UKM-IDS20 dataset. The highest accuracy and lesser model building time machine classifier are decided to select for the proposed framework. The preferred feature form accomplished IG feature evaluation method achieves superior accuracy on the Hoeffding tree based classifier compared with the JRip rule based classifier. The Hoeffding classifier proceeds the model within 0.13 seconds using 23 selected features. The proposed IDS framework compared to the existing systems.

Keywords: Feature Selection, Hoeffding, JRip, Info Gain (IG), Intrusion Detection System(IDS), Machine learning.

INTRODUCTION

Cybersecurity refers to defending internet-connected devices from internet threats. It includes safeguarding data, hardware, software, and other elements to prevent intruders from getting the privilege of using devices over networks. The open network presents the different latest types of vulnerabilities. The cybernetic devices must be protected over the network from these types of security threats, and information is then processed, stored, and transported through the network. The work [1] has included information technology organization security and maintenance rules on the cyber network. Digital systems lack built-in security measures that would allow them to fend against attacks, making them insecure. The attackers have slowed down or crashed the machine quickly via port scanning and DoS assaults by sending numerous connection requests. Therefore, the system signifies the necessity of robust intrusion detection(IDS) to prevent the vulnerable latest types of cyber-attacks. The objective of this paper is as follows.

- The effects of attacks and importance of effective intrusion detection (IDSs)over network.
- For the UKM-IDS 20 dataset, this work suggests noisy subset of features using the aid of filter based selection techniques.
- The irrelevant attributes are deleted from the captured network traffic and improved IDS evaluation on the UKM-IDS20 with the machine learning classifier.

Section II refer to the Literature work in IDS. Section III describe the importance of IDS. Section IV presents proposed system. Section V describes execution and result analysis. Lastly, section VI objective conclusion.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background and Methodology

The number of insecure digital device users are increasing daily, and the intruders can take the benefit to take advantage of the attention present over the network traffic. Figure I shows the influence of attacks on the system and the necessity of detecting intrusion over a network.

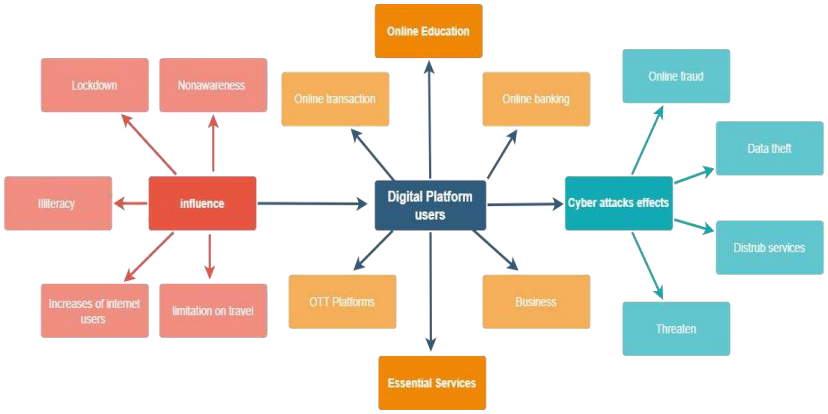

Figure I: Effects and methodology for detection of intrusion over network

In the table, I illustrate the example of attacks on the operating system(OSI) layer and the effects of attacks on a particular layer. The exploration process consists of actions necessary to carry out evaluation effectively. We must go through specific methodologies. The build- up time for every Machine Learning model is significantly reduced by enhanced feature selection in efficient intrusion detection, which plays a dominant role. It aids in simplifying the model by getting rid of extraneous features. Reduced features lead to a simpler model, which reduces the amount of computation needed to identify DoS attacks. The three main categories of FS algorithms are Filter, Wrapper, and Embedded. The earlier approach has a lower computational cost but only proceeds a subset of features based on the generated datasets characteristics. Wrapper-based produces high-quality results for chosen features by using the probabilistic accuracy of a learning algorithm that has already been trained.

