IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT
Impact Factor: 7.301

Volume 3, Issue 2, April 2023

Monitoring Network Traffic for Suspicious Activity

B. Ravi Teja[1], Josyabhatla Sreejani[2], Jonnada Sasidhar[3], Kalapala Devakiran[4], Gunna Revanth[5]

Faculty, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4,5]
Raghu Institute of Technology, Visakhapatnam, India

Abstract: *An intrusion detection system investigates hostile behavior within a network or an approach. Software or a gadget called intrusion detection scans a network or system for an untrustworthy action. As computer connectivity increases, intrusion detection becomes increasingly important for network security. Many Intrusion Detection Systems have been built to defend the networks using statistical and machine learning technologies. Accuracy is a crucial factor in how well an intrusion detection system performs. To decrease false detections and boost detection rates, the accuracy of intrusion detection needs to be improved. In recent works, many strategies have been employed to enhance performance. The Intrusion detection system's primary task is to analyze network traffic data. To solve this problem, a structured classification system is needed. This problem is approached in the suggested manner. Classification methods are often used to address related issues. NSL-KDD knowledge discovery Dataset is used to evaluate the results of these systems. This research aims to find an efficient classifier that detects anomaly traffic with a high accuracy level and minimal error rate by experimenting with possible machine-learning techniques.*

Keywords: Intrusion Detection, Deep Learning approaches, anomaly-based network intrusion detection, Classifiers, NSL-KDD

I. INTRODUCTION

The Internet and enterprise networks primarily create and support new business opportunities. As a result, assaults are more likely to target today's networks. The various attacks and their dynamic nature necessitate a flexible approach to developing network security. To accurately detect a wide range of assaults, a flexible security system is necessary. Intrusion detection techniques are a helpful way in this situation to identify threats in intrusion detection systems. Anomaly detection provides a technique to identify potential risks by continuously monitoring and modeling the networks' usual activity.

In anomaly detection, anomalies are significant because they indicate severe but rare events. For example, unusual network traffic patterns could mean that your computer has been attacked and that data is sent to unauthorized destinations. A significant factor of anomaly detection is the nature of the anomaly. An anomaly can be categorized in three ways [1], including point anomaly, contextual anomaly, and collective anomaly. These different types of anomaly have a relationship with the attacks in network security, including DoS, Probe, U2R, and R2L.

1.1 Intrusion Detection System

The DoS attack's traits coincide with the overall oddity. Attacks known as probes are focused on gathering information and performing surveillance, making them compatible with contextual anomalies. User to Root (U2R) attacks is unauthorized access to the administrator account that takes advantage of one or more security holes. In remote-to-local (R2L) attacks, the attacker employs trial and error to determine the password before gaining local access and the ability to send network packets. Both U2R and R2L attacks are complex and condition-specific. Network intrusion detection systems must find all these anomaly kinds, analyze them, and group them into different categories of network attacks. However, in many instances, abnormal activities in the system could be outdated versions of expected behaviors. Anomaly-based NIDSs have utilized a variety of methodologies over the years, including statistical, knowledge-based,

DOI: 10.48175/IJARSCT-9153

and machine learning-based techniques. However, specific research issues need to be carefully examined to enhance performance and make them compatible with contemporary network data characteristics.

B. Machine Learning

Building analytical models is automated using machine learning. It is a method for analyzing data. It is one of the applications of artificial intelligence that relies less on human interaction as a machine learns, makes judgments, and recognizes patterns. The two most popular machine learning strategies are supervised and unsupervised learning. Unsupervised learning's two main goals are to find some structure in the data and to explore the data. Here, models for classification, regression, and prediction are applied. The three main elements of this learning are the agent, environment, and actions. The objective is for the agent to choose those activities that take advantage of the predictable payoff.

