ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

COMPARATIVE ANALYSIS OF CLASSIFICATION OF MACHINE LEARNING MODELS ON KDDCUP99 AND CIC-IDS DATASETS FOR THE INTRUSION DETECTION SYSTEM.

[1]Pranali R. Landge, [2]Dr. Swati S. Sherekar

[1]Research scholor P.G. Department of Computer Science Sant Gadge Baba Amravati University, Amravati Maharashtra, India pranalilandge9@gmail.com

[2]Professor P.G. Department of Computer Science Sant Gadge Baba Amravati University, Amravati Maharashtra, Indiaswatisherekar@sgbau.ac.in

Abstract

Intrusion Detection System (IDS) is the key part of system defense for growing attacks. The inconsistent and unreliable performance of anomaly-based intrusion detection methods can be attributed to outdated test and validation datasets. This study investigates the performance of various machine learning models on the KDDCup99 and CIC-IDS popular datasets, focusing on their predictive performance accuracy, computational efficiency and generalizability. A suite of models including logistic regression, decision trees, random forest, support vector machines, gradient boosting were evaluated. Results indicate that different ML models and deep learning methods outperform traditional algorithms in terms of accuracy, but they require higher computational resources. Model performance varied significantly between the two datasets, highlighting the impact of data characteristics on model efficiency.

Keywords: machine learning models, performance, intrusion detection systems

I. Introduction

Today information technologies vigorously involve in every part of life and machine learning methodologies plays important role in intrusion detection(attacks) which furthermore helpful for prevention. Our dependability increases on new technologies such that big data and Internet of Things (IOT) in the military, business, healthcare. Digital devices store and handles huge amount with variety of data. Now a days intrusion taken on a new dimension by jumping from cyberspace to physical space.[1] Misuse based and anomaly-based intrusion detection are two main techniques used for IDS. Misuse based technique to detect intrusion relies on set of attacks' signatures. Misuse based gives low false alarm and high detection rate. The main drawback of misuse based is that they are incapable against unidentified attacks. Model describing the normal behavior is built by anomaly-based detection technique. Anomaly-based detection technique capable to detect unknown attacks with high false alarm rate. Machine Leaning (ML) models are increasingly used in diverse domains for predictive tasks. In this paper we propose to use five machine learning approaches Naïve Bayes, Decision Trees, Random Forest, Support Vector Machine, Gradient Boosting Machine. However, their performance is influenced by dataset-specific characteristics such as data size, feature distribution and noise. This study compares the performance of popular ML models on KDDCup dataset. The goal is to provide insights into the suitability of different models for these datasets and to identify generalizable patterns in ML performance. Through this paper we conduct a comparative study that aims to evaluate machine learning models performance for intrusion detection. We use accuracy detection rate and false alarm rate as performance metrics and the labelled KDDCup99 and CIC-IDS as a dataset. CIC-IDS also called CCID. In this paper section II provides the review of related work. Section III highlights performance evaluation measures of methodologies. Section IV discusses the experimental result and discussion of KDDcup99 and CIC-IDS datasets. Section V concludes the paper for future research.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

II. Related Research

Using the KDDCup99 dataset, the effectiveness of machine learning models in IDS has been the subject of several studies. IDS can be deployed for the detection of a variety of attacks. According to [2], cyber security attacks can be categorized based on purpose, legal classification, based on severity of involvement, based on scope, and based on network types. Reconnaissance attacks, access attacks, and denial of service attacks are examples of attacks with a purpose. A reconnaissance attack is a dangerous type of attack as the attacker trap victims into becoming their friend to extract sensitive information from them [3]. Packet sniffers, port scanning, and internet information queries are all examples of these attacks. The intruder is able to access a device during access attacks. Man-in-the-middle attacks, phishing, social engineering, and attacks on secret code are all examples of these kinds of attacks. A denial of service (DoS) attack is the third kind of attack that falls under this category. A DoS attack simply involves flooding a victim site with a large number of requests, taking advantage of the internet's connectivity to cripple the victim site's services. It can come from a single source or from multiple sources, the latter of which is known as a distributed denial of service attack (DDoS) [4]. Examples of DoS attacks include Smurf, SYN flood, and DNS attacks.

