IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12719

Intrusion Detection System Using Ensemble Learning Approaches

J. Vimal Rosy[1*] and Dr. S. Britto Ramesh Kumar[2]

Head & Assistant Professor, Soka Ikeda College of Arts & Science for Women, Chennai, St. Joseph's College (Autonomous), Trichy[1]

Assistant Professor, Soka Ikeda College of Arts & Science for Women, Chennai, St. Joseph's College (Autonomous), Trichy[2]

Abstract: A lot of information systems are protected by and have damage minimized by intrusion detection systems. It defends computer networks, both virtual and physical, from dangers and weaknesses. Machine learning methods are currently being widely expanded to create efficient intrusion detectionsystems. Machine learning techniques for intrusion detection include rule learning, ensemble approaches, statistical models, and neural networks. Machine learning ensemble approaches stand outamong them for their effectiveness in the learning process. This study aims to increase detection rate accuracy for all attack kinds and individual attack types, which will aid in the identification of attacksand specific categories of attacks. K-fold cross validation is used to assess the suggested approach, and the experimental outcomes of all three classifiers are examined. UNSW-NB15 dataset is used tomeasure the performance of the proposed approach in order to guarantee its efficiency.

Keywords: Ensemble Learning, Network Intrusion Detection, , Multi-classification, Random Forest.

I. INTRODUCTION

IDS are referred to as a software tool that provides management-level experiences by identifying potentially dangerous system actions. A security measure to identify intrusions that endanger the familiarity, availability, and integrity of data sources is an intrusion detection system [1]. The groupsare utilizing IDS with the goal of identifying issues with security procedures and cataloging current threats. Since it was first created with the most amount of network traffic data possible, traffic analysis of networks has grown increasingly extensive due to the development of internet technologies, applications, and protocol.

Anti-hazard application software, such as antivirus software, firewalls, and spyware detectionprograms, are installed on each computer linked via a network with two-way access to the external environment, such as the internet, in HIDS (Host based Intrusion Detection System). It takes a snapshot of the process documents and compares it to the previous snapshot. If we put a firewall nextto it, even if they both provide protection, the IDS architecture differs from the firewall. Firewall limits the way approaches are taken between strategies to prevent interceptions and to stop an attack from being flagged along the process. As soon as a suspected interference occurs, an IDS analyzes itand flags a warning. A substructure known as an interference counteractive action substructure endsrelationships.

Cyber Security

In order to reduce the danger of cyber intrusion and safeguard network components and data againstassaults and malicious events, a collection of procedures, tools, and systematic rules and practices areused. The purpose of cyber-security is to reduce risks and provide network environments with safetyand privacy (Babar, 2018). In order to maintain information confidentiality, information integrity, and information availability (CIA), a great deal of work is put forth by cyber-security experts and professionals in building a variety of cyber-defense systems and technologies (Du, 2016). A generalized view of network intrusion detection systems (NIDSs) in the cyberenvironment is shownin Figure 1.1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12719

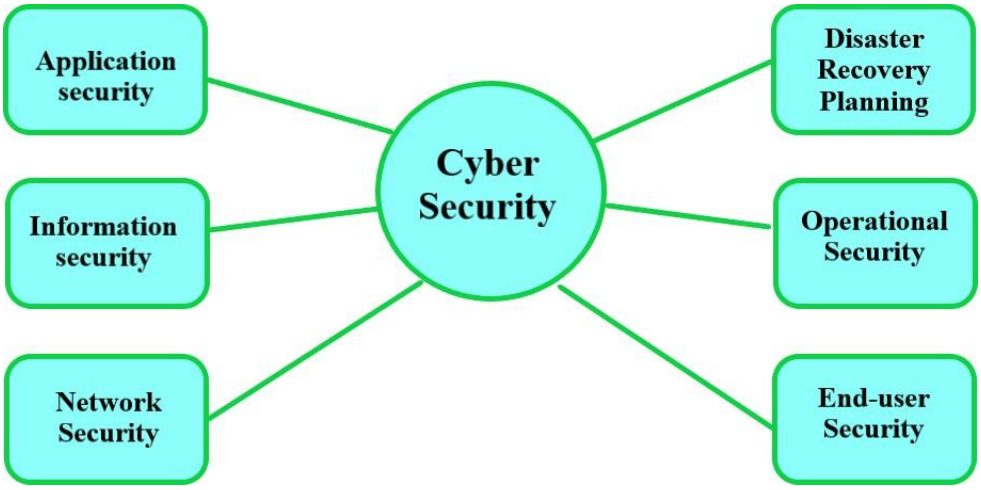

Fig 1.1 Elements of Cyber Security

Additionally, the IDS can be divided into two categories based on where it is deployed; these categories are host-based and network-based [2]. The network-based IDS (NIDS) tends to capture intruders at the network level, whereas the host-based IDS (HIDS) is put in the area that is closer to the host in order to capture the intruders . The majority of cloud storage today uses software definedtechnologies to increase its usability and dependability for all application services. This hybrid environment increases the likelihood of a large number of malicious assaults being launched [2].

The anomaly-based IDS analyzes the current system using a specified normal profile to find the deviation. However, the problem of false alarms in real-world implementation affects it [3]. By integrating anomalies-based and signature-based detection, the hybrid IDS offers security. Byutilizing adaptive algorithms, the problem of intelligent false alarm technique is overcome. [4]

The paper organization is given as follows. Section II describes the related work, and Section III describes the proposed methodology. The experimental results of the proposed intrusion detection are given in Section IV. Section V describes the significant aspects of the proposed methodology and conclusion.

Machine Learning:

Learning is the process of creating a scientific model based on patterns gleaned from transactions ina training set (Guven, 2016). Machine learning is a sophisticated computing method that uses provided data to automatically identify patterns that may be relevant to decision-making. Accordingto Du (2016) and Guven (2016), machine learning algorithms can be roughly divided into supervised,unsupervised, and reinforcement learning.

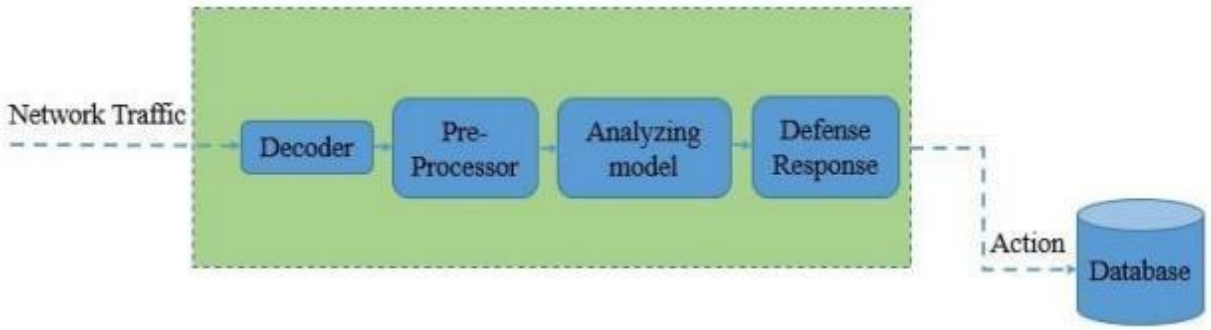

Fig 1.2 Machine Learning Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12719

II. RELATED WORKS

Using classification techniques such Naive Bayes, K-means, neural networks, RF, SVM, and DT, Abhishek Divekar et al. (A. Divekar, 2018) analyzed the results for different KDD'99 alternatives. They discovered that the UNSW-NB15 is a superior and more recent replacement for the KDD'99. The study's findings demonstrated that classifiers trained using the f1-score performed significantly better than those trained using KDD'99 and NSLKDD.

A survey of several deep learning and machine learning-based intrusion detection systems has been provided by Anns Issac et al. (ANN, 2022). The proposed system adopts an ensemble learning approach to improve accuracy.

An efficient method was created by S. Kejriwal et al. (KEJ, 2022) to recognize any unauthenticated activity on the inside or outside. Some of the models that have been examined (MLP) include a Multi-Layer Perceptron Classifier, a Logistic Regressor, a Random Forest Classifier, a K Nearest NeighborClassifier, an XGBoost Classifier, and a Gaussian Naive Bayes Classifier. In this particular use case,the Random Forest Classifier model produced the best results, with an accuracy of 99.8% and a macroaverage F1-Score of 0.98.

The best features for the UNSW-NB15 dataset were chosen using the Sine Cosine feature selection algorithm, which was developed by Vimalrosy et al. Once an assault has been detected, the proposedNovel CVAR-k fold Cross verified Artificial neural network weighted Random Forest classification is used to make the proper classification. The proposed method exceeds the existing strategies to classify and detect different assaults, as demonstrated by the proposed system SC-CVAR's maximumaccuracy rate of 0.9987 for the UNSWNB15 Dataset and detection rate. So the system is quickly andsuccessfully repaired.

Vimalrosy et al. optimized the Sine Swarm to find the best features for the UNSW-NB15 dataset. TheRandom forest classification method is used to make the correct classification. 98.15% accuracy wasreached with the suggested system, OSS-RF.

III. DATASET DESCRIPTIONS

According to (Slay N. M., 2016), UNSW-NB15 was developed in the Australian Centre for Cybersecurity's cyber range lab in 2015. One of the dataset's formats is CSV files. The original CSVfiles, which were divided into four files and had more than 2.5 million records, are not used.

Since the training and testing sets of the polished CSV files contain 175,341 transactions and 82,332entries, respectively, we are employing them in our study. The dataset has 47 features, encompassingcategory, nominal, and numeric data types. It is a multi-class, binary labelled dataset. Table 1 displaysthe distribution of each assault in training and test sets.

IV. PROPOSED METHODOLOGY

Three steps make up the design process for the suggested scheme: normalization, featureselection, and ensemble technique.

Table 1. Number of records in training and testing subsets for each class

Classes	Training Subset	Testing Subset
Normal	56,000	37,000
Analysis	2,000	677
Backdoor	1,746	583
DoS	12,264	4,089
Exploits	33,393	11,132
Fuzzers	18,184	6,062
Generic	40,000	18,871
Reconnaissance	10,491	3,496
Shellcode	1,133	378
Worms	130	44
Total Number of Records	175,341	82,332

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12719

Normalization: To minimize the dimensionality of the dataset and prepare frameworks for theprimary analysis using all 41 features, the following processes are taken.

Where z' is normalized value and z is initial value. Max and min value for attribute A beforenormalization.

Step 1: 47 feature UNSW-NB15 dataset.
Step 2: Dataset normalization using a formula.
Step 3: Choosing Features Based on Information Gain.
Step 4: Enter the feature and labels into Random forset, Adaptive Boost, and Naive Bayes to create three models.
Step 5: Test these models, then determine their precision, recall, and accuracy.
Feature Selection: The Filter Method ranks the features based on statistical methods. Following therating, the user can select the top 24 or 47 features for the experiment, depending on their needs. In the third study, feature selection is carried out using Information Gain to select the best componentsrather than using all 47 inputs. The trial is then run using straight Nave Bayes, Random forest and Adaptive Boost, and the results are examined. We choose the correlation method because it reduces the unbalanced distribution of data while representing the similarity of characteristics based on distribution. The stages that follow during this feature selection phase are depicted in steps.

Step1: UNSW-NB15 dataset with 47 features.
Step2: Normalization of Data set with help of formula.
Step3: Feature Selection by Information Gain.
Step4: Input the feature & labels into Random Forest, Naïve Bayes & Adaptive Boost and make threemodels.
Step5: Perform the test on these models and calculate the precision, recall & accuracy.

Ensemble Approach : In the third study, the Ensemble Approach is used to perform the experimentusing Random forest, Naive Bayes and Adaptive boost while selecting a few of the best componentsas opposed to using all 47 include [2]. We employ the bagging approach because it illustrates the similarity of characteristics based on distribution while minimizing imbalance, biasing, and varianceof features in form of distribution. Using an ensemble technique, we averaged or aggregated the results of numerous classifiers. The variance error is lessened by bagging. The Bagging can be used as a bootstrap aggregation as well. We select 'n' observations or 'n rows' from the original dataset forbootstrapping. However, the primary determining element is replacing each row with a chosen original dataset so that every iteration's probability of choosing a row is the same for every row. Thesampling Technique is comparable to this. The following steps occur during the ensemble approach phase, as depicted in Figure 1.

Step 1: 47 features UNSW-NB15 dataset.
Step 2: Using a formula, normalize the data set.
Step 3: Choosing Features Based on Information Gain.
Step 4. Enter the feature and labels into the Random Forest and Naive Bayes models.
Step 5: Run the ensemble models test and determine the precision, recall, and accuracy.

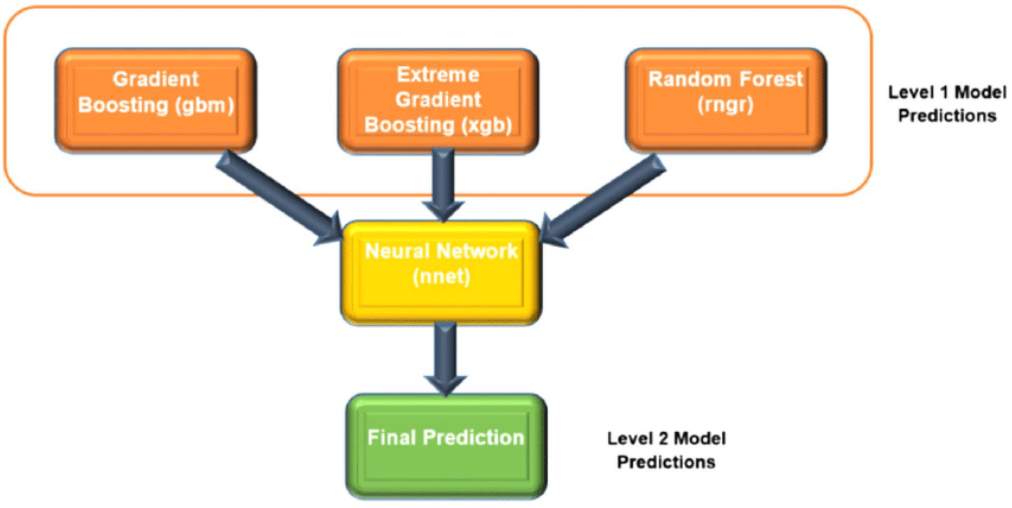

Fig1.3 Ensemble Learning Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12719

Ensembles can give us boost in the machine learning result by combining several models. Basically, ensemblemodels consist of several individually trained supervised learning models and their results are merged in various ways to achieve better predictive performance compared to a single model. Ensemble methods can bedivided into following two groups: Sequential ensemble methods As the name implies, in these kind of ensemble methods, the base learners are generated sequentially. The motivation of such methods is to exploitthe dependency among base learners. Parallel ensemble methods As the name implies, in these kind of ensemble methods, the base learners are generated in parallel. The motivation of such methods is to exploit the independence among base learners.

Ensemble Learning Methods The following are the most popular ensemble learning methods i.e. the methods for combining the predictions from different models:

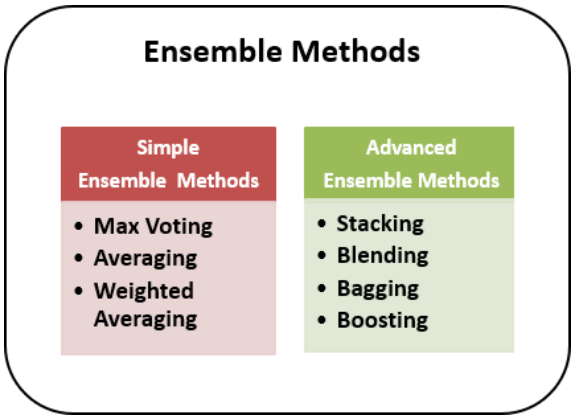

Fig 1.4 Ensemble Methods

Bagging

The term bagging is also known as bootstrap aggregation. In bagging methods, ensemble model tries to improve prediction accuracy and decrease model variance by combining predictions of individual models trained over randomly generated training samples. The final prediction of ensemble model will be given by calculating the average of all predictions from the individual estimators. One of the best examples of bagging methods are random forests.

Boosting

In boosting method, the main principle of building ensemble model is to build it incrementally by training eachbase model estimator sequentially. As the name suggests, it basically combine several week base learners, trained sequentially over multiple iterations of training data, to build powerful ensemble. During the training of week base learners, higher weights are assigned to those learners which were misclassified earlier. The example of boosting method is AdaBoost.

Voting

In this ensemble learning model, multiple models of different types are built and some simple statistics, like calculating mean or median etc., are used to combine the predictions. This prediction will serve as the additional input for training to make the final prediction

V. EXPERIMENTAL RESULTS

In this section, we recognize certain supervised learning techniques for their performance. Here, we made significant network anomaly investigations using well-identified UNSW-NB15 data [20]. We do three different trial configurations in this. The frameworks are created using all forty-seven elements in the initial study.

Classifier	Accuracy	Precision	Recall
Naïve Bayes	91.4852	97.50	91.20
Random Forest	99.9601	99.56	99.90
Adaptive Boost	96.5810	95.15	96.14

Table -2 Performance Evaluation with all 47 features for UNSW-NB15 dataset

DOI: 10.17148/IJARCCE.2023.12719

In the second study, feature selection is carried out using Information Gain to select the best components rather than using all 47, then the trial is carried out using Random Forest, Naive Bayes,and Adaptive Boost, and the results are analyzed.

Classifier	Accuracy	Precision	Recall
Naïve Bayes	91.9715	97.50	91.20
Random Forest	99.9605	99.56	99.90
Adaptive Boost	97.9043	96.10	96.14

Table -3 Performance Evaluation with applying feature selection with information gain forUNSW-NB15 dataset

In the third analysis, the ensemble approach is used by voting to select some of the best componentsrather than using all 47 included and running the experiment with Random forest, Naive Bayes, and Adaptive Boost then evaluate the results Comparing the Naive Bayes, Random Forest, and AdaptiveBoost classifiers first. Without using feature selection, Table 2 shows how Naive Bayes, Random Forest, and Adaptive Boost algorithms performed on the UNSW-NB15 data set for all 47 features. 47 features are relevant in the Random Forest outcome. Additionally, Adaptive Boost outperformed Naive Bayes in terms of results.

Classifier	Accuracy	Precision	Recall
Naïve Bayes	91.9715	97.50	91.20
Random Forest	99.9605	99.56	99.90
Adaptive Boost	97.9043	96.10	96.14
Ensemble Approach	99.9723	99.86	99.91

Table -4 Efficient analysis making use of Ensemble method for UNSW-NB15 dataset

VI. CONCLUSION

The focus of this research is on creating an optimized classifier model for two classes—attackand do not attack—but the data imbalance problem is resolved using the ensemble approach. We run three different test configurations. The frameworks are available using all 47 of the key investigation'shighlights. In the second trial, we do feature selection using information gain to choose the appropriate components rather than using all 47 factors. We next conduct the experiment using NaiveBayes, Random forest and Adaptive boost and consider the outcomes. The third study involves usingthe Ensemble Approach and bootstrapping to select the best components rather than using another classifier. This analysis also involves testing Nave Bayes, Random forest and Adaptive Boost and analyzing the results . These findings show that Ensemble Approach performs on average better thanother classifiers.

REFERENCES

[1] Vimal Rosy and Dr. S. Britto Ramesh Kumar, “OSS- RF: Intrusion Detection using optimized Sine swarm based random forest classifier on UNSW-Nb 15 dataset”, International Journal of Technical & physical problems of Engineering (IJTPE) Issue : 51, Vol.14, No. 2, PP, 275-283, ISSN:207723528 June 2022.

[2] J. Vimal Rosy and Dr. S. Britto Ramesh Kumar, “SC-CVAR : Intrusion detection using Feature selection and Machine Learning Techniques on UNSW-NB15 dataset”, International Journal of Computer Science and Network Security (IJCSNS), ISSN: 1738-7906 , Vol 22, No.4 April 2022.

[3] Dua, S., & Du, X. (2016). Data mining and machine learning in cybersecurity. CRC press.

[4] Moustafa, N., & Slay, J. (2015). A hybird feature selection for network intrusion detection systems: central points and association rules. In Australian Information Warfare Conference, 5- 13.

[5] S. Kejriwal, D. Patadia, S. Dagli and P. Tawde, "Machine Learning Based Intrusion Detection," 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India,

DOI: 10.17148/IJARCCE.2023.12719

2022, pp. 1-5, doi:10.1109/ICAECC54045.2022.9716648.

[6] Anns Issac, Aswathy Reghu, Aswathy S, Jinu P Sainudeen, "A Review on Application ofMachine Learning and Deep Learning for Intrusion Detection", International Journal ofEngineering Research & Technology ,ICCIDT 2022 (Volume 10 – Issue 04), ISSN (Online) : 2278-0181.

[7] Gharaee, H., & Hosseinvand, H. (2016). A new feature selection IDS based on genetic algorithmand SVM. In 8th International Symposium on Telecommunications (IST), 139-144.

[8] Hooshmand, M.K., & Gad, I. (2020). Feature selection approach using ensemble learning fornetwork anomaly detection. CAAI Transactions on Intelligence Technology, 5(4), 283-293.

[9] Types of Intrusion Detection System." [Online]. Available: https://en.wikipedia.org/wiki/Intrusion_detection_system.

[10] 2. K. S. Desale, C. N. Kumathekar, and A. P. Chavan, "Efficient Intrusion Detection System using StreamData Mining Classification Technique,," in International Conference on Computing Communication Control and Automation,, 2015.

[11] A B. Athira, V. Pathari, "Standardisation and Classification of Alerts Generated by Intrusion Detection Systems", IJCI, International Journal on Cybernetics & Informatics, Vol 5 Issue 2, 2016.

[12] Ren, Y. Python Machine Learning : Machine Learning and Deep Learning With Python ,. International Journal of Knowledge-Based Organizations, 11(1), 67–70. 2021.

[13] Suraya, S., & Sholeh, M. Designing and Implementing a Database for Thesis Data Management by Usingthe Python Flask Framework. International Journal of Engineering, Science and Information Technology,2(1), 9–14, 2021 https://doi.org/10.52088/ijesty.v2i1.197.

[14] Çelik, Ö. A Research on Machine Learning Methods and Its Applications. Journal of Educational Technology and Online Learning. 2018. https://doi.org/10.31681/jetol.457046 [3].

[15] Chandiramani, A. Management of Django Web Development in Python. Journal of Management and Service Science (JMSS), 1(2), 1–17, 2021. https://doi.org/10.54060/jmss/001.02.005.

[16] Brownlee, J. (2016, August 16). What is Deep Learning? Retrieved December 30, 2017, from Machine Learning Mastery: https://machinelearningmastery.com/what-is-deeplearning/

[17] Chen, S. (2017, June 16). A Basic MAchine Learning Workflow In Production. Retrieved January 5, 2018,from Medium: https://medium.com/eliza-effect/how-machines-learnd9e9a3e6f97c.

[18] C. A. Ronao and S. B. Cho, "Anomalous query access detection in RBAC-administered databases with PART and PCA," Information Sciences, vol. 369, pp. 238–250, 2016.

[19] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, and Y. L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," Information Sciences, vol. 378, pp. 484–497, 2017.

[20] Idhammad, M., Afdel, K., & Belouch, M. (2017). Dos detection method based on artificial neural networks. International Journal of Advanced Computer Science and Applications, 8(4), 465-471.

[21] Deshpande, A., & Sharma, R. (2018). Multilevel Ensemble Classifier using Normalized Feature based Intrusion Detection System. International Journal of Advanced Trends in Computer Science and Engineering, 7(5), 72-76. https://doi.org/10.30534/ijatcse/2018/02752015.

[22] A. Iftikhar, M. Basheri, M. Javed Iqbal, A. Raheem; "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", IEEE ACCESS, Survivability Strategies for Emerging Wireless Networks, 6 ,pp.33789-33795, (2018).