A significant issue in network intrusion detection is the availability of labeled data for the training and validation of models. Labels for normal behavior are usually available, while labels for intrusions are not. Therefore, motivated by this, we propose a system using deep learning. With deep learning, we expect to tackle issues in network intrusion detection, such as high intrusion detection rate, ability to adapt to dynamic network environments, and unavailability of labelled data. Deep learning demonstrates the effectiveness of generative models with high-accuracy classification and the capacity to extract information from sparse training data partially. There aren't many studies employing deep learning for intrusion detection, but those that do haven't fully tapped into the technology's potential. Our work studied and contrasted deep learning algorithms in anomaly-based NIDS, processed massive data, and took advantage of it.

II. RELATED WORK

Protecting computer network information of organizations and individuals became problematic because compromised information can cause huge losses. Numerous studies have been conducted on intrusion detection systems. The emergence of Big Data makes it increasingly difficult for traditional ways to handle it. As a result, many academics want to develop an accurate and quick intrusion detection system using big data approaches. This section contains several examples of research that handled intrusion detection using machine learning algorithms. The proposed model was tested using the full KDDCup1999 dataset. The KDD Cup 1999 is used to train and test the suggested approaches.

A hybrid model is created by incorporating machine learning methods like Extreme Learning Machine (ELM) and SVM [4]. High-quality datasets are created using modified K-means. This step shortens the classifier’s training period. It displays a 95.75% accuracy rate.

Artificial Bee Colony (ABC) and Artificial Fish Swarm (AFS) are the subjects of a new hybrid classification technique that has been suggested [6]. Due to the ubiquitous use of the internet, systems are vulnerable to various information thefts, which has prompted the development of IDS. Training datasets are separated, and unnecessary features are removed using Fuzzy CMeans Clustering (FCM) and Correlation-based Feature Selection (CFS) [6]. While generating if-then rules, the CART technique distinguishes between normal and abnormal records according to the chosen attributes.

KDD99 and NSL-KDD intrusion detection datasets were compared using a Self Organization Map (SOM) artificial neural network by Laheeb et al. [7]. They employed a hierarchical anomaly intrusion detection system with an unsupervised artificial neural network. SOM neural nets are used to distinguish between regular traffic and attacking traffic. The report also assessed SOM's effectiveness in detecting anomalous intrusions.

Bhupendra et al. analyzed the performance of the NSL-KDD dataset using an ANN. The ANN accuracy is presented [8]. The result was based on various performance measures for five classes and binary class classification on attacking types.

A study by Verma et al. [9] demonstrates an opportunity for improvement in anomaly-based intrusion detection, particularly regarding the false positive rate. On the NSL-KDD dataset, extreme gradient boosting (XGBoost) and adaptive boosting (AdaBoost) learning techniques were used. Although an accuracy of 84.253 was achieved, increasing performance using hybrid or ensemble machine learning classifiers is still necessary.

The cluster machine learning technique was employed by Ferhat et al. To assess whether the network traffic is an attack or regular, the authors used the k-Means approach in the machine. A clustering method for IDS based on Mini Batch K-means and principal component analysis was proposed by Peng et al. (PCA). Before using the tiny batch K-means++ approach for data clustering, the processed dataset's dimension is reduced using the principal component analysis method. The authors did not employ feature selection in this proposed strategy. Machine learning for classification was utilized by Peng et al. The authors suggested a decision tree-based IDS system for use with big data in a foggy environment. The researchers developed a preprocessing technique to identify the strings in the provided dataset. They subsequently normalized the data to ensure the input data's accuracy and improve detection effectiveness. They used the decision tree method for IDS and contrasted it with the KNN and Naive Bayesian methods. The experimental findings using the KDDCUP99 dataset demonstrated the efficiency and accuracy of the suggested methodology.

III. DATASET ANALYSIS

The dataset's selection and use significantly impact the algorithm's performance. The NSL-KDD dataset is resolving the KDDCUP99 dataset's fundamental issues... To help classifiers generate a fair outcome, redundant records were deleted from the training and test datasets of the NSL-KDD dataset.

