WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

NETWORK INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

V.RAMA LAKSHMI [1] , SWARNA RAMYA [2] , VULLINGALA LOHITHA NAGA SAI [3] , PAMIDIPALLI KIRAN MANOGNA [4]

Assistant Professor , Dept of CSE , PSCMRCET [1] , ramalakshmi@pscmr.ac.in
Assistant Professor , Dept of CSE , PSCMRCET [2] , ramyaswarna6@gmail.com
Student , Dept of CSE , PSCMRCET [3] , lohithanagasai06@gmail.com
Student , Dept of CSE , PSCMRCET [4] , kiranmanogna@gmail.com

ABSTRACT

In these days all the services are available on internet and malicious users can attack client or server machine through this internet and is to avoid such attack request IDS (Intrusion Detection System)will be used ,IDS will guide request data and then check if it contains normal or attack signatures then request will be dropped or stopped. In the previous approach they use Naïve Bayes and Random Forest algorithms. but the accuracy rate will be very less. So to overcome that we use SVM and ANN algorithms. Upon receiving new request IDS will apply that request on that train model to predict it class whether the request belongs to normal class or attack class. During these process the accuracy will be increased by >5.77% compare to previous approach. To train such models and prediction various data mining classification or prediction algorithm will be used.

Keywords–SVM(Support Vector Machine), Algorithm, Random forest, IDS(Intrusion Detection System), Machine learning, Naïve bayes, Classification, ANN(Artificial Neural Networks)

Introduction

An **Intrusion Detection System (IDS)** is a system that monitors network traffic for suspicious activity and issue alerts when such activity is discovered. It is a software application that scans a network or a system for harmful activity or policy breaching. With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies.

IDS detects attacks from a variety of systems and network sources by collecting information and then analyzes the information for possible security breaches. The existing system is like while network IDS that works based on signature have seen commercial growth and widespread adoption of technology-based organization throughout the globe. The main problem is that sometimes false positive signatures will be high in this approach. This is particularly an issue with zero day or emerging threats that rely on new exploits and attack techniques that IDS is unfamiliar sometimes. The promise and the contribution machine learning did till today are fascinating, There are many real-life applications we are using today offered by machine learning. It seems that machine learning will rule the world coming days. Here we developed a supervised machine learning model that can classify unseen network traffic based on what is learnt from the seen traffic. We used both SVM and ANN learning algorithm to find the best classifier with high accuracy and success rate. During these process the analyses the large volume of network data and considers the complex properties of attack behaviours to improve the performance of detection speed and detection accuracy.

According to previous research , most of the proposed systems worked on Signature-based method. As we all know that Signature-based IDS detects the attacks on the basis of the specific patterns such as number of bytes in a network traffic. In these approach it is difficult to detect the new malware attacks as their patterns like signatures is not identified. So I decided to choose Anamoly-based method to detect the unknown malwares. Basically it also uses machine learning to create trustful activity model and anything coming is compared with that model and it declares if any suspicious activity if it is not found in model. Generally in machine learning Anomaly-based method has a better generalized property in comparision to signature-based method.

Types of IDS

The IDSs are mainly categorized as follows 1.Network Intrusion Detection System (NIDS) 2.Host Intrusion Detection System (HIDS) 3.Protocol-based Intrusion Detection System (PIDS) 4.Application Protocol-based Intrusion Detection System (APIDS) 5.Hybrid Intrusion Detection System (HIDS)

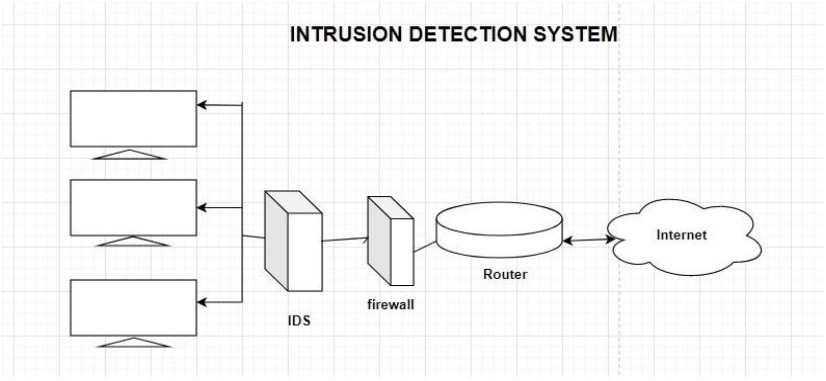
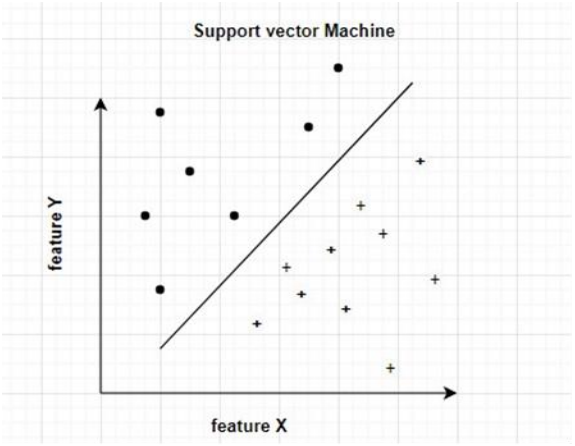


Fig : Intrusion Detection System



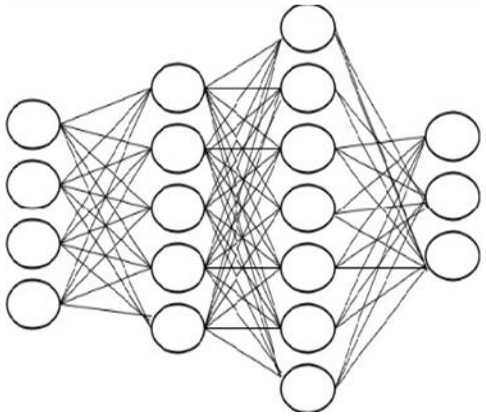


Fig :Support Vector Machine Fig :Artificial Neural Network

Literature Survey

Rachid Tahri et.al uses classification algorithms like SVM (Support Vector Machine), NB(Naïve Bayes), KNN (K-Nearest Neighbor) and compared performances of previous approaches. Two different datasets like NSL-KDD and UNSW-NB 15 are used to measure the performance of the proposed approach in order to guarantee of efficiency of project. The result of study showed that classifiers were much better than those who trained with KDD'99.

Zeeshan Ahmed et.al explains by using algorithms like decision tree, KNN (K-Nearest Neighbor) ,K-means clustering techniques. During these process both discrete and continuous traffic characteristics could be extracted using these techniques. They provided various research challenges and provide future scope for the research improving ML and DL based NIDS. The result of study includes that 60% of proposed methodologies were tested by KDD' Cup and NSL-KDD. And it gives quite accuracy than previous approach.

Bello Nazifi Kagara et.al implemented an intrusion detection method with the range of NSL-KDD dataset by splitting the input dataset into different subsets, on the basis of threats. They mainly focuses on explaining the distinguish between intrusion and regular network traffic is quite challenging. And here they use signature-based IDS which refers to a device known list of earlier attack signatures and security bugs. During these process they often find the security breaches and anomalies. To give better accuracy they use two types of datasets i.e, NSL-KDD ,KDD'99.

Ameera S. Jaradat et.al perform evalution of intrusion detection based on machine learning approach. Here they use MATLAB environment is utilized to try variety of algorithms. The experiment conducted by using CICIDS-2017 dataset. The main aspect of conducting this experiment is to build intrusion detection systems with higher accuracy and precesion.

Thomas Rincy N et.al proposed by using machine learning techniques like Bayesian Network,
Hidden Markov Model, KNN(K-Nearest Neighbor), K-means Clustering. Here they utilize the Hybrid NID-Sheild with relates to the characteristics of UNSW-NB15 and NSL-KDD datasets. As we all know that hybrid NID-Sheild NIDS is most efficient of all approaches found in the existing literature studies. And it obtains a comprehensive excellent performance among all other metrics.

Patrick Vanin et.al proposed an improved solution for network intrusion detection system.In this process they use deep learning techniques like CNN(Convolutional Neural Networks), DNN(Dense Neural Networks). And they uses datasets of KDDcup99 and NSL-DD. Recent studies shows that limitation of NIDS to detect is zero-day attacks. During process of executing all the solutions were tested using public dataset. It gives best solution comparing with existing methods.

Lirim Ashiku explains about IDS by using both machine and deep learning techniques like Random
Forest, Decision trees, SVM. By this process they adopted the datasets from KDDCup99 , UNSW-NB15 NSL-DD. These model uses keras library as a prototype working on top of tensor flow framework. To increase the optimization momentum, they use deep learning models on GPU enabled framework of Google colab. Hence preserving the best accuracy than the existing approach.

Usman Shuaibu Musa et.al has used anamoly-based approach to explain about HIDS(Host-based Intrusion Detection System) and NIDS(Network-based Intrusion Detection System). We all know that most of the previous approaches should done by using signature-based approach . In this process the datasets is taken from The main aspect is that in signature based approach the accuracy limit is very less so in this approach to improve the efficiency he chooses Adaboost (Adaptive boosting) learning algorithm is applied on NSL-KDD dataset.

Hafiza Anisa Ahmed et.al has explained Intrusion Detection \System using oversampling technique and machine learning algorithms. In this process the datasets is taken from KDDCUP99 ,KDD98,NSL-KDD7. Consequently ,due to revolution of network traffic, and traffic data available on those dataset is different fom modern-day existing data traffic. Here they use machine learning algorithms like Random Tree, Logistic Regression(LR),Artificial neural networks. Here they use novel combination of different preprocessing techniques inorder to resolve all the underlying issues of the dataset and develop fast and efficient network security intrusion detection system.

Emad E. Abdallah et.al introduce intrusion detection using supervised machine learning method by using feature selection technique. Main goal is provide a taxonomy for linked IDS by using supervised learning approach. Here they use popular datasets like KDD'99,NSL-KDD,CICDS2017and UNSW-NB15. Finally , for good performance they used large intrusion detection datasets form deep learning technique. However the support vector machine and random forest perform very well on this datasets.

Table

Author Name	Algorithm	Merits	Limitations	Accuracy
Rachid Tahri	**SVM(Support Vector Machine)** **KNN(K-Nearest Neighbor)**	**SVM is used to handle huge amount of data.** **KNN is simple and we can understand easily.**	**SVM works very slow.** **KNN is quite expensive compare to others.**	**SVM = 97.777%** **KNN = 93.333%**
Zeeshan Ahmed	**K-means Clustering** **KNN(K-Nearest Neighbor)**	**K-means mainly fit for numerical data.** **KNN can solve problems efficiently.**	**K-means always gives Spherical clusters.** **Sometimes KNN will throw storage issues.**	**K-means = 63.57%** **KNN = 77.06%**
Bello Nazifi Kagara	**NB(Naïve Bayes)** **SVM(Support Vector Machine)**	**Naive Bayes is simple to implement** **SVM has high flexibility.**	**Naive bayes can encounter zero frequency problems.** **SVM will only works for complete dataset without any redunancies**	**NB = 81.74%** **SVM=89.46%**

Ameera S. Jaradat	**SVM(Support Vector Machine) K-means Clustering**	**SVM gives high Quality results. K-means is used to partitioning datasets in groups.**	**SVM doesn't work for large dataset. K-means always requires to specify the clusters before perform or execute.**	**SVM = 90.81% KNN = 91.00%**
Thomas Rincy N	**KNN(K-Nearest Neighbor) K-means Clustering**	**KNN has non-parameter approach so we can easily use it. K-means is simple and convenient to use easily.**	**KNN is slight slow performing while performing K-means doesn' t work for outliers.**	**KNN = 95.55% K-means = 98.1%**

Partrick Vanin	**CNN(Convolutional Neural Network) DNN(Deep Neural Network)**	**CNN has high accuracy. DNN uses multiple layers of Neural networks**	**CNN will always obtain large datasets DNN will always requires huge amount of data.**	**CNN = 87.65% DNN=84.72%**
Lirim Ashiku	**SVM(Support Vector Machine) Decision Tree**	**SVM gives faster results. Decision Tree gives more information**	**SVM can't fit for missing values. Decision Trees sometimes occur risks due to overfitting**	**SVM = 74.78% Decision Tree=72.56%**
Usman Shuaibu Musa	**Random Forest SVM(Support Vector Machine)**	**The random Forest is applicable for both classification and regression problems. SVM has high Interpretability**	**Random Forest is slightly complex compared with other. SVM consume lots of memory**	**Random Forest = 69.76% SVM = 83.26%**
Hafiza Anisa Ahmed	**Logistic Regression ANN(Artificial Neural Network)**	**Easier to implement machine learning method. ANN can handle large amount of data.**	**Sometimes LR will be underfitting or overfitting ANN has burden issues while performing with large data.**	**LR = 56.2% ANN = 71.5%**

Emad E. Abdullah	**Random Forest SVM(Support Vector Machine)**	**Random Forest is also works when it contains empty and missing values. SVM gives high accuracy.**	**In random forest the training time is high. SVM doesn't set for big datasets.**	**= Random forest 87.01% SVM = 81.79%**

Proposed Methodology

The promise and contribution machine learning did till today are fascinating. There are many real-life applications we are using today offered by machine learning . It seems that machine learning will rule the world in coming days. Hence we came out into a hypothesis that the challenge of identifying new attacks or zero-day attacks facing by the technology enabled organizations today can be overcome using machine learning techniques. Here we developed a supervised machine learning model that can classify unseen network traffic based on what is based on what is learnt from the seen traffic. We used both SVM and ANN learning algorithm to find the best classifier with higher accuracy and success rate.

Advantages of Proposed System

The new proposal was innovative as Hidden Naïve Bayes which shows more advantage then Traditional Naive Bayes.
And it analyses the large volume of network data and considers the complex properties of attack behaviors to improve the performance of detection speed and detection accuracy.

System Architecture

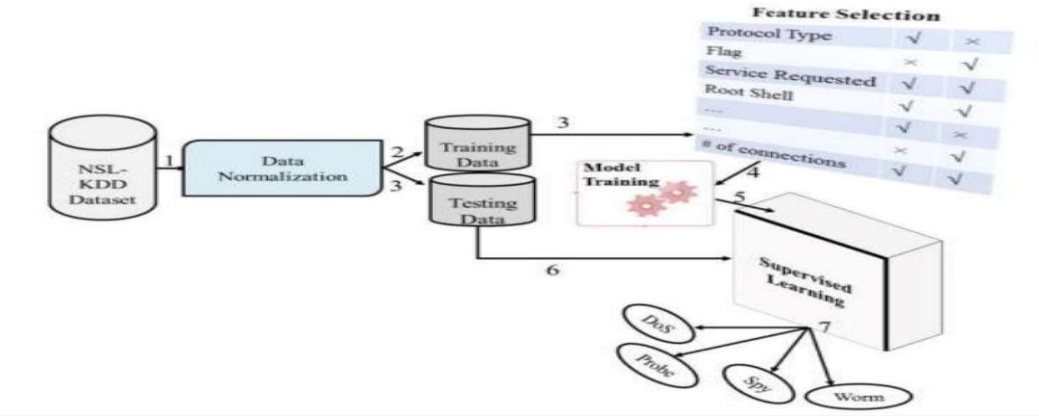


Fig : System Architecture

Working with Machine Learning

The IDS is an essential technique used for maintaining and improving network security. And as we all know that IDS-ML is an open source code repository which is written in python for developing IDSs from public network data by using traditional and advanced ML (Machine Learning) algorithms. Basically while implementing the process with Machine Learning algorithms it gives high accuracy compare to previous approaches.

Working Process for Intrusion Detection System Using Machine Learning

- These process includes seven steps to execute

3.1 Uploading Dataset
3.2 Preprocessing
3.3 Generate Training Model
3.4 Run SVM Algorithm
3.5 Run ANN Algorithm
3.6 Upload Test Data and Detect Attack

3.7 View Accuracy graph

3.1 Uploading Dataset

Initial step is to upload a valid dataset. For these different datasets are collected from Kaggle licensed software. And here we used "intrusion_dataset.txt" as main dataset which gives accurate readings.

3.2 Preprocessing

After Uploading dataset preprocessing will be done to clean dataset by remove string values and covert string attack names to numeric values such as normal signatures as '0' and anomaly signatures as '1' during these process.

3.3 Generate Training Model

After preprocessing the dataset is cleaned. We can easily generate the training model which is used to train the Algorithms like machine learning.

3.4 Run SVM Algorithm

Once training model is generated we can use the algorithms like SVM. While using SVM Algorithm we got 84.73% of accuracy.

3.5 Run ANN Algorithm

Here we use another algorithm i.e, ANN . While using ANN Algorithm we got accuracy of 96.88% of accuracy.

3.6 Upload test data and Detect Attacks

After successful completion of running both algorithms .Upload test data file which contains test records. After prediction we will get the results.

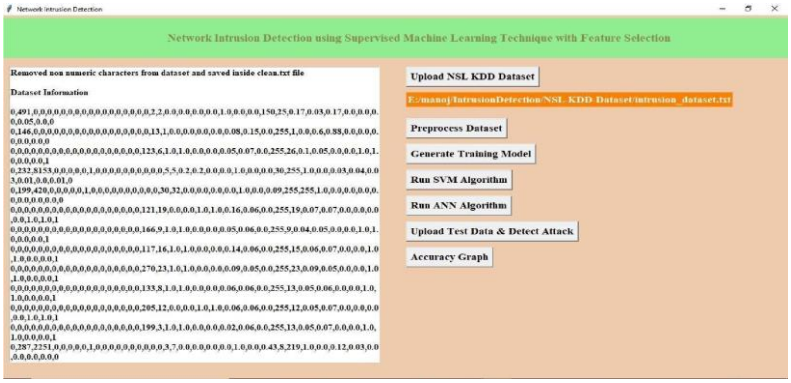
3.7 Accuracy Graph

After successful results obtains the accuracy graph is generated as follows. The finalized result will be resulted as graph.

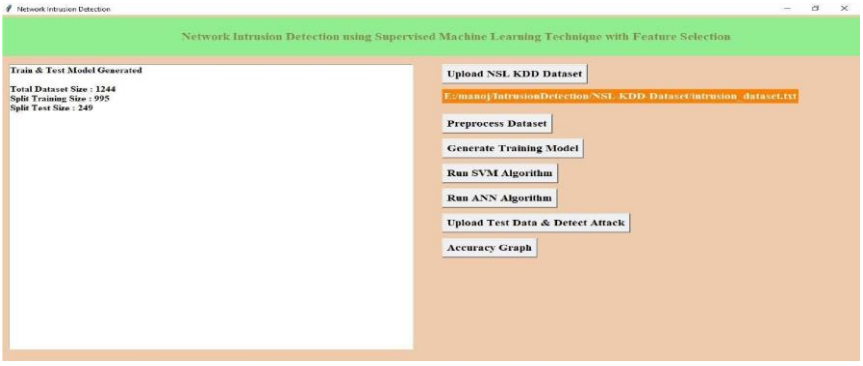
Results and Discussion

SCR 4.1: HOME SCREEN

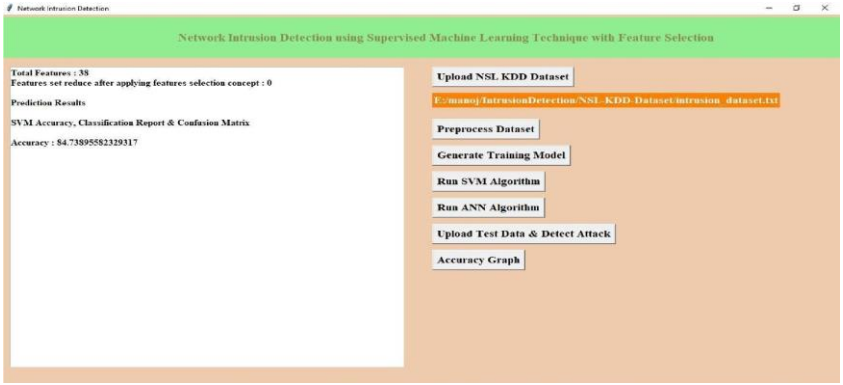
SCR 4.2: UPLOADED DATASET



SCR 4.3: PREPROCESSING THE DATA

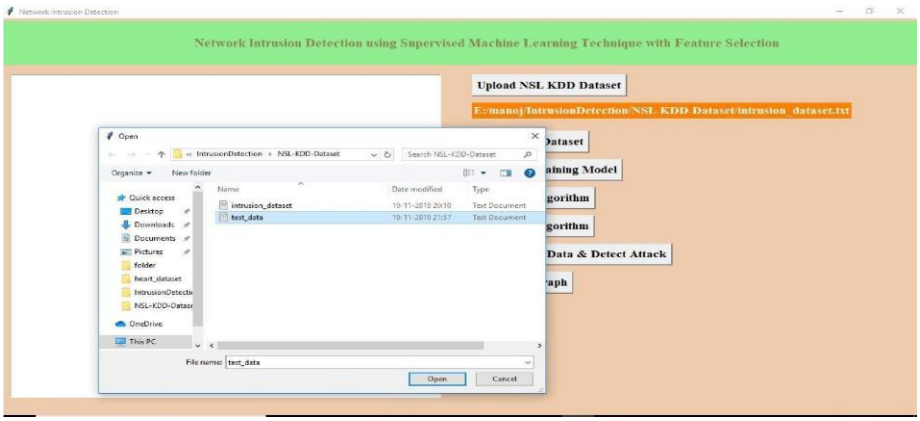


SCR 4.4: GENERATING TRAINING MODEL

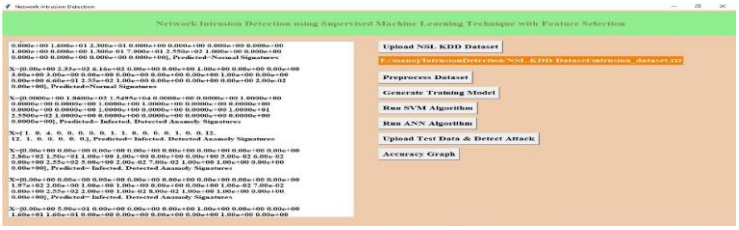


SCR 4.5: RUN SVM ALGORITHM

SCR 4.6: RUN ANN ALGORITHM



SCR 4.7: UPLOADING TEST DATA



SCR 4.8: DETECTING ATTACKS

SCR 4.9: ACCURACY GRAPH

Conclusion

In this project, we have presented different models machine learning models using different
machine learning algorithms and different feature selection methods to find a best model. The analysis of a result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%. We believe that these findings will contribute to research further in the domain of building a detection system that can detect known attacks as well as novel attacks. The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero-day attacks still remains a research topic due to high false positive rate of the existing systems.