In the second type of categorization, legal classification attacks, the attacks include cybercrime, cyber espionage, cyber terrorism, and cyberwar. Cybercrime attacks example is identity theft which involves the use of an account without the owner's permission [5]. Another type within this category is cyber espionage, or cyberspying attack, which involves the use of computer networks for gaining illegal access to confidential information especially associated with governments [6]. Cyber terrorism attacks are carried by extremists using cyberspace. Lastly, cyberwars are wars fought between nations using cyberspace. Attacks can be divided into two categories under the third type of classification based on the degree of involvement: active attacks and passive attacks [2]. Simply put, the difference between these two types of attacks is that in the first, the attacker seeks to alter the operation or resources of the system, while in the second, the attacker makes use of the information without any alteration or modification of resources or operations. Examples of active attacks include spoofing, man-in-the-middle attacks, buffer overflow, and others. Keystroke logging [7] and backdoors are two examples of passive attacks. Cyber-attacks can be classed into malicious and non-malicious attacks in the fourth category of categorization. In order to carry out an attack with the intention of causing harm, malicious assaults employ various types of software, including viruses, worms, Trojan horses, spyware, adware, botnets, and others. Non-malicious assaults, on the other hand, are unintentional attacks carried out by untrained staff that may result in modest data loss [8]. The final type of categorization of cyber security attacks is based on network types where attacks are classified according to the network types such as mobile ad hoc networks (MANET) and wireless sensor networks (WSN) [9]. Black hole, flood rushing, and Byzantine attacks are all examples of attacks on MANET. Application-layer attacks, network layer attacks, and other network layer attacks are additional examples on WSN. As a result of this work, we have evaluated 74 intrusion detection classification methods. The purpose is to guide a searcher's initial efforts to detect intrusions using data mining methods. This paper shows the best twenty classifiers for each attack type of KDDCup99 dataset as well as the best twenty overall classifiers. We can conclude from our findings that there is no one classifier that consistently performs better than the others. However, generally speaking, rule based and decision tree-based methods got sufficient results for intrusion detection. We have proposed a feature selection strategy in this paper. for IDS to produce the optimal subset of features that can be used to classify the instances of KDDCup99 and UNSW-NB15 datasets. The proposed approach is based on three stages: a preprocessing stage, a feature selection stage, and a classification stage. The preprocessing stage consists of reducing the size of the datasets through resampling, changing the attribute values in a way to be handled with LR

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

classifier, and removing redundant records if they exist. The experimental results are promising with an accuracy of classification equal to 99.90%, 99.81% DR and 0.105% FAR with a subset of only 18 features for the KDDCup99 dataset. In addition, the selected subset has a 99.98% DR for the DoS category. Table 15 [10] contains the UNSW-NB15 results that were obtained. it has been observed that the DT classifier is more successful than the other classifiers used. The DT success rates for the CSE-CIC IDS-2018, ISCX-2012, NSL-KDD, and CIC-IDS-001 data sets are comparable to those reported in the literature. In the study, a categorization process was made for the UNSW-NB15 data set. As a result, the performance rates achieved for the UNSW-NB15 dataset in all classifiers are ahead of studies in the literature [11]

III. Methodology

3.1 Datasets and features

- KDDcup99 was produced for IDS evaluation and incorporates four types of attacks such as DoS, R2L, U2R, and probing. Dataset contains 41 features describing network connections, labeled as normal or specific attack types. It is high-dimensional and imbalanced.

- CIC-IDS Dataset: Comprises 30 features related to consumer demographics and financial behavior, labeled as default or non-default. It is relatively balanced and smaller in size.

3.2 Classifier

Classification is the way toward foreseeing the lesson of given information focuses. Every strategy grasps a learning calculation to recognize a show that best to the relationship between the preparing information and the testing information [12]. Following diagram shows machine learning classifiers.