This study's training and test dataset includes the two target values, normal and abnormal. While the remaining network traffic was labelled normal traffic, the known attack types were grouped as anomaly traffic.

The original NSL-KDD dataset contained a label and 41 features. As the KDD dataset has three features of object values that need to be converted to numeric format before applying classifiers, the NSL preprocessing step is carried out. These are the three characteristics: There are three distinct categories for "protocol_type," 70 distinct categories for "service," and 11 distinct categories for "flag."

The dataset has 122 characteristics after using the one-hot encoding technique, with a label assigned to each instance. There are 125,973 total cases in the dataset, which is divided into a training and test dataset. There are 100,778 occurrences in the training dataset and 25,195 instances in the test dataset. The training and test datasets' respective normal and anomalous instance counts are shown in Figures 1 and 2.

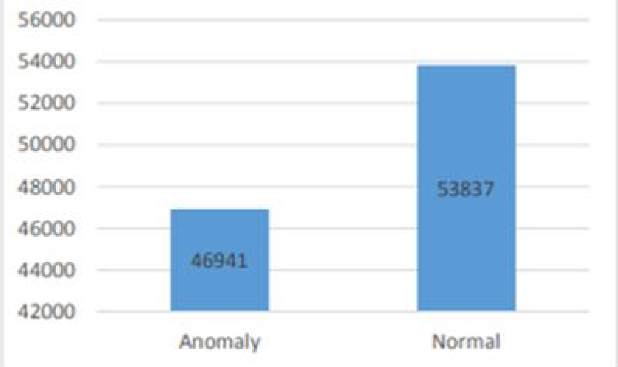

Fig 1. Training dataset target counts

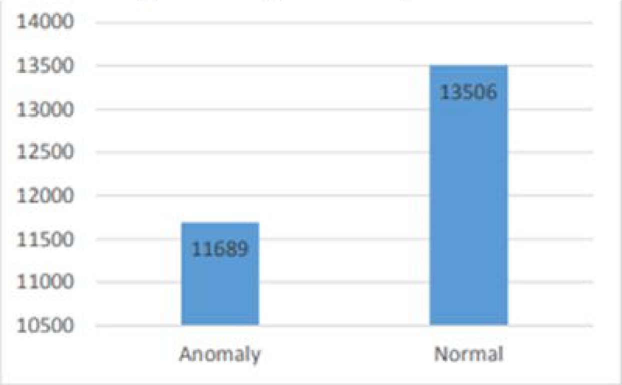

Fig 2. Test dataset target counts

DOI: 10.48175/IJARSCT-9153

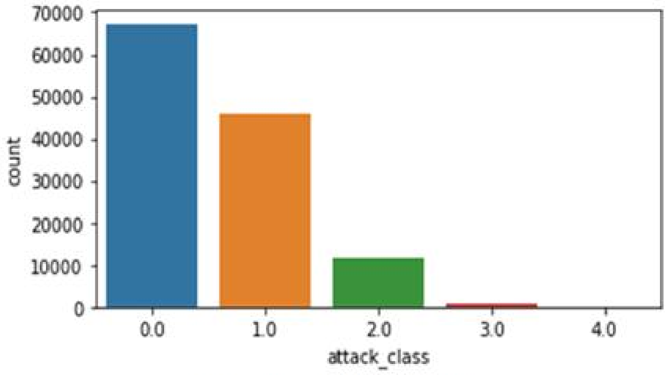

Fig 3. attack class counts

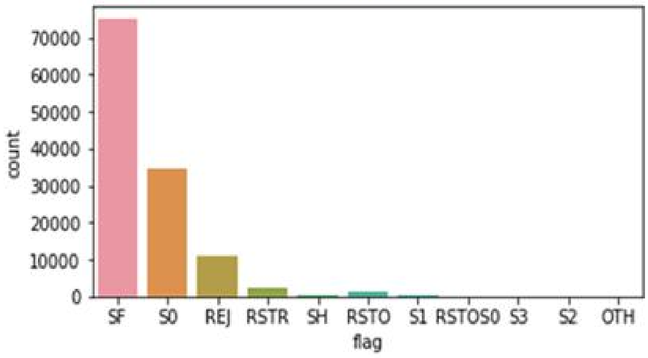

Fig 4. Flag class counts

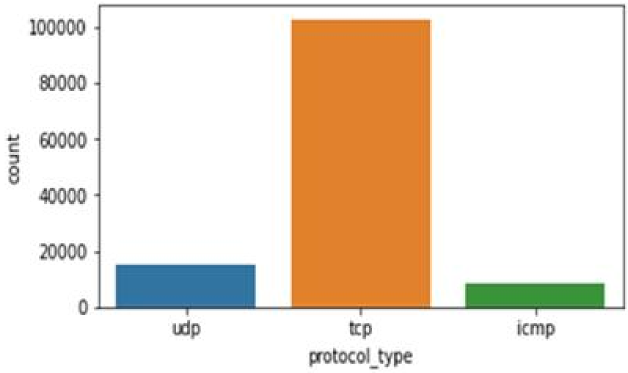

Fig 5. protocol type counts

IV. PROPOSED APPROACH

The vital approaches of the proposed model are Dataset pre-processing, classification, and custom evaluation. Each phase of the proposed system is essential, heavily enriches the performance, and influences efficiency.

