An Effective Method for Detecting Cyber Attacks on Computer Networks from the NSL-KDD Data Set

Aseena Babu Shaik[1], *Dr. Rajeswara Reddy*[2], *Dr. Nagagopi Raju Vullam*[3], *Dr. Gondi Konda Reddy*[4], and *Dr. Subhani Shaik*[5]

[1]Dept. of AIML, Samskruthi College of Engineering and technology(A), Hyderabad, India.
[2]Dept. of S & H, Samskruthi College of Engineering and technology(A), Hyderabad, India.
[3]Dept. of AIML, Chalapathi College of Engineering and technology(A), Guntur, A.P., India.
[4]Dept. of ME, Sreenidhi Institute of Science and technology(A), Hyderabad, India.
[5]Dept. of IT, Sreenidhi Institute of Science and technology(A), Hyderabad, India.

Abstract. Cybercrime is rapidly increasing and exploits various vulnerabilities in these computing environments. Ethical hackers pay more attention to determining vulnerabilities and recommending mitigation methods. Due to the effectiveness of machine learning in solving cybersecurity problems, machine learning is of great importance to cybersecurity. Machine learning modelsare used to advance the techniques to detect and solve cybersecurity problems. Machine learning methods help detect more cyber attacks more efficiently than other software-oriented techniques, reducing the burden on security analysts. Adaptive methods such as machine learning can improve detection rates. Logistic regression is used to resolve the issue of intrusion identification and a novel research model for intrusion identification. Logistic regression models can fully favor network traffic structure information to capture features more comprehensively. Experimental outcomes show that the algorithm behaves better than traditional methods.

1 Introduction

Internet technology is advancing and improving with the new generations, internet gives the population several practical opportunities. However, we also aspect many security threats. Network viruses, malicious attacks, and eavesdropping are on the advancement, and building network security is a booming matter for civilization and government departments. Fortunately, intrusion detection can easily solve these complications. Intrusion detection plays a crucial role in establishing the security of network intelligence. However, with the uncontrollable expansion of Internet employment, the traffic varieties within the networks are rising day to day, and the operating components of the networks are becoming more and more complicated, which poses a great trial for intrusion detection [1,2]. Identifying various kinds of malevolent network traffic,

especially abrupt malevolent network traffic, is an importantand unavoidable problem.

Network traffic could be divided into two divisions. They are normal and malicious traffic. Additionally, network traffic could be classified into five categories: probe attack, user-to-root, root-to-local, and denial of service. Intrusiondiscovery could be viewed as a distribution problem. Increasing the performance of classifications that adequately identify malicious traffic can greatly improve the accuracy of intrusion detection.

Machine learning models [3,8] are generally used to find malicious traffic in intrusion detection. Although these techniques fit into a narrow study and frequently indicate feature construction and selection. It has poor feature assortment and can't adequately resolve the problem of data classification in the case of large interlopers, resultingin low detection efficiency and a false alarm rate in height. In recent years, intrusion recognition techniques on deep learning have been recommended one after another. The author proposes a convolution-based classification method for malware traffic.