Figure 1: ML classifiers

- Naïve Bayes (NB) : Based on Bayes' theorem, a probabilistic machine learning algorithm known as a Naive Bayes classifier is used for classification tasks. It makes the "naive" assumption that all features in a dataset are independent of one another, allowing for quick and accurate predictions. It is especially useful for text classification and spam filtering; it is regarded as a straightforward and simple supervised learning model.
- Decision Tree (DT): Supervised learning algorithm utilized for regression and classification issues. It is depicted as a tree structure with each leaf node representing a class label or a predicted value, each internal node representing a test on an attribute, and each branch representing the test's outcome.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

- Random Forest (RF): The ensemble learning technique known as random forests is used for classification, regression, and other tasks. It works by creating a large number of decision trees during training. For classification tasks, the output of the random forest is the class selected by most trees.
- Support Vector Machine (SVM): SVM is a type of supervised learning algorithm used in machine learning to solve classification and regression tasks. SVMs excel at solving binary classification problems, which require dividing a data set into two distinct groups.
- Gradient Boosting Machine (GBM): GBM is a method of machine learning that targets pseudo-residuals rather than residuals in a functional space, unlike traditional boosting [13].

3.3 Performance Metrics

- Accuracy: Accuracy shows how often a classification machine learning model is correct overall.
- Precision shows how frequently a machine learning model correctly predicts the target class.
- Recall shows whether an machine learning model can find all objects of the target class.
- F1-score serves as a crucial evaluation metric frequently utilized in classification tasks to assess the effectiveness of a model.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): is a vital instrument for assessing the effectiveness of binary classification models. It graphs the True Positive Rate (TPR).
- Training and Inference Time: Inference time computation pertains to the computational resources needed to generate predictions from a trained model. In contrast to training a model, which requires analyzing extensive datasets to identify patterns and connections, inference is the phase where the model applies its learned knowledge to make predictions on fresh, unseen data[14].

IV. Experimental Result and Discussion

Experimental Setup

Data preprocessing included normalization, handling missing values, and encoding categorical features. Hyper-parameter tuning was performed using grid search with cross-validation. Each model was trained and tested using an 80-20 train-test split.

Results

The positive impact of the connected strategy on KDDCup99 and CIC-IDS datasets utilized in the think about can be seen in Table 1 and Table 2. There are numerous writings considers utilizing KDDCup99, CIC-IDS datasets. In this area, the KDDCup99 and CIC-IDS datasets are compared with the Literature study

KDDCup99 Dataset

The following results are generated by the models on the training set.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

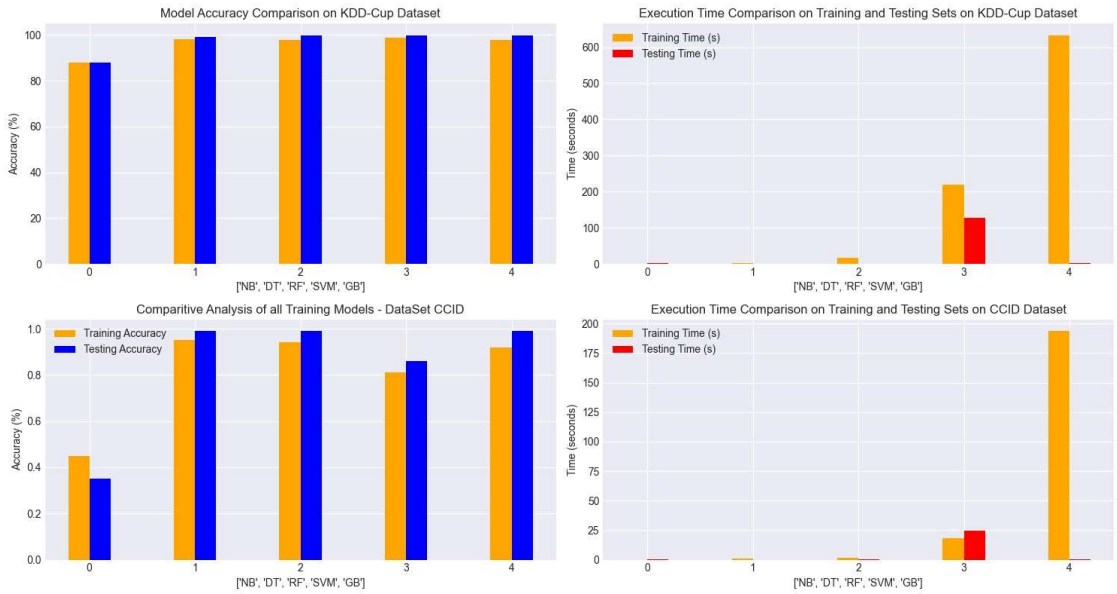
Mode	Naïve Bayes (NB)	Decision Tree (DT)	Random Forest (RF)	Support Vector Machine (SVM)	Gradient Boost (GB)
Training Set	87.95	98.05	97.99	98.87	97.79
Testing Set	87.90	99.05	99.96	99.87	99.77