Non-numeric or symbolic features must be removed or replaced as part of the overall pre-processing procedure because they play no significant role in intrusion detection. Protocol, service, and flag are symbolic qualities that might change or disappear. Eventually, the occurrences are categorized into the following four groups: R2L, Probe, DoS, and Normal

4.1 Data Pre-Processing and Preparation

The trends and patterns of the dataset are discussed earlier. The data points need to be better structured for the required analysis. The Dataset contains symbolic features, which the classifier is unable to process. Pre-processing consequently happens. All symbolic or non-numeric aspects are modified or eliminated during this phase. In the pre-processing phase, symbolic or non-numeric features are removed or replaced.

DOI: 10.48175/IJARSCT-9153

The field of machine learning (ML), a subset of artificial intelligence (AI), can be defined as teaching computers to automatically learn, enhance, or optimize performance criteria without explicit programming. To forecast distinct classes, machine learning models concentrate on training sets of data that correspond to the desired attributes[2]. Algorithms for supervised, unsupervised, and reinforcement learning comprise the broad categories of machine learning.

Non-numeric or symbolic features must be removed or replaced as part of the overall pre-processing procedure because they play no significant role in intrusion detection. Protocol, service, and flag are symbolic qualities that might change or disappear. The instances are then divided into four categories: Normal, DoS, Probe, and R2L.

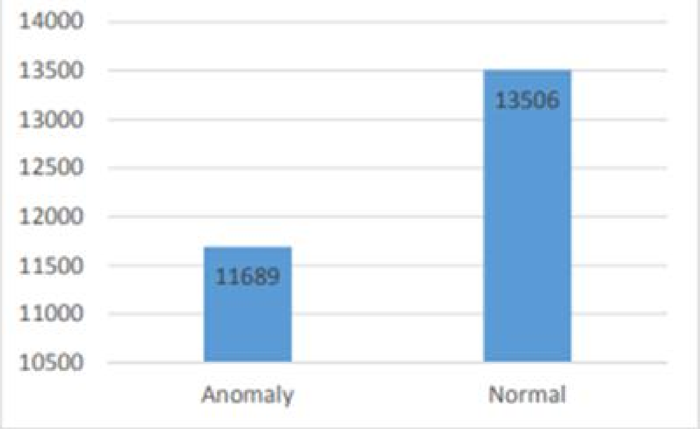

Fig 2. Test dataset target counts

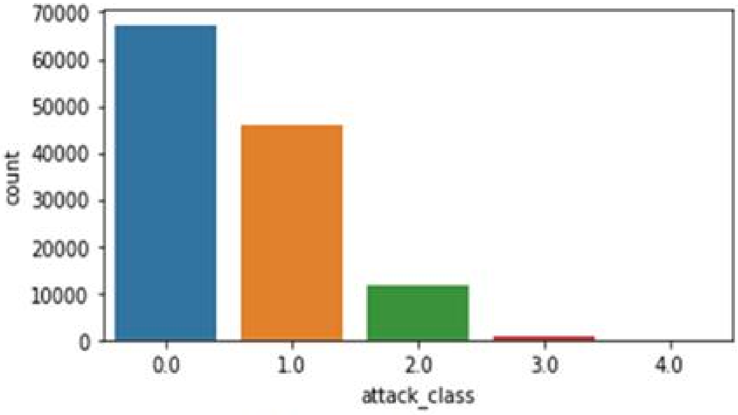

Fig 3. attack class counts

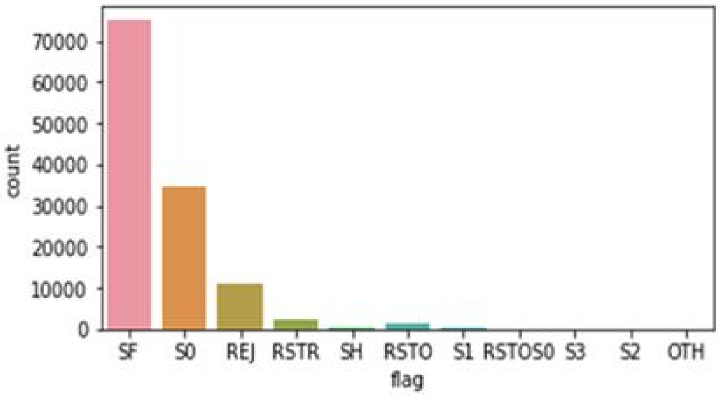

Fig 4. Flag class counts

DOI: 10.48175/IJARSCT-9153

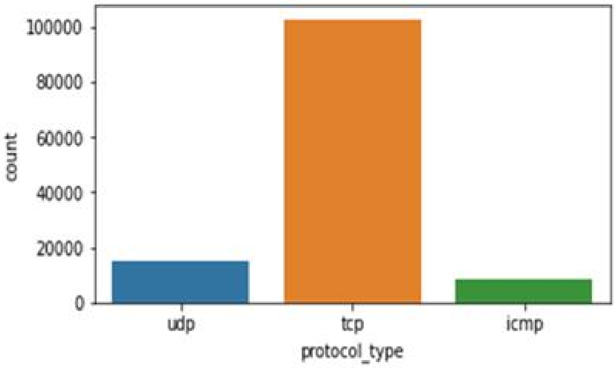

Fig 5. protocol type counts

V. PROPOSED APPROACH

The vital approaches of the proposed model are Dataset pre-processing, classification, and custom evaluation. Each phase of the proposed system is essential, heavily enriches the performance, and influences efficiency.

Non-numeric or symbolic features must be removed or replaced as part of the overall pre-processing procedure because they play no significant role in intrusion detection. Protocol, service, and flag are symbolic qualities that might change or disappear. Eventually, the occurrences are categorized into the following four groups: R2L, Probe, DoS, and Normal

5.1 Data Pre-Processing and Preparation

The trends and patterns of the dataset are discussed earlier. The data points need to be better structured for the required analysis. The Dataset contains symbolic features, which the classifier is unable to process. Pre-processing consequently happens. All symbolic or non-numeric aspects are modified or eliminated during this phase. In the pre-processing phase, symbolic or non-numeric features are removed or replaced.

The field of machine learning (ML), a subset of artificial intelligence (AI), can be defined as teaching computers to automatically learn, enhance, or optimize performance criteria without explicit programming. To forecast distinct classes, machine learning models concentrate on training sets of data that correspond to the desired attributes[2]. Algorithms for supervised, unsupervised, and reinforcement learning comprise the broad categories of machine learning.

Non-numeric or symbolic features must be removed or replaced as part of the overall pre-processing procedure because they play no significant role in intrusion detection. Protocol, service, and flag are symbolic qualities that might change or disappear. The instances are then divided into four categories: Normal, DoS, Probe, and R2L.

A probabilistic approach called Naive Bayes can be used by NIDS to categorize network data as either malicious or benign. The foundation of Naive Bayes is the Bayes theorem, which says that the likelihood of a hypothesis given some evidence,, is inversely proportional to the likelihood that the hypothesis will be correct. The likelihood of one feature does not change depending on the presence or absence of another feature. Naive Bayes can still perform well if the features are approximately independent or not overly strong, even though this assumption might not hold true in practice.

Network intrusion detection systems can benefit from the potent deep learning method known as convolutional neural networks (CNNs). By automatically identifying pertinent features from the raw data, Convolutional, pooling, and fully linked layers are among the many layers that make up CNNs. With a series of filters or kernels that glide over the input and extract features, CNNs operate by performing convolutions on the input data. The pooling layer, which shrinks the size of the feature maps and removes the most pertinent features, is applied after the convolution layer's output.

One benefit of using CNNs for NIDS is that they can automatically learn and extract complex features from the raw network traffic data without the need for manual feature engineering. Techniques such as regularization and early stopping can be used to prevent overfitting. The fully connected layers then take the extracted features and classify the input as benign or malicious.

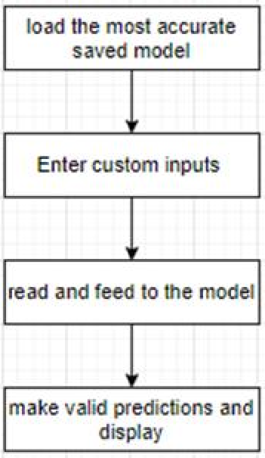
load the most accurate saved model
Enter custom inputs
read and feed to the model
make valid predictions and display

Fig 7. Flow chart of testing custom inputs

5.2 Best of the Lot

The examination of the several machine learning models developed for NIDS revealed that Logistic Regression outperformed the others. The accuracy metrics collected from the models show that this is the case. A statistical technique called logistic regression is well known for being able to handle both linear and nonlinear correlations between input and output variables. It is also well-known for its capacity to simulate event probability based on input data. Because NIDS must identify potential security threats in network traffic and detect unusual activity, this makes it an appropriate option. Overall, Logistic Regression's performance in IDS shows that it has the potential to be an effective tool for network security applications

VI. PERFORMANCE ANALYSIS

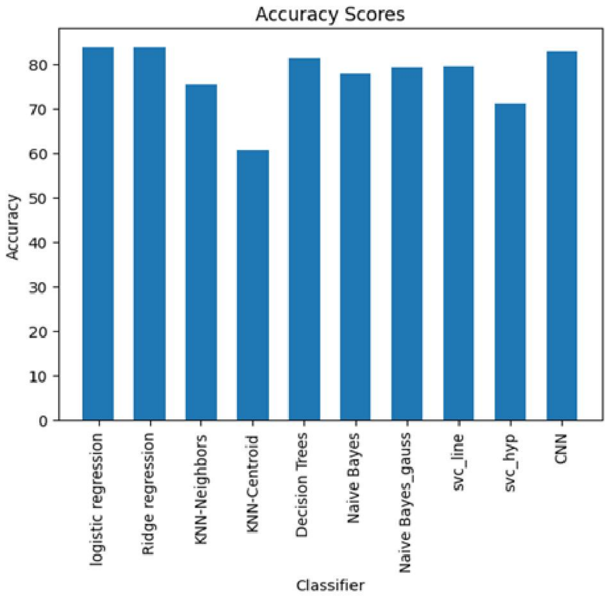

Fig 8. A Comparison Of The Accuracy Of Models

The graph above compares the classification accuracy of different machine learning models given the same data set after feature reduction. From the graph, we can conclude that logistic regression achieves an accuracy of 88% and CNN achieves an accuracy of 87.6,slightly lower than the previously mentioned model.

DOI: 10.48175/IJARSCT-9153

VII. CONCLUSION

In conclusion, network intrusion detection systems (NIDS) that utilize machine learning (ML) techniques offer significant advantages in detecting and mitigating cybersecurity threats. ML-based NIDS leverages the power of data analysis, pattern recognition, and learning algorithms to detect anomalous activities and potential intrusions in real-time, enhancing the security of networks and systems. They can detect complex attack patterns and adapt to new threats by continuously learning from new data, making them more robust and resilient against evolving cyber threats. ML-based NIDS can also reduce false positives and false negatives compared to traditional rule-based NIDS, resulting in fewer false alarms and more accurate alerts.

VIII. FUTURE SCOPE

Machine learning-based network intrusion detection systems (NIDS) have already demonstrated considerable promise in identifying network intrusions and anomalies. Research and development can be improved in several areas to improve the efficiency and effectiveness of these systems. Some potential future applications for machine learning-based network intrusion detection systems: NIDS utilizing hybrid machine learning models, which combine supervised and unsupervised learning techniques, can be more accurate in identifying both known and unidentified threats. To improve the precision of intrusion detection, deep learning models such as convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) Networks, and Autoencoders can be integrated with conventional machine learning models. Defending machine learning models from adversarial attacks is the focus of the developing field of adversarial machine learning. It can be used to strengthen NIDS' resistance to attacks intended to avoid detection.

REFERENCES

[1]. .H.Wang,J.Gu,andS.Wang,‘‘An effective intrusion detection framework based on SVM with feature augmentation,’’ Knowl.-Based Syst., vol. 136, pp. 130–139, Nov. 2017

[2]. Farah N. H et al. (2015). Application of Machine Learning Approaches in Intrusions Detection Systems. International Journal of Advanced Research in Artificial Intelligence. IJARAI. (9-18).

[3]. S. Revathi and Dr. A. Malathi, “A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection,” International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, ISSN: 2278-0181, December – 2013 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.6760&rep=r ep1&type=pdf

[4]. Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman , Mohd Zakree Ahmad Nazri; “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, ELSEVIER, Expert System with Applications, Volume.66,Jan 2017,pp.296-303.

[5]. Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, and Twae-kyung Park, “Feature Selection for Intrusion Detection using NSL-KDD,” Recent Advances in Computer Science, ISBN: 978-960-474-354-4 http://www.wseas.us/e-library/conferences/2013/Nanjing/ACCIS/ACCIS30.pdf.

[6]. Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu; “A Feature Reduced Intrusion Detection System Using ANN Classifier”, ELSEVIER, Expert Systems with Applications,Vol.88,December 2017 pp.249-247

[7]. Laheeb M. Ibrahim, Dujan T. Basheer and Mahmod S. Mahmod, “A Comparison Study for Intrusion Database (KDD99, NSL-KDD) Based on Self Organization Map (SOM) Artificial Neural Network,” Journal of Engineering Science and Technology, Vol. 8, No. 1, pp. 107 – 119, 2013 https://core.ac.uk/download/pdf/25739889.pdf

[8]. Bhupendra Ingre and Anamika Yadav, “Performance Analysis of NSL KDD dataset using ANN,” Signal Processing and Communication Engineering Systems (SPACES), 2015 International Conference, 2015, Page(s):92- 96

[9]. Verma P, Shadab K, Shayan A. and Sunil B. (2018). Network Intrusion Detection using Clustering and Gradient Boosting. International Conference on Computing, Communication and Networking Technologies (ICCCNT). (pp. 1-7). IEEE.

DOI: 10.48175/IJARSCT-9153