Table1: Operating system layer and example of attacks

Layer	Attacks	Example	Effects of attacks
Application (7)	HTTP floods, DDoS, cross-site scripting, Slow Loris, SQL injections, Exploits etc..	POPS, SMTP, WWW browsers, HTTP, SNMP, Telnet, FTP, Domain DNS, ICMP, DHCP	No user can access network resources during an attack.
Presentation (6)	Phishing, malformed Secure Socket Layer requests SSL hijacking etc....	ASCII, TIFF, GIF, JPEG, MPEG, PICT, EBCDIC,	Stop accepting SSL connections, restart.
Session (5)	DDoS-attackers, Session Hijacking, exploit a flaw on Telnet server, Packet Sniffers and Cross Site Scripting (XSS) etc..	RPC, PAP, HTTP, FTP, SMTP, Secure Shell, SQL, NetBios.	Disable management operations.
Transport (4)	Session Hijacking, SYN Flooding, packet sniffing, ping sweeping, port scan, phishing, etc..	TCP, UDP, SPX	Connection limits of hosts.
Network (3)	IP address spoofing, Man in the middle, Warm-hole, Black-hole, Grey-hole, Byzantine, Flooding, Jelly-fish, Fabrication, Scan.	IPv4, IPv6, IPsec, AppleTalk, ICMP, IPX, IP, ARP and Routing protocol	effect on network bandwidth and add to the firewall's workload.
Data-Link (2)	MAC flooding, ARP Spoofing, VLAN, spanning tree attacks, Brute force attacks, Traffic analysis, etc..	PPP, IEEE 802.2, L2TP, ATM, ISDN, Frame Relay	Disrupts n/w flow of data by flooding on all ports.
Physical (1)	Sniffing, Jamming, Interceptions, Eavesdropping etc....	Digital, USB, BT, Ethernet, NIC, H/w	Data destroyed,

LITERATURE REVIEW

The objective of feature selection is to select a simple feature subset, like the class labels in classification problems, that significantly increases relevance and reduces redundancy to the targeted set. We will illustrate the existing literature on feature selection over network traffic.

The study [2] performed the SVM, decision tree, naive Bayes, and random forest classification algorithms for intrusion detection. The random forest classifier performs an accuracy of 97.49% with 0.08 seconds prediction time on the UNSW-NB15 dataset. The system does not accomplish the feature selection technique to remove the non-relevant features. The study [3] uses the five-level hybrid classification on flow statistics of hierarchical extreme learning machine (H-ELM) methodology with k-nearest neighbour approach (kNN) algorithms on the NSL-KDD benchmark dataset. The experiments give a level of accuracy of 84.29% for intrusion detection. The study [4] perform the naïve Bayes technique with support vector machine(NB-SVM) model for intrusion detection. The IDS framework with SVM classifier achieves an accuracy of 98.92% on the CICIDS2017 dataset and an accuracy of 98.58% on the Kyoto2006+ dataset.

The study proposes the model on [5] tree-based stacking ensemble technique for feature selection that comprises decision tree, XGBoost, and random forest. The model shows 99.9% and 95.26% of accuracy for the NSL-KDD and UNSW-NB15 datasets. In the [6] proposed the deep neural network (DNN) model based focal loss intrusion detection system (FL-NIDs). The model proposed to overcome the imbalanced problem of the benchmark dataset for better IDS. The research train the model in 25, 50, and 75% split for dataset training. The NSL-KDD, UNSW-NB15, and Bot-IoT datasets outperform the higher accuracy of 77.69% over scale 50 %, 73.39% over 100%, and 99.76 over 75%, respectively.

The article proposed [7] the adaptive particle swarm optimization and support vector machine (APSO-SVM) algorithm is included in the hybrid network IDS model to detect attacks accurately. The SVM parameters have been improved using the APSO algorithm. The KDD- CUP 99 dataset is used to evaluate the proposed IDS performance analysis, which has a detection accuracy of 97.687%. A IDS based on feature selection(FS) with HOE-DANN was presented in the study [8]. The results were evaluated using the UKM-IDS20 dataset using a discrete cuttlefish algorithm known as RS-DCFA. On the training set of the new UKM-IDS20, the HOE-DANN classifier attains a greater accuracy of 96.46%.

The study [9] selects the 30 important features using the RelifF filter built FS techniques on the UKM-IDS20 captured dataset to detect the intrusion. The proposed system achieves 99.969 % accuracy with 1.94 seconds to build the framework using FURIA classifier. The article [10] proposed the genetic algorithms(GA) for feature selection for evaluation of binary class in addition to multiclass classification using a random forest classifier(RF). The proposed GA-RF model gives 96.12% accuracy on the NSL-KDD dataset. The GA-RF model validated on the UNSW-NB15 dataset and performed an accuracy of 92.06%. The model consists of representation learning(RL) and a network intrusion detection system by explicit and implicit feature interaction, i.e. RL-NIDS. The study [11] outperforms feature selection with a multiclass classification accuracy of 95.72% and 81.38% on the AWIDS and NSL-KDD datasets, respectively.

The article [12] proposed the multiple IG, CFS, GR, SU, and ReliefF feature selection algorithms. The proposed model removes the irrelevant features having score of zero obtained from the score assigned by feature selection methods. The classifier PART achieves the suitable accuracy of 99.9591% on 67 relevant features of CICIDS-2017 dataset. The study [13] uses 46 features from the UKMIDS20 dataset and uses the outputs of a Furia rule-based classifier to reach a higher accuracy, F1 score, and recall of 99.969%. The research [14] further clarifies the model's adaptability by accounting for percentage variations in yields with an accuracy of 99.63%.

PROPOSED IDS SYSTEM

The proposed intrusion detection with feature selection is presented in figure II. The system contains a combined cleaned capture dataset, a feature selection technique, and a machine learning classifier to detect the attacks.The compact dataset is the network traffic captured and used in the suggested IDS. There is no missing or 'NaN' values in the dataset. Hence, the system does not carry out pre- processing on the captured dataset. The captured, cleaned dataset is immediately applied to feature selection techniques and classifiers. Due to a large number of features in

the dataset, machine learning classifiers necessitate additional time to model building. Some of the features are noisy, which reduces IDS performance.

The filter, wrapper, and embedded methods are available in machine learning for feature selection. The suggested IDS selected the filter methods of Gain Ratio (GR), Chi- square, symmetric uncertainty (SU), Information Gain (IG), and Correlation(CR) on the prepared captured dataset for feature selection. The system determines the score of individual features available in the available dataset by the exclusive FS technique and examines the results. The features with score values of zero or the closest to zero, that is, 0.1, are eliminated from the feature set. For intrusion detection purposes, the rest of the features are considered as appropriate features. Ten-fold cross-validation is used to analyze the relevant features chosen using the Hoeffding tree based machine learning classification algorithms.

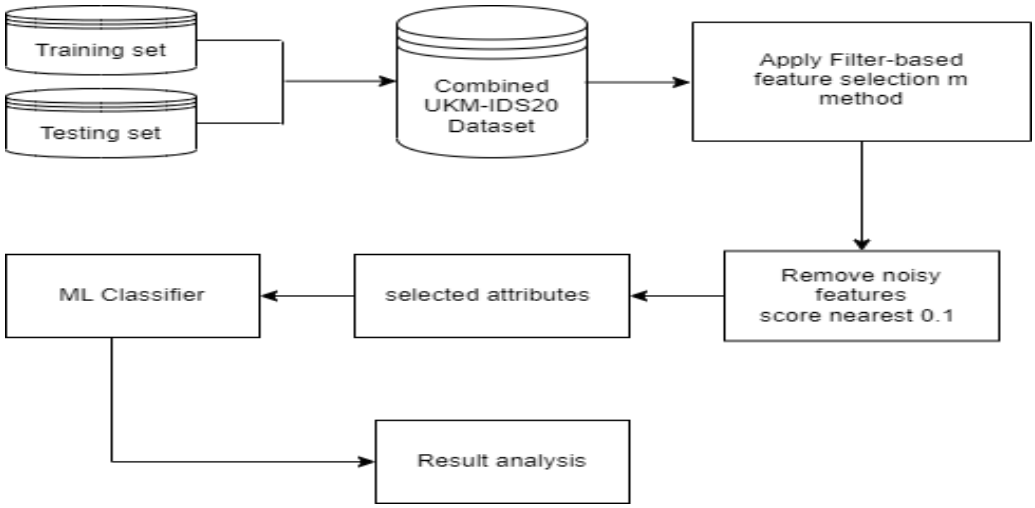

Figure II: Proposed feature selection IDS

SYSTEM IMPLEMENTATION

The system executed the recommended IDS with attribute selection as presented in the figure II to detect attacks using the Waikato tool. The tool Waikato 3.8.6 is used for classification and feature selection. The intrusion detection framework is tested on HLBS CM 44, 11th Gen Intel(R) Core(TM) i9-11900 @ 2.50GHz with 16 GB RAM on Windows10 Pro. The suggested FS framework is endorsed on the latest intrusion detection assessment UKM- DS20 dataset.

The most recent captured network traffic evaluation dataset, known as the [17] UKM- IDS20, is used to validate on the suggested IDS for feature selection. The University of Kebangsaan Malaysia's generated and collected the UKM-IDS20 captured dataset [8]. With the service of Windows Server 2012, penetration testing is carried out in a simulated environment using the tools Nmap, John the Ripper and Metasploit. There are 10,308 and 2579 total cases in the training and test datasets, respectively. The dataset has 46 attributes, including examples of attacks and normal behavior.

The number of normal and attack cases in the training dataset is 7,140 and 3,168, respectively. The test dataset has 1,769 examples of normal behavior and 810 instances of network traffic, respectively. The collection includes common cases of scans, ARP Poisoning, DoS, and exploited attacks. The proposed system is implemented to detect the attacks for binary multi-classification. The system is primarily established with Hoeffding Tree, Decision Stump, and LMT tree-based classifiers with 46 features.

Table I indicate the evaluation of tree-based classifiers on the newest UKM-IDS20. The results are evaluated to choose a solitary classifier for implementation. These classifiers will be used to create models using 10-fold cross-validation.

ANALYSIS OF RESULTS

As shown in Table 1, the LMT classifier has a superior accuracy of 99.9845% to the Hoeffding Tree, Decision Stump classifier but takes the additional time of 2.59 seconds to build the IDS model. The Hoeffding Tree takes less time to build the model and reaches a better accuracy of 99.9534% than the decision stump classifier. The system results associated with on CICIDS 2019 [16] dataset on random forest classifier. Therefore, our proposed system practices the Hoeffding Tree classifier for experimentation with the IDS model.

Table I: Analysis of tree-based classifier on multi classification

classifier	# Feature	# ACC %	ICI %	Building Time
Hoeffding Tree	46	99.9534	0.0466	0.25
Decision Stump	46	82.6492	17.3508	0.03
LMT	46	99.9845	0.0155	2.59

The system also evaluated on the JRip, Decision table, OneR and ZeroR. classifier with all features. Table II shows the analysis of multi classification on UKM-IDS20 dataset using classifiers. The results are analysed to select the single rule based classifier for further experiments.

Table II: Rule based classifier on multi classification

#classifier	# Feature	ACC %	ICI %	Building Time
JRip	46	99.9690	0.0310	0.39
Decision table	46	99.8836	0.1164	1.61
OneR	46	90.5797	9.4203	0.03
ZeroR	46	69.1317	30.8683	0.10

Table II shows that the rule based ZeroR classifier on multi classification performs the lowest accuracy of 69.1317% but takes lesser 0.1 seconds model building time with 46 features. The OneR classifier and Decision table achieve the accuracy of 90.5797% and 99.8836%, with 0.03 and 1.61 seconds of model building time, respectively. The system performs the higher accuracy of 99.9690% with 0.39 seconds model built up time on JRip classifier. So we have considered the JRip classifier for further research with the proposed model.

Later, IG, GR, CR, SU, and RelifF with ranker filter based attribute selection are also applied to the dataset for each feature; IG provides a score that ranges, in decreasing order, from 0.000263 to 0.851. The scores provided by GR for every feature vary from 0.0075300 to 0.3308100 in decreasing order. Each feature is given a score by ReliefF, with values ranging from 0.0000155195157911072 to 0.927372536 in order. The score range provided by Chi-square is 0 to 40706.64. SU offers a score for every attributes that ranges from 0 to 0.51309. In decreasing order, CR deliver a score of every attributes ranging from 0.0053 to 0.54068. The 46 features are cleaned up of noisy features by removing those with scores that are closest to zero or 0.1. The existence of irrelevant features in UMK-IDS20 is represented in Table III.

Table III: List of removed features

#Method	# Features	#Noisy Feature
IG	23	2,4,5,6,7,8,9,10,13,16,19,32,35,36,37,38,39,40,41,42,44,45,46
GR	17	2,4,5,6,8,9,16,19,25,28,32,33,34,35,36,38,45
Chi	4	32,38,40,46
SU	16	4,5,6,10,13,32,35,36,37,38,39,40,41,42,45,46
CR	17	1,3,5,13,17,28,32,33,34,35,36,37,38,39,40,45,46

Later, the IG, GR, Chi-square, SU, and CR features 23,17,4,16 and 17, respectively, are removed from the dataset before applying the Hoeffding tree based and JRip rule-based classifier on to the UMK-IDS20 dataset. The framework evaluation of the Hoeffding classifier on the UKM-IDS20 set is shown in Table III.

Table IV: Analysis of Hoeffding tree classifier

Method	# Feature	Accuracy (%)	# ICI %	Building Time(seconds)
IG	23	99.9922	0.0078	0.13
GR	29	99.9922	0.0078	0.20
Chi	42	99.9534	0.0466	0.25
SU	30	99.9845	0.0155	0.19
CR	29	99.9534	0.0466	0.17

Further, the system is evaluated on the LMT tree based classifier to analyse the performance. The LMT classifier applied on features selected by using the filter based feature selection technique of IG and GR. The LMT classifier has given the same accuracy as shown in table I. whereas the Hoeffding tree classifiers performance is higher than the other classifier mentioned above. Hence, we have applied the Hoeffding tree classifier to the feature selected from the filter based feature selection methods. Table IV shows the analysis of the Hoeffding tree classifier. The Hoeffding tree classifier performed the highest accuracy of 99.9922% with 0.2 seconds to build the model using 29 selected features from GR. The Hoeffding tree classifier achieves the same accuracy on IG using 23 features with 0.13 seconds to build the model. The classifier performed the accuracy of 99.9534%, 99.9845%, and 99.9534 using Chi- square, SU, and CR's 42, 30, and 29 features, respectively, compared with all features. Simultaneously system also tested on the JRip rule based classifier for evaluation.

Table V: Analysis of JRip rule-based classifier

Method	# Feature	# Accuracy (%)	# ICI %	Building Time
IG	23	99.9767	0.0233	0.23
GR	29	99.9767	0.0233	0.30
Chi	42	99.9612	0.0388	0.42
SU	29	99.9767	0.0233	0.33
CR	28	99.9767	0.0233	0.27

Table IV shows the performance of the JRip rule based classifier on selected features from the feature selection techniques. The JRip classifiers give a superior accuracy of 99.9767% using 23 selected features from IG and take 0.23 seconds to build the model. The JRip classifier performed the same accuracy of 99.9767% using CR's, SU's, and GR's 28, 29, and 29 selected features, respectively, with comparison from 46 features of UKM-IDS20. The chi-squared 42 selected features performed the lesser accuracy of 99.9612% with 0.42 seconds model built-up time.

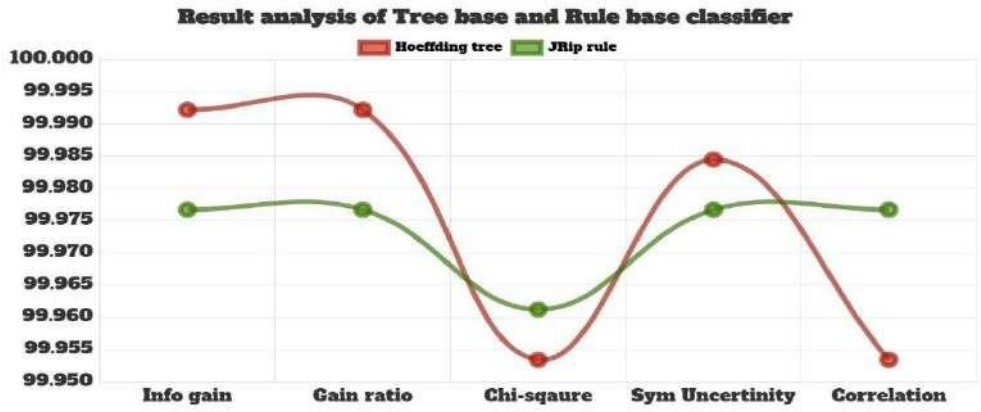

Figure III: Results of tree and rule base classifier

Figure III indicates the comparison analysis of the Hoeffding tree base classifier and JRip rule base classifier, subsequently choosing the important features to form all the captured traffic of the UKM-IDS20 dataset. As shown in Table VI, the current framework is also compared with existing IDSs. The system framework compared with existing IDSs is shown in Table VI.

Table VI: Compare with existing IDS

#Work	#Feature selected	#Classifier	#Dataset	Accuracy (%)
[8]	RS-DCFA	HOE-DANN	UKM-IDS20	96.46
[9]	ReliefF	FURIA	UKM-IDS20	99.9690
[12]	IG	PART	CICIDS-2017	99.9591
Proposed	IG	Hoeffding	UKM-IDS20	99.9922
		JRip	UKM-IDS20	99.9767

Table VI recommends that the proposed IDSs detect the latest types of attacks with 23 important features with the mentioned machine learning classifier for binary multi classification. The proposed system performed the greater accuracy compared with the existing attacks detection models. The outcomes analysis of proposed systems are listed below:

- The proposed framework model with feature selection methods resulting in the machine learning classifiers are performing potentially faster and simpler.
- The proposed IDS accuracy prediction improved by discarding irrelevant or noisy features from the network traffic database.
- Identifying the relevant features gives more insight into the nature of the corresponding classification problems.
- The proposed feature selection framework model building time decreased compared with traditional Intrusion Detection models.

CONCLUSION

The article suggested using the UKM-IDS20 feature selection to find the most recent attacks. The suggested system used filter based FS techniques, rule-based, then tree based classifiers to evaluate the UMK-IDS20 dataset. The IG obtained the 23 important feature for the UKM-IDS20 captured dataset and performed the greater accuracy with minimum time for building the model using the Hoeffding tree based classifier compared with all the features of the UKM-IDS20 dataset. The suggested IDS framework detects the most recent kinds of network traffic that are not included in the old traffic datasets. In the scope of the research, we intend to use evolutionary algorithms using hybrid ensemble framework for FS in IDS for better accuracy, FP rate and precision improvement.

In future, we will implement the evolutionary methods by combination with algorithms on benchmark datasets.

REFRENCES

[1] Mohite, Babasaheb J., and D. M. Kumthekar. "Awareness of IT security laws and security maintenance policies: two pillars of information security management." (2013).
[2] Belouch, M, El Hadaj, S., & Idhammad, M. Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science, 127, 1-6 (2018).
[3] Latah, M., & Toker, L. (2020). An efficient flow-based multi-level hybrid intrusion detection system for software-defined networks. CCF Transactions on Networking, 3(3), 261-271].
[4] Gu, Jie, and Shan Lu. "An effective intrusion detection approach using SVM with naïve Bayes feature embedding." Computers & Security 103 (2021): 102158.
[5] Rashid, Mamunur, et al. "A tree-based stacking ensemble technique with feature selection for network intrusion detection." Applied Intelligence (2022): 1-14.
[6] Mulyanto, M., Faisal, M., Prakosa, S. W., & Leu, J. S. (2020). Effectiveness of focal loss for minority classification in network intrusion detection systems. Symmetry, 13(1), 4.
[7] Luo, Yin. "Research on Network Security Intrusion Detection System Based on Machine Learning." International Journal of Network Security 23.3 (2021): 490-495.
[8] Muataz Salam Al-Daweri, Salwani Abdullah, Khairul Akram Zainol Ariffin, "An adaptive method and a new dataset, UKM-IDS20, for the network intrusion detection system", Computer Communications, Volume 180, 2021.

[9] Pawar, K., Mohite, B., & Kshirsagar, P. Analysis of Feature Selection Methods for UKM- IDS20 Dataset. In International Conference on Computing in Engineering & Technology (pp. 461-467). (2022). Springer, Singapore.
[10] S Liu, Zhiqiang, and Yucheng Shi. "A hybrid IDS using GA-based feature selection method and random forest." Int. J. Mach. Learn. Comput 12.02 (2022): 43-50.
[11] Wang, Wei, et al. "Representation learning-based network intrusion detection system by capturing explicit and implicit feature interactions." Computers & Security 112 (2022): 102537.
[12] Vaidya, Atharva, and Deepak Kshirsagar. "Analysis of Feature Selection Techniques to Detect DoS Attacks Using Rule-Based Classifiers." Applied Information Processing Systems. Springer, Singapore, 2022. 311-319.
[13] Pawar Kiran., Mohite Babasaheb. Performance analysis of UKM-IDS20 dataset on machine learning algorithms, Journal of Statistics & Management Systems, ISSN 0972-0510 (Print), ISSN 2169-0014 (Online) Vol. 27 (2024), No. 5, pp. 997–1008, DOI : 10.47974/JSMS-1296.
[14] Manneh, M., Ansah, P., Tetarave, S.K., Mishra, M.R., Kalaimannan, E. A Comparative Analysis of Random Forest and Support Vector Machine Techniques on the UNSW-NB15 Dataset. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24). ICR 2024. Lecture Notes in Networks and Systems, vol 1058. Springer, (2024). https://doi.org/10.1007/978-3-031-65522-7_18
[15] WEKA: Software machine learning, the University of Waikato, Hamilton, New-Zealand https://www.cs.waikato.ac.nz/ml/weka/ accessed in (2022).
[16] Kiran S. Pawar, Babasaheb J. Mohite, Design Framework Model for Network Intrusion Detection With Feature Selection Algorithms Using Machine Learning". *Machine Intelligence Research*, vol. 19, no. 1, Feb. 2025, pp. 46-63.
[17] The Kaggle repository, "https://www.kaggle.com/datasets/muatazsalam/ukm-ids20" accessed by 08:00 PM 07/01/2023.
[18] Kiran S. Pawar, Babasaheb J. Mohite, Framework for IDS using Machine Learning algorithms on UKM-IDS20 dataset, Industrial Engineering Journal 53 (Issue 7, No.2), 43-47 (2024).