Table 1: Table show the accuracy percentage in predicting attack by ML models on training and testing set of KDDCup99 Dataset

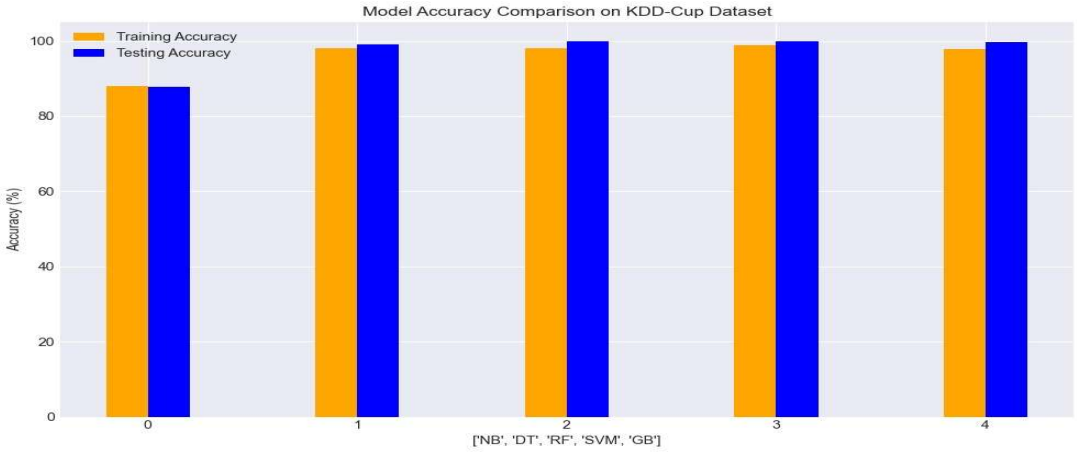
CIC-IDS Dataset

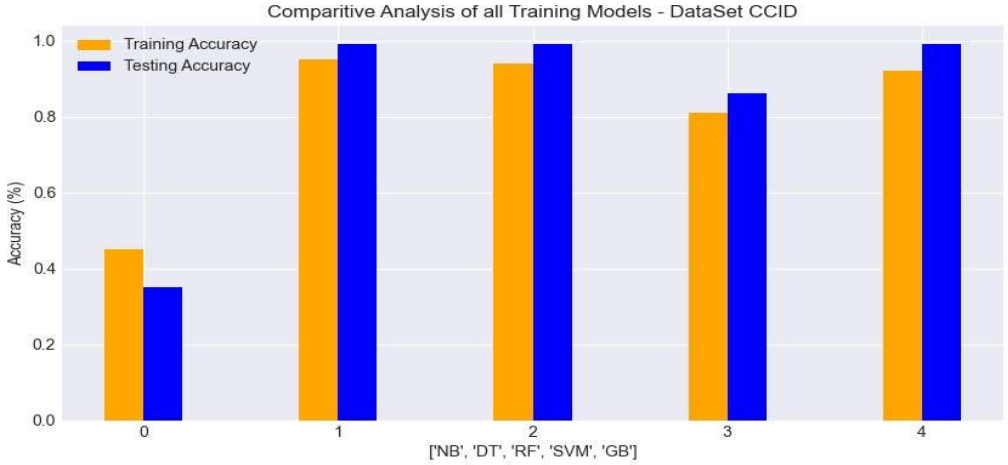
Mode	Naïve Bayes (NB)	Decision Tree (DT)	Random Forest (RF)	Support Vector Machine (SVM)	Gradient Boost (GB)
Training Set	35.46	99.96	99.64	86.51	99.72
Testing Set	35.95	99.40	99.18	86.21	99.51