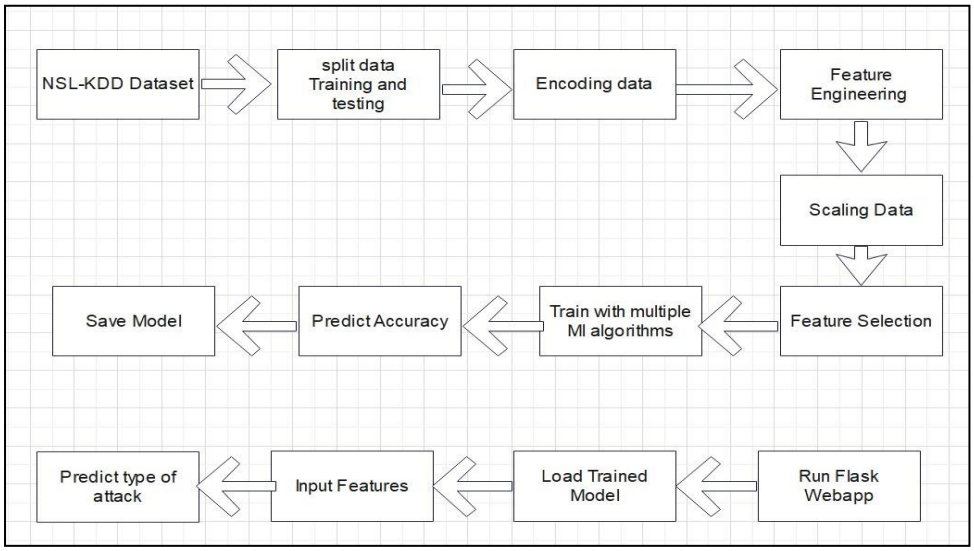

Fig. 1. Block Diagram

2 Background Work

Existing methods detect network traffic nature by analyzing the strength of recurrent neural networks, modeled as a set of conditions that change with respect to time. Existing methods check the execution of Long Short-Term Memory networks when distributing intrusive traffic. Experimental outcomes show grasp all attack classes are masked inside the training data.

In machine learning-based network attack detection research, scientists mainly focus on distinguishing between normal and anomalous network traffic through attribute depletion, grouping, and categorization to achieve malicious attack identification

[10,11]. Pervez recommended the latest technique for feature assortment and distribution integration for multiclass NSL-KDD-Cup99 datasets by using SVM, classifier distribution under different dimensional features. Accuracy was discussed [12]. Shiraz inspected several latest technologies to boost thedistribution work of CANN intrusion detection techniques and assessed their execution using the NSL-KDD Cup99 dataset [13].

He classifies data by using K Farthest Neighbor and K Nearest Neighbor. if the nearest and farthest nearby residents have the same class designation and use the second nearest neighbor (SNN). Results show CANN detection rate, reduced error rate, and improved or comparable performance with alarm rate. Bhattacharya recommended a ML model established on hybrid Principal Component Analysis Firefly [14]. The dataset is collected from Kaggle and it is an open data set. First, the model achieves key encoding and transforms the intrusion detection system dataset, then the hybrid principal component analysis Firefly algorithm is used to truncate the dimensionality, and the XGBoost algorithm distributes the shortened dataset.

Many analysts apply deep learning to traffic classification for intruder discovery, which is a hotspot of current research. Deep learning techniques notice latent features in high-dimensional data via training models, turning network traffic abnormality detection into a segmentation issue [15]. Concluded a broad range of pattern data training, adaptation learning in the middle of normal and anomalous network traffic adequately developed real-time intruder handling. Torres et al. [16] initially convert network traffic properties into a list of elements, then use a recurrent neural network to gain the temporary properties, which are after used to encounter mischievous network traffic.

Wang et al. [17] present a convolutional neural network-based algorithm for classifying malicious software traffic. A network traffic figure is produced by mapping traffic properties to pixels, and this picture is fed into a CNN to accomplish traffic distribution. R.C. Staudemeyer and Shamsinejad [13] advised a Long Short-Term Memory based on an intrusion detection algorithm that practices characteristic time sequences in the KDD-Cup99 dataset to detect DoS and probe attacks. Kwon et al. [18] have done similar work on deep learning techniques absorbing data resolution, dimensionality contraction, distribution, and additional techniques proposing a fully convolutional network model. Traditional machine learning comparison methods confirm that the network traffic investigation uses the FCN model. Tama et al. offered an abnormality-based IDS on a two-level meta-classifier that customs a fusion feature-choosing technique to attain the correct feature statement.

2.1 Supervised Machine Learning Techniques

In this paper, our main goal is to achieve an accurate cyber-attack prediction system using some supervised machine learning models. The algorithms used here are Logistic Regression, Gaussian Naive Bayes, Decision Tree, Support Vector Machine, and Gradient Boosting Classifier are deliberated below.

2.1.1 Logistic Regression

Logistic regression is the most popular machine learning model. It is used to anticipate an absolute dependent variable accepting specified customary sovereign variables. Predict the outcome of an absolute dependent variable. Therefore, the outcome must be absolute or distinct. It may be represented in yes/no, 0/1, and true/false. However, rather than giving accurate values like 0 and 1, it gives contingency values in the limits between 0 and 1. In logistic regression, rather than fixing the regression boundary, we fix an "S" shaped logistic regression activity that anticipates the two maximal values (0/1).

A logistic function arc establishes contingency of whether a cell is destructive, whether a mouse is overweight upon its weight, and so on. Logistic regression is an essential machine learning algorithm since it produces possibilities and can arrange modern data accepting continuous and discrete data sets. Use logistic regression to classify observations situated on different data types and efficiently determine which variables are most adequate for distribution. The following figure shows the logistic function. The general logistic function {\display style p:\mathbb {R} \right arrow (0,1)} can be written as [18]:

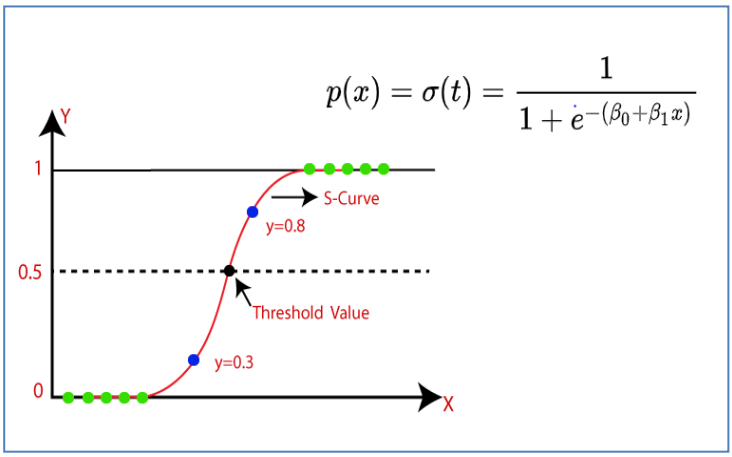

Fig. 2. Logistic Regression graph

2.1.2 Gaussian Naive Bayes

The Naive Bayes algorithm is an effective supervised learning model that corrects distribution issues and isgrounded on the Bayes theorem. It is especially used in content distribution with high-dimensional training datasets. This classifier is one of the fundamental and most adequate distribution algorithms for building quick machine-learning models able to make fast predictions. It's a probability classifier, in essence, it makes forecasting upon object probabilities.

$$P(M/N) = P(N/M) * P(M)/P(N) \qquad (1)$$

2.1.3 Support Vector Machine

A support vector machine is employed under a supervised learning model for both distribution and regression complications. Although it is mainly operated for machine learning distribution complications. The main aim of the SVM model is to build an excellent line or agreement borderline that can separate the n-dimensional volume

into classes so that the latest data points can be comfortably settled in the exact category later. The optimal decision borderlineis called a hyperplane. SVM determines vectors that assist generate hyperplanes. These maximum instances are called support vectors [16].

2.1.4 Decision Tree Classifier

In a decision tree, every inner node specifies a test for an element of every branch revert test results in every leaf node supplies a class label. Decision tree classifiers are pre-owned for grouping and regression. The goal is to frame a model that learns simple decision regulations copied from the properties of the data to predict the values of the destination variable. You can imagine a piecewise constant approximation of the tree [17].

2.1.5 Classification of Attacks

Attacks are classified into four broad categories.

i) DOS

A denial-of-service attack is a cyber-attack upon hardware, software, or alternative network assets that denies authorized users from using entitled resources and services. Typically, this is done by saturating the target host or network with traffic until it becomes passive or fails. DoS attacks could waste money and time while an organization's assets and services are inaccessible. They can last from hours to months.

ii) R2L (Remote to Local)

R2L is a type of computer network assault when a hacker sends a couple of packets to another computer or server via anetwork where he or she didn't have authorization to entry as a local user. It is used to get unauthorized remote access to a specific network address.

ii) U2R (User to Root)

The attacker first gains access to a typical user account before using system flaws to take control of the root.

3 METHODOLOGY

All types of data that are in text form are translated to numeric form during the preprocessing stage. Testing data and training data are two categories of preprocessed data. Logistic regression is used for building a model. This model is employed to forecast the testing data labels. The forecasted labels and the actual labels are contrasted. The calculation of correctness, true positive rate (TPR), and false positive rate (FPR). These factors are used to compare the models' performances.

After Data collection and Pre-processing the Dataset is cut into training data and testing data

now here, we are takingtraining data for building the classifier model i.e. Logistic Regression. Then, the classifier model has to test on training data and enumerate and contrast F1-scores and Accuracy for all the models.

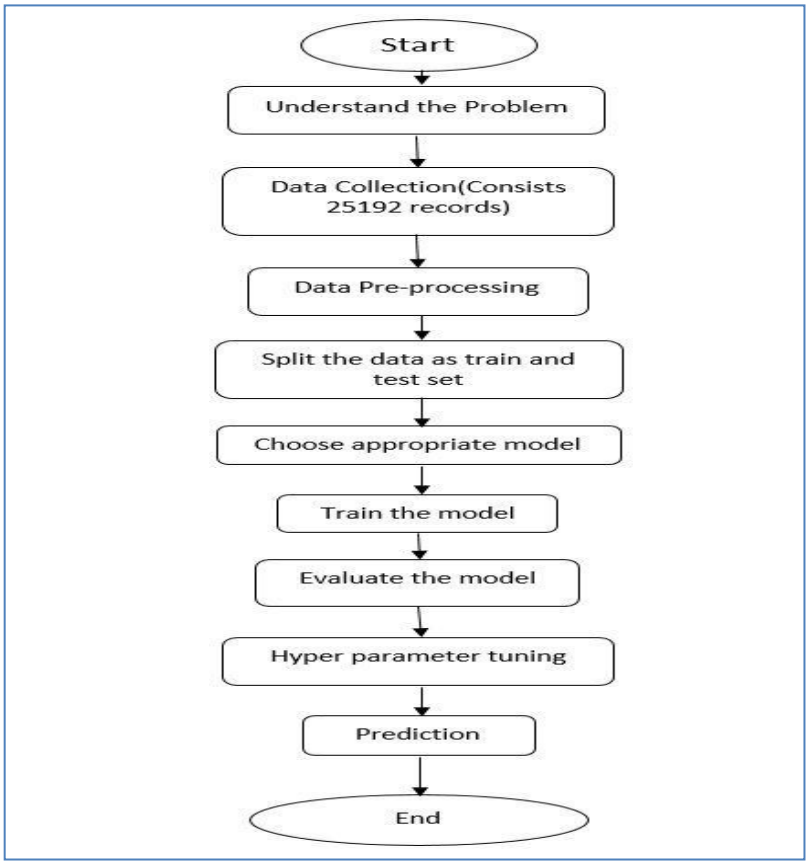

Fig. 3. Implementation process of Detecting cyber attacks

3.1 Proposed System

We proposed an end-to-end logistic regression for machine learning containing logistic regression and attentional mechanisms. Logistic regression successfully solves the issue of

intrusion recognition and provides the latest investigation approach for intrusion exposure.

Comparing the work of logistic regression with popular deep learning techniques, the model could extract knowledge from every packet. Logistic regression models can take full advantage of network traffic structure intelligence to capture features more comprehensively. We figure out the proposed network on the actual NSL-KDD dataset (train set test set). Experimental outputs show that the algorithm executes more than popular methods.

4 Results and Analysis

4.1 DOS Attack input

The following Figure 4 gives specific input of DOS. It shows that the time duration is 7638 sec. The used service is telnetand the protocol type is TCP with destination bytes 44. Port count and destination count are the same as 1.

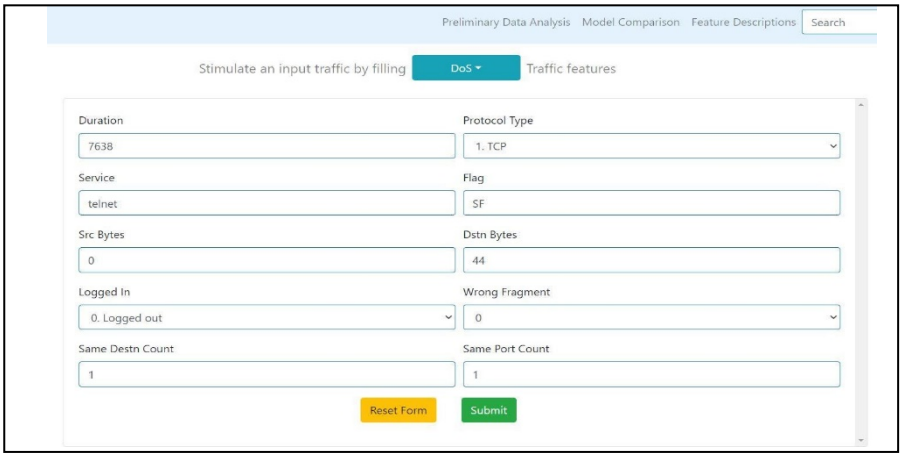

Fig. 4. DOS attack input

4.2 DOS Attack Output

The following Figure 5 shows that the Dos attack is 99% detected and the probe attack is 1%.

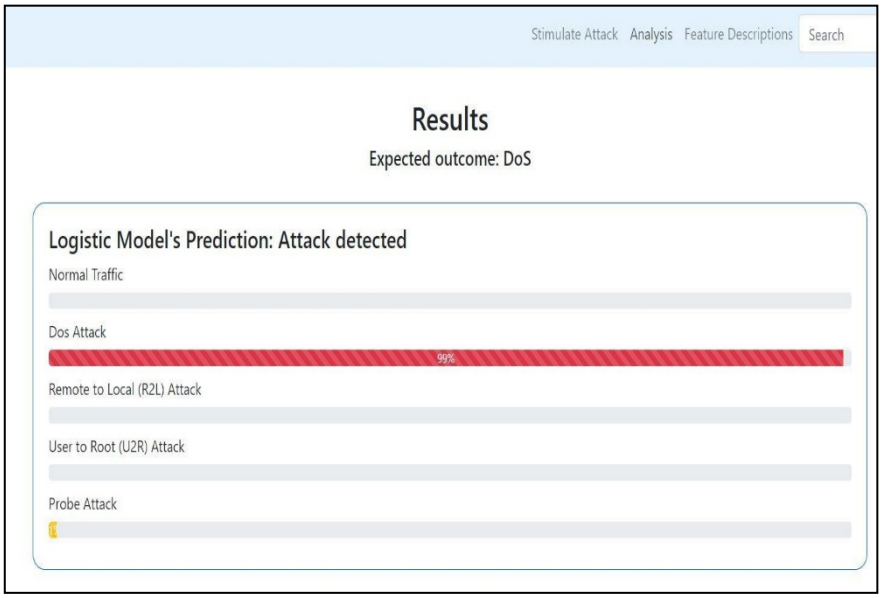

Fig. 5. DOS attack output

The following Figure 6 describes the remote to local input with duration 0. The source byte is 124 and the Destination bytesare 174. The protocol type is TCP. Destination counts and port counts are the same.

4.3 R2L Attack Input

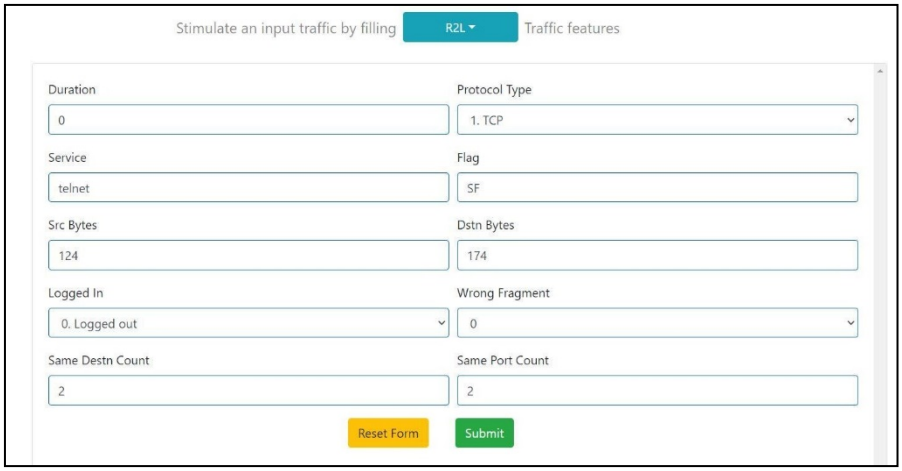

Fig. 6. R2L attack input

4.4 R2L Attack Output

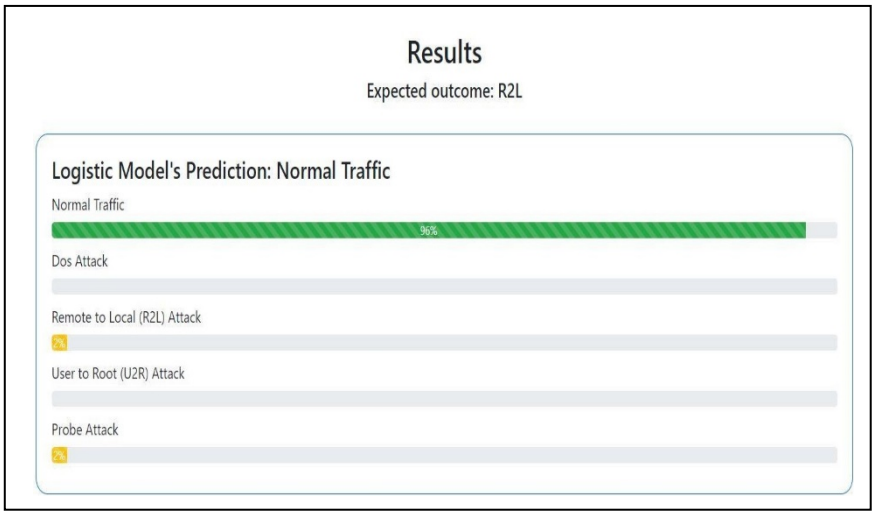

Fig. 7. R2L attack output

The following Figure 7 shows the logistic model prediction that traffic is normal, 2% remote, 2% local, and 2% Probeattack is detected.

5. Conclusion and Future Scope

As network attacks evolve, so does the need to detect cyber- attacks. In particular, issues created by unbalancednetwork traffic make it ambitious for intrusion detection systems to conclude the spread of malevolent attacks, posing an important threat to cyberspace security. This research paper recommended a novel difficult-set sampling model that empowers distribution models to enhance learning on unbalanced network data. By intentionally increasing the number of the few samples to learn, we can reduce network traffic imbalance, enhance minority learning under difficult samples, and increase classification efficiency. We used several machine-learning classification mechanisms and mixed them with alternative sampling methods. Experimentations appear that our technique can precisely identify which samples should be scaled with asymmetric network traffic to enhance attack detection more adequately.

In the investigation, we found that logistic regression with machine learning improved accuracy after samples from an unbalanced training set were sampled with the algorithm. Neural networks enhance data representation, but modern popular datasets have earlier pre-extracted features of data by making machine learning more limited to grasp pre-processed features and its automatic feature removal is not available. Thus, the later step is to straight utilize the machine learning replica to execute feature separation, model sharpening on the initial network traffic data, leverage machine learning in feature separation, and improve imbalance. We plan to moderate the effect of unreliable data and gain extra detailed classification.

References

1. D. E. Denning, ‘‘An intrusion-detection model,’’ IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222–232, Feb.1987.

2. N. B. Amor, S. Benferhat, and Z. Elouedi, ‘‘Naive Bayes vs decision trees in intrusion detection systems,’’ in Proc.ACM Symp. Appl. Comput. (SAC), 2004, pp. 420–424.

3. M. Panda and M. R. Patra, ‘‘Network intrusion detection using Naive Bayes,’’ Int. J. Comput. Sci. Netw. Secure.,vol. 7, no. 12, pp. 258–263, 2007.

4. M. A. M. Hasan, M. Nasser, B. Pal, and S. Ahmad, ‘‘Support vector machine and random forest modeling forintrusion detection system (IDS),’’ J. Intell. Learn. Syst. Appl., vol. 6, no. 1, pp. 45–52, 2014.

5. N. Japkowicz, ‘‘The class imbalance problem: Significance and strategies,’’ in Proc. Int. Conf. Artