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

Performance analysis of UKM-IDS20 dataset on machine learning algorithms

Kiran S. Pawar *
Babasaheb J. Mohite †
Zeal Institute of Business Administration, Computer Application & Research
Savitribai Phule Pune University, Narhegaon
Taluka – Haveli
Pune 411041
Maharashtra
India

Abstract

Machine learning algorithms are essential in classification and regression because they might a significant outcome on the accuracy of the classifier. It decreases the total of features of the traffic records. The algorithms adaptively improve their performance, reducing the processor and memory norms. This work proposes the machine learning algorithms on the available well-known mathematically proven classifiers. The proposed system for intrusion detection is tested and validated on UKM-IDS20 datasets with the suite of the classifiers. The system uses the BayesNet, Lazy Kstar, NaiveBayes, NavieBayesMultinomialText, NavieBayesUpdateable, Function LibSVM, Function SGDText, and Function Voted Perception classifier from the Bayes and function-based classifier to detect intrusion detection. The BayesNet classifier performed the higher accuracy of 99.9845% with 0.27 seconds to build the model on the dataset. The proposed system also served on the rule-based classifiers to achieve the accuracy, F1 score and built up a time to detect the intrusion detection on UKM-IDS20 datasets. The proposed system uses the rulebased machine learning algorithms JripOneR, Decision Table, Furia, PART, and ZeroR classifiers for network intrusion detection. The Furia rule-based classifier achieves the higher accuracy, F1 score, and Recall of 99.969%, 99.9775%, and 99.9888%, respectively, with 46 features of the UKMIDS20 dataset. The proposed framework achieve a superior accurateness of 99.9845% and an F1 score of 99.9887% with 1.6 second model built-up time with a Random Forest tree-based classifier

* *E-mail:* **ksp.comp@coeptech.ac.in (Corresponding Author)**
† *E-mail:* **babasaheb.mohite@zealeducation.com**

on the UKMIDS20 dataset used in NIDS. The system applied on filter-based algorithms info gain (IG) with top ranked and threshold model to analyze the performance.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Network intrusion detection system (NIDS), Web attacks, Info gain, Random forest, FURIA Classification.*

1. Introduction

Cyber security plays a vibrant role in getting secure over the digital network every new day. The electronics transaction is increasing as internet users are utilizing global networks. The OTT platform has not under control over the web. This network is not secure as they do not have inbuilt security protection. Hence, the intruders can benefit from it to penetrate the network through different types of attacks. The press report of renowned organization Gartner [1] pretend by 2025. Cyber theaters have weaponries functioning tools surroundings to effectively mischief or destroy the data. To overcome that, organizations can reduce risk by implementing prevention control frameworks.

The wired [2] release the report on data breaches, digital scams, ransomware attacks, and hacks around the world in the first half of the year 2022. After covid 19 pandemic, the cyber security vulnerabilities have proven that human rights are enmeshed in all aspects globally. The hacktivists are attentive to increasing the disruptive hacks and DDoS attacks affecting the services. This underlined the importance and prevention against digital firepower.

The attackers have used ARP poisoning, scans, exploits, and DoS attacks to crash the machine. DoS attacks are carried out by overflowing with traffic, sending it data that triggers a crash, or getting control over the network. The UKMIDS20 dataset contains ARP poisoning, scans, exploits, and DoS attacks. So there is a need to secure the networks through machine learning methods and artificial intelligent algorithms. The machine learning methods and classification techniques give higher performance and better accuracy on different datasets.

The study [3] discussed the ensemble feature selection technique for classification of attacks using J48 classifier on correlation, gain ratio, ReliefF, Information gain, and chi-square. This study examines the model on a UNSW-NB15 dataset, the majority of current attacks. Therefore, we have performed a classification method on to the UKM-IDS20 dataset

with different classifiers with threshold and top ranked filter based feature reduction methods and get better accuracy.

The contribution of this article is as follows.

Propose machine learning-based classification methods in IDS

- To detect mentioned attacks in the UKM-IDS20 dataset.
- The classification methods achieve higher accuracy on to the UKM-IDS dataset among Random Forest Tree-based classifier.

The Section 2 describes the traditional datasets used for IDS. Section 3 contributes the projected IDS with classification methods. Section 4 describe a Framework, Execution and result examination.

2. Literature Survey

This study mentions the related mechanisms on Network Intrusion detection systems (NIDS) on web attacks datasets. The study [4] used Recursive Feature Elimination (RFE) algorithm to perform feature reduction with random forest classifier and achieved the accuracy of 91% by Deep Multilayer Perception (DMLP) model with obtained features on the CICIDS2017 dataset.

The article [5] used the Multi-Layer Perceptron (MLP) and Random Forest(RF) technique to detect the DoS attacks. The experiment uses the Weka tool for the evaluation of the CICDS 2017 dataset at the application layer to detect the DoS attacks. The Random Forest(RF) achieves higher accuracy compared to MLP. The study [6] approaches based on the performance for feature selection using JRipalgorithms on the CICIDS17 Friday dataset for recognition of DDoS LOIC (Low Orbit Ion Cannon) attack. The systems performed a 99% accuracy rate with minimum building time using 18 reduced relevant features out of 86.

In this study, the author [7] explored the effects of feature reduction on the three classifiers (Naive Bayes, Multilayer Perceptron and Decision Tree) on J48 world network traffic.The study [8] selects 38 features with PSO feature selection method and The system performed the highest accuracy of 99.12% with Naïve Bayes classifier on the KDD Cup99 data.

The study [9] described the original feature size as reduced to essential features that can carry out better accuracy, and testing time on the CIDIDS-2017 Friday dataset. The system performed greater accuracy of 96.36% for 15 reduced feature selections using the J48 classifier. The study

[10] used IDS future selection methods on the CIDIDS-2017 dataset on PART classifier. The system PART achieves the upper correctness and F1 score of 99.9191% and 99.5 %, respectively, with 11.48 seconds built up time with PART classifier to detect web attacks.

The study [11] uses Genetic Algorithm (GA) to reduce the features with random forest gives a higher accuracy of 86.5% for feature reduction on the NSL-KDD data. Analysis of rule- based classifier is provided in Table II. The study mentioned [12] the novel approach for Threshold Determinated in IDS on the CIC-DDoS2019 dataset. The article represents the IDS approach by combining cumulative sum algorithms (CUSUM), k-nearest neighbor algorithm, and exponentially weighted moving average (EWMA) in one Triple Modular Redundancy (TMR) algorithm using python. The results analysis shows improvement in accuracy and precision using the TMR approach.

The system [13] performed with a threshold of filter-based info gain(IG) and correlation (CR) feature reduction for the DDoS attacks detection. The literature study suggests the related traditional methods that provide solutions on web attacks and do not perform on the UKMIDS20 attacks. therefore, the analysis motivates us to detect the latest types of attacks succinctly.

3. Proposed IDS with Bayes and Function Based classification algorithms

The Proposed Intrusion Detection System using rule-based ensemble classifiers. For designing and pre-processing the system, has to go through the following stages [see Figure 1]. The combined training and testing set of the UKMIDS20 dataset with ensemble rule-based classification

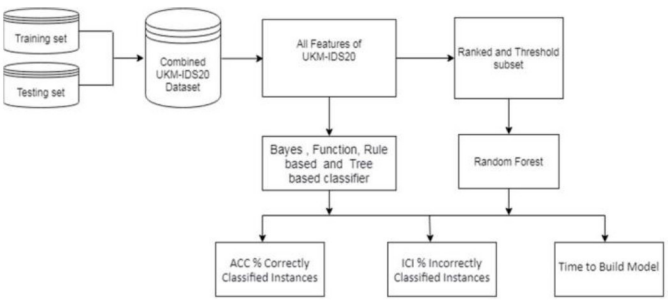

Figure 1

Classification with different machine Learning Algorithm

methods. The [14] DoS, ARP Poisoning, port Scan, and Exploits attacks have been compared with the new UKMIDS20 dataset. As shown in figure1. first, select all the features and calculate the accuracy ACC% and ICI% and time to build the model. The UKMIDS20 feature dataset applied on the classifies from the suit of machine learning algorithms available in Weka [15].

4. Framework Execution and analysis

We have executed the suggested method for feature reduction as presented in Figure 1. Weka [19] 3.9.3, an machine learning open source tool, is used for classification and feature selection. This method is performed on HPZ230 Workstation on Windows7 Professional. The proposed classification method in network intrusion detection is performed on the modern IDS evaluation data (UKMIDS20 Dataset). The dataset consists of four attacks with 46 featured attributes, including the one class Binary feature. The[1] UKM- IDS20 has released in year 2020 in the UniversitiKebangsaan Malaysia. The total number of Connection records is 12887 contains records of traffic Normal, ARP Poisoning, Dos, Scans, Exploits including (Testing set of 2579 instances and training set of 10308 instances). The presented network intrusion detection system is executed, validated, and verified on the UKMIDS20 data. The system selects all 46 features, excluding the class label from the preprocessed combined training set and testing set of the UKMIDS20 dataset. The feature classification performed on BayesNet machine learning algorithms

Table I

Analysis of bayes and function based classifier

#Bayes and Function Based	# Features	ACC %	ICI %	Building Time (sec)
Bayes Net	46	99.9845	0.0155	0.27
lazy.Kstar	46	85.3341	14.6659	0.01
Naïve Bayes	46	91.1306	8.8694	0.08
NaiveBayesMultinomialTex	46	69.1317	30.8683	0.03
NaiveBayesUpdateable	46	91.1306	8.8694	0.08
Function LibSVM	46	70.0008	29.9992	177.98
Function SGDText	46	69.1317	30.8683	0.57
Function Voted Perception	46	75.7042	24.2958	1.69

achieves a higher accuracy of 99.9845% with 0.27 second built-up time than lazy.Kstar, NaiveBayes, NaiveBayes MultinomialText, NaiveBayes Updateable Bayes based classifiers and Function LibSVM, Function SGDText, Function Voted Perception function- based classification algorithms. [Table I] shows the analysis of Bayes and function-based machine learning algorithms.

A few performance evaluation measurements have been recorded in IDS literature and used by researchers to approve their IDS methods for anomaly detection. F1-score is the harmonic mean (in percentile) calculated by the formulas given below in confusion matrix [16] These are abbreviated as True Positives(TP), False Negatives(FN),False Positives(FP), True Negatives(TN). The recall is the ratio of records suitably categorized above the number of total actual attacks. The formula of calculating the recall is given below.

$$\textbf{Recall} = \frac{TP}{TP} + FN \tag{1}$$

The calculation of suitably classified records above the number of total records. And accuracy is calculated using the given formula

$$\textbf{Accuracy} = \frac{(TN+TP)}{(TP+TN+FP+FN)} \tag{2}$$

The proportion of records suitably classified above the number of total records categorized as attacks. The precision is consideredwith the given formula.

Table II

Analysis of rule- based classifier

#Rule Based	# Feature	ACC %	ICI %	# Time	# Precision	#F1	Sensitivity/ Recall, or TPR, DR
Decision Table	46	99.9224	0.0776	2.74	100	99.94390846	99.88787981
Furia	46	99.9690	0.031	3.41	99.96632619	99.97754827	99.98877288
Jrip	46	99.9146	0.0854	0.69	99.94387698	99.93826814	99.93265993
OneR	46	99.1464	0.8536	0.09	99.2142777	99.38160558	99.54949882
PART	46	99.9379	0.0621	0.47	99.94387698	99.95509654	99.96631863
ZeroR	46	69.1317	30.8683	0.1	100	81.74894476	69.13168309

$$\textbf{Precision} = \frac{TP}{TP+FP} \tag{3}$$

F1-Score is calculated with the below formulas

$$\textbf{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The proposed system was performed on rule-based classifiers with 46 features to detect intrusions on the UKMIDS20 dataset. The FURIA classification model achieves the higher accuracy, F1-score, and Recall of 99.969%, 99.9775%, and 99.9887%, respectively, compared with decision table, Jrip, OneR, PART, ZeroR are with 46 features of UKMIDS20 dataset. The decision table performed the highest precision of 100% with a zero FAR rate compared to the mentioned rule-based classifiers with the 2.74-second model built up time. The system achieved a lower accuracy of 69.1317% and did not calculate the FAR value using the ZeroR rule-based classification model.

Table III shows the investigation of the Tree-based classifier. The Random Forest tree-based algorithm achieves 99.9845% accuracy with 1.6 seconds time using 46 features of the UKM-IDS20 dataset. The random forest algorithms also performed the 100% Detection rate (DR) with the lowest false alarm rate (FAR) on the newly mentioned dataset. Therefore, the random forest algorithms are proposed for the further investigation. The decision stump algorithm takes less 0.09 time but performed the lowest 86.2885% accuracy and highest FAR of 0.304093567 to detect the

Table III

Analysis of tree- based classifier

#Rule Based	# Feature	ACC %	ICI %	# Time	F1%	#Recall, DR	#FAR
Decision Stump	46	86.2885	13.7115	0.09	89.06085557	99.29596908	0.304093567
Forest PA	46	99.9534	0.0466	5.9	99.96632619	99.96632619	0.000754148
Hoeffding Tree	46	97.3384	2.6616	0.39	98.10591419	96.55434783	0.007051804
J48	46	99.9534	0.0466	0.3	99.96632619	99.96632619	0.000754148
Random Forest	46	99.9845	0.0155	1.6	99.988774	100	0.000502513
Random Tree	46	99.9146	0.0854	0.06	99.93826121	99.94387068	0.001507917
REPTree	46	99.9457	0.0543	0.13	99.96071609	99.95510662	0.000754337

Table IV

Analysis of comparison of classifiers with random forest

Method	# Feature	Classifier	ACC %	Building Time (sec)	FAR
[17]	24	PART	99.9593	133.66	0.000542
Proposed RF	46	RF	99.9845	1.6	0.0005025

UKM-IDS20 Dataset. Most researchers used the J48 for precise predictions that performed 99.9534% accuracy with less built-up time in 0.3 seconds compared with the mentioned tree-based machine learning classifier. Table III shows the performance of the UKM-IDS dataset on tree-based machine learning classifiers to detection of the attacks.

The Table IV shows the comparison of classifiers with traditional system using random forest tree based classifier. The traditional [17] system with part rule based classifiers performed the 99.9593 % accuracy with 0.000542 False alarm rate on CICIDS 2017 dataset to detects the DoS based attacks.

The system applied algorithms like info gain (IG), correlation (CR), gain ratio (GR), Symmetric Uncertainty (SU), and RelifF on the UKMIDS20 dataset. The method gives the score of features in the range of 0 to 1 in descending order mentioned in the previous work [18]. The model uses the ranker-based selection and selects the top 23 and top 38 features from IG. The proposed model selects scores above the 0.2 range from the filter-based info gain. The tree-based classifier applied to analyze the performance. Similarly, the system also explored several filter-based models like CR, GR, SU, and RelifF.

The performance shown in Table V with a Random Forest (RF) classifier for the detection of DDoS attacks. Table V shows that the proposed technique produced the same accuracy of 99.9845% with 16 selected features. The system takes the less built-up mode time of 1.28

Table V

Analysis of feature selection on random forest classifier

Method	# Feature	ACC %	Building Time (sec)
IG Top 23	23	99.9845	1.34
IG Top 38	38	99.9845	1.61
IG- 0.2	16	99.9845	1.28

Table VI

Analysis of comparison of classifiers with traditional system

Method	# FS	Classifier	ACC %
[14]	RS-DCFA	HOE-DANN	96.4600
[13]	IG and CR Threshold	J48	99.8846
Proposed Threshold	IG- 0.2 Threshold	Random forest	99.9845

seconds compare than the traditional topranked methods on the UKMIDS20 dataset.

The presented machine learning filter methods include Top ranker 23 and 38 features present in weka tools. The IG-0.2 performed the same accuracy, precision, recall, F1-score, and FAR with minimum 1.28 seconds model building time compared to IG Top 23, IG Top 38, CR Top 38, GR Top 38, SU Top 38, and RelifF Top 38.

The study [18] with ReliefF obtain 30 selected features for dataset and performed the similar accurateness using the FURIA classifier with 1.94 seconds built-up time. Table VI shows that the proposed systems detect the latest types of attacks with a Random Forest (RF) classifier compared to the traditional method stated in [13]. The suggested system performed the high accurateness compared to [14].

5. Conclusion

This article analyzes the best relevant feature of the UKM-IDS20 data to build a multi-class IDS to cyber-atmosphere. The system set the example of a top-ranked and threshold approach for feature reduction with algorithms. The proposed system objective is to detect the traditional and recently discovered exploits, scans, ARP poisoning, and DoS types of attacks. The system was explored with several classifiers on the UKM-IDS data. The Random Forest(RF) classifier achieves better performance on the proposed model.

The proposed feature selection technique produced the same accuracy of 99.9845% with 16 selected features. The system takes a less built-up mode time of 1.28 seconds to compare than the traditional top-ranked featured selection methods on to the UKMIDS20 dataset. In the future, we plan to analysis of IDS with combination of evolutionary machine learning algorithms for the dataset for feature selection.

Dataset accessibility statement

The data are available in [20] kaggle [public dataset repository] at [https://www.kaggle.com/datasets/muatazsalam/ukmids20], The Muataz Salam Al-Daweri and etl.. has released UKM- IDS20 [1] in year 2020 in the Universitiy Kebangsaan Malaysia (UKM).

References

[1] 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).

[2] The WIRED release, https://www.wired.com/story/worst-hacks-breaches-2022/ 7.00 AM July 4, 2022, last accessed by 07/07/2023.

[3] Esra Mahsereci Kara bulut, Selma Ayşe Özel, TurgayİbrikÇi, "A comparative study on the effect of feature selection on classification accuracy", *Procedia Technology*, Volume 1, (2012).

[4] R. B. Adhao and V. K. Pachghare, "Performance-Based Feature Selection Using Decision Tree," 2019 International Conference on Innovative Trends and Advances in Engineering and Technology (ICITAET), pp.135-138, (2019).

[5] Gartner Press Release, "https://www.gartner.com/en/newsroom/press-releases/2021-07-21- gartner-predicts-by-2025-cyber-attackers-will-have-we" STAMFORD, Conn., July 21, (2021), last accessed by 13/05/2022.

[6] Kumar, V., Sinha, D., Das, A.K. et al. An integrated rule based intrusion detection system: analysis on UNSWNB15 data set and the real time online dataset. *Cluster Comput* 23, 1397– 1418 (2020).

[7] Kshirsagar, D., & Kumar, S. An ensemble feature reduction method for web-attack detection. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(1), 283-291. (2020).

[8] O. I. Ahmed and C. Varol, "Detection of Web Attacks via PART Classifier," 2021 9th International Symposium on Digital Forensics and Security (ISDFS), pp. 1-4, (2021).

[9] Talita, A. S., Nataza, O. S., &Rustam, Z. Naïve Bayes Classifier and Particle Swarm Optimization Feature Selection Method for Classify-

ing Intrusion Detection System Dataset. In *Journal of Physics: Conference Series*, Vol. 1752, No. 1, p. 012021, (2021).

[10] Shah, S., Muhuri, P. S., Yuan, X., Roy, K., & Chatterjee, P. (2021, April). Implementing a network intrusion detection system using semi-supervised support vector machine and random forest. In Proceedings of the 2021 ACM Southeast Conference, pp. 180-184. (2021)

[11] S. Ustebay, Z. Turgut and M. A. Aydin, "Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier," 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), 2018, pp. 71-76, (2018).

[12] Babu, Md. A., Ahmmed, Md. M., Ferdousi, A., Mostafizur Rahman, M., Saiduzzaman, Md., Bhatnagar, V., … Poonia, R. C. (2022). The mathematical and machine learning models to forecast the COVID-19 outbreaks in Bangladesh. *Journal of Interdisciplinary Mathematics*, 25(3), 753–772.

[13] Kshirsagar, Kumar, A feature reduction based reflected and exploited DDoS attacks detection system. *J. Ambient. Intell. Humaniz. Comput.* 13, 393–405 (2021).

[14] Deepak Kshirsagar, Sandeep Kumar, An efficient feature reduction method for the detection of DoS attack, ICT Express,Volume 7, Issue 3, 2021, Pages 371-375, ISSN 2405- 9595. (2021).

[15] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. Deep learning approach for intelligent intrusion detection system. Ieee Access, 7, 41525-41550. (2019).

[16] Pawar, K., Mohite, B., &Kshirsagar, P. Analysis of Feature Selection Methods for UKM-IDS20 Dataset. International Conference on Computing in Engineering & Technology (pp. 461-467). Springer, Singapore. (2022).

[17] Wankhede, Shreekh, and Deepak Kshirsagar. "DoS attack detection using machine learning and neural network." Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, (2018).

[18] G. V. Patil, K. V. Pachghare and D, "Feature Reduction in Flow Based Intrusion Detection System," 2018 3rd IEEE International Confer-

ence on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2018, pp. 1356-1362, (2018).

[19] The university of Waikato, "https://www.cs.waikato.ac.nz/ml/weka/" last accessed by 10:00 AM 07/07/2023.

[20] The Kaggle repository, "https://www.kaggle.com/datasets/muatazsalam/ukm-ids20" accessed by 08:00 PM 07/01/2023.

[21] Zheng, Jun, and Yujie Zhou. "Intrusion detection algorithm based on improved principal component analysis." *Journal of Discrete Mathematical Sciences and Cryptography* 20.5 (2017): 1007-1016. (2017).

[22] Anish Khobragade, Rushikesh Mahajan, Hrithik Langi, Rohit Mundhe & Shashikant Ghumbre Effective negative triplet sampling for knowledge graph embedding, *Journal of Information and Optimization Sciences*, 43:8, 2075-2087, (2022).

Received May, 2024