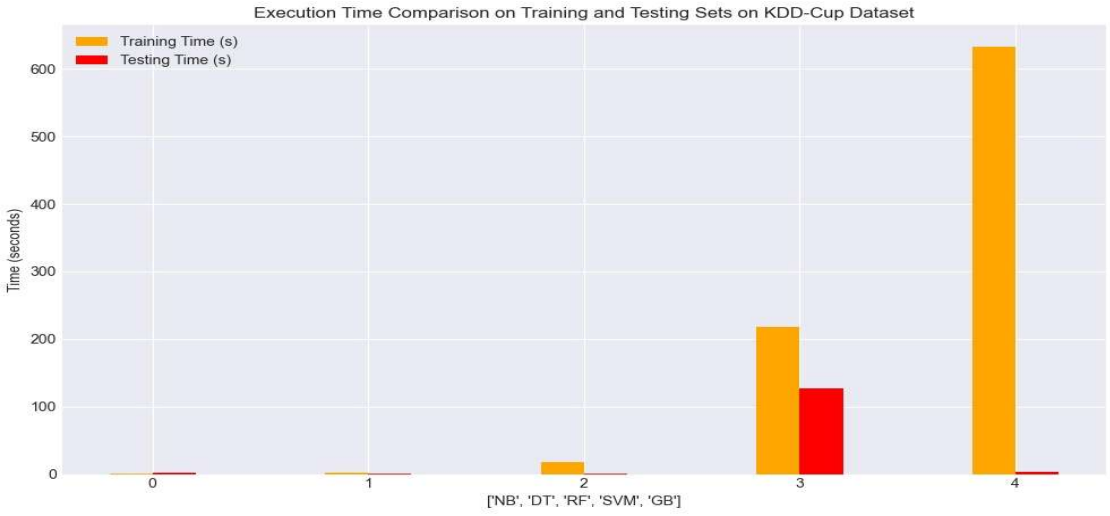
Table 2: Table show the accuracy percentage in predicting attack by ML models on training and testing set of CIC-IDS Dataset



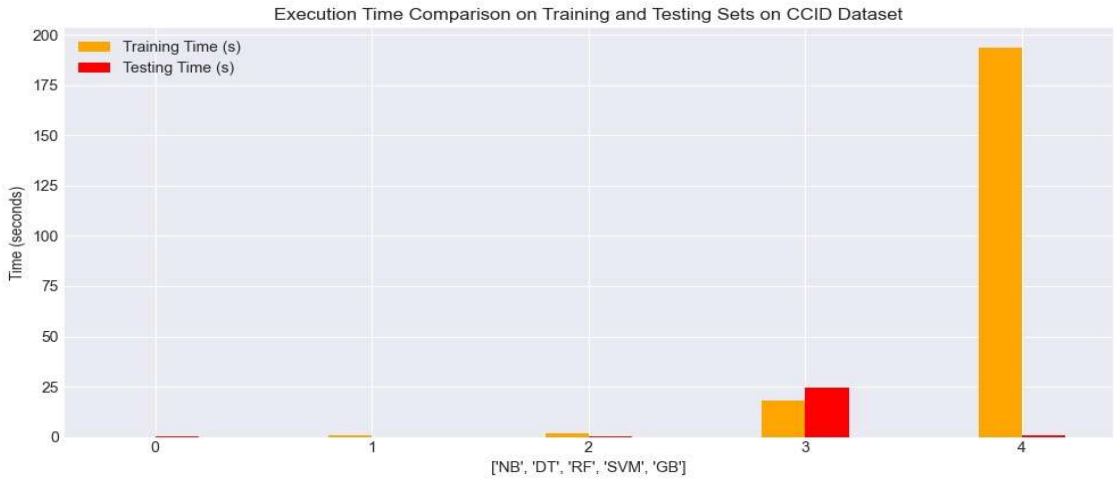
ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)







ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)