5.1 Future Work

The IDS plays an important role in upcoming days, because the ultimate goal of IDS is to give security for the data which is highly confidential. For example if we consider an organization has some specific security policies and the information must be shared within the organization and no third person from outside shouldn't allows and access the info or make any attacks between network. As we all know that the total expected cost of an IDS is the sum of consequential and operational costs. Here we use two different machine learning approaches to reduce operational cost.

5.2 Limitations

Every approach and models have some limitations and even here there are some limitations as follows
The cost metrics changes model to model.
If we use new approach or model we have to restart or reconstruct from initial step to final step.
Sometimes the re-training is applicable if the adopted resource is not fit for an approach or a model.

REFERENCES

1. Tahri, R., Balouki, Y., Jarrar, A., & Lasbahani, A. (2022). Intrusion Detection System Using machine learning Algorithms. In M. Sbihi, A. Mounadi, & M. Garoum (Eds.), ITM Web of Conferences (Vol. 46, p. 02003). EDP Sciences. https://doi.org/10.1051/itmconf/20224602003

2. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2020). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies. https://doi:10.1002/ett.4150

3. Kagara, B. N., & Md Siraj, M. (2020). A Review on Network Intrusion Detection System Using Machine Learning. In International Journal of Innovative Computing (Vol. 10, Issue 1). Penerbit UTM Press. https://doi.org/10.11113/ijic.v10n1.252

4. Jaradat, A. S., Barhoush, M. M., & Easa, R. S. B. (2022). Network intrusion detection system: machine learning approach. In Indonesian Journal of Electrical Engineering and Computer Science (Vol. 25, Issue 2, p. 1151). Institute of Advanced Engineering and Science. https://doi.org/10.11591/ijeecs.v25.i2.pp1151-1158

5. Rincy N, T., & Gupta, R. (2021). Design and Development of an Efficient Network Intrusion Detection System Using Machine Learning Techniques. In P. Fournier-Viger (Ed.), Wireless Communications and Mobile Computing (Vol. 2021, pp. 1–35). Hindawi Limited. https://doi.org/10.1155/2021/9974270

6. Vanin, P., Newe, T., Dhirani, L. L., O'Connell, E., O'Shea, D., Lee, B., & Rao, M. (2022). A Study of Network Intrusion Detection Systems Using Artificial Intelligence/Machine Learning. In Applied Sciences (Vol. 12, Issue 22, p. 11752). MDPI AG. https://doi.org/10.3390/app122211752

7. Ashiku, L., & Dagli, C. (2021). Network Intrusion Detection System using Deep Learning. In Procedia Computer Science (Vol. 185, pp. 239–247). Elsevier BV. https://doi.org/10.1016/j.procs.2021.05.025

8. Musa, U. S., Chhabra, M., Ali, A., & Kaur, M. (2020). Intrusion Detection System using Machine Learning Techniques: A Review. In 2020 International Conference on Smart Electronics and Communication (ICOSEC). 2020 International Conference on Smart Electronics and Communication (ICOSEC). IEEE. https://doi.org/10.1109/icosec49089.2020.9215333

9. Ahmed, H. A., Hameed, A., & Bawany, N. Z. (2022). Network intrusion detection using oversampling technique and machine learning algorithms. In PeerJ Computer Science (Vol. 8, p. e820). PeerJ. https://doi.org/10.7717/peerj-cs.820

10. Abdallah, E. E., Eleisah, W., & Otoom, A. F. (2022). Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey. In Procedia Computer Science (Vol. 201, pp. 205–212). Elsevier BV. https://doi.org/10.1016/j.procs.2022.03.029