V. Discussion and Conclusion

Performance trends across both datasets underscore the importance of dataset characteristics. RF and GB consistently outperformed, algorithms LR and DT excelled in accuracy but at the cost of increased computational demand[15]. SVM performed well on smaller datasets but struggled with scalability in larger datasets. On KDDCup dataset Naïve Bayes achieved moderate accuracy i.e. 87.95% but struggled with imbalanced data, decision tree overfit on the training data, resulting in reduced generalizability i.e. accuracy is 99.05%. Random Forest performed well with high accuracy i.e 99.99% and robust handling of imbalances. SVM: Struggled with scalability due to dataset size, achieving 98.87% accuracy.
On CIC-IDS dataset GBM delivered good performance with 99.79%. Naïve Bayes demonstrated very poor performance with 45.40% accuracy. Decision Tree Showed the highest accuracy of 95%. Random Forest depicted a balanced accuracy of 94.52%. SVM achieved relatively fair accuracy of 81.84%. and GBM performed with 92.78% accuracy

This study highlights that model selection should be guided by dataset properties. Given methods are robust for diverse datasets, while suitable for scenarios demanding the highest accuracy. Future work will focus on exploring automated ML (AutoML) tools and domain-specific feature engineering for enhanced performance.

VI. References

1. Imad Bouteraa, Makhlouf Derdour, Ahmed Ahmim, “Intrusion Detection using Data Mining: A contemporary comparative study” 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS) 2018 DOI:10.1109/PAIS.2018.8598494.
2. Uma, M., & Padmavathi, G., “A Survey on Various Cyber Attacks and their Classification” IJ Network Security, 15(5), 390-396, 2013
3. Li, X., Smith, J. D., Dinh, T. N., & Thai, M. T., ” Privacy issues in light of reconnaissance attacks with incomplete information.” IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 311-318). IEEE. October 2016
4. Hussain, A., Heidemann, J., & Papadopoulos, C., “A framework for classifying denial of service attacks”, Conference on Applications, technologies, architectures, and protocols for computer communications (pp. 99-110) August 2003

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

5. Forcht, K. A., Kieschnick, E., Thomas, D. S., & Shorter, J. D., " Identity Theft: The Newest Digital Attack. Issues in Information Systems" 8(2,297-302).2007
6. https://en.wikipedia.org/wiki/Cyber_spying#Examples, accessed 20-5-2021
7. Bhardwaj, A., & Goundar, S., "Keyloggers: silent cyber security weapons. Network Security" 2020(2), 14-19.
8. Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E.," Understanding nonmalicious security violations in the workplace: A composite behavior model. Journal of management information systems" 28(2), 203-236., 2011
9. Simmons, C., Ellis, C., Shiva, S., Dasgupta, D., & Wu, Q. "AVOIDIT: A cyber attack taxonomy" 9th Annual Symposium on Information Assurance (ASIA'14) (pp. 2-12). Jun2014
10. Chaouki Khammassi, Saoussen Krichen,,"A GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection " c o m p u t e r s & s e c u r i t y 7 0 (2 0 1 7) 2 5 5 – 2 7 7 Elsevier 2017
11. Ilhan Firat Kilincer a, Fatih Ertam b,*, Abdulkadir Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study" Computer Networks 188 (2021) 107840
12. V. Kanimozhi, Dr. T. Prem Jacob, "CALIBRATION OF VARIOUS OPTIMIZED MACHINE LEARNING CLASSIFIERS IN NETWORK INTRUSION DETECTION SYSTEM ON THE REALISTIC CYBER DATASET CSE-CIC-IDS2018 USING CLOUD COMPUTING" International Journal of Engineering Applied Sciences and Technology, 2019 Vol. 4, Issue 6, ISSN No. 2455-2143, Pages 209-213 Published Online October 2019 in IJEAST (http://www.ijeast.com)
13. Mrutyunjaya Panda, Ajith Abraham, Swagatam Das, Manas Ranjan Patra, "Network intrusion detection system: A machine learning approach"13th International Conference on Ambient Systems, Networks and Technologies (ANT) March 22-25, 2022,
14. Emad E. Abdallah *, Wafa' Eleisah, Ahmed Fawzi Otoom, "Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey" Procedia Computer Science Elsevier Volume 201, Pages 205-212, 2022 https://doi.org/10.1016/j.procs.2022.03.029
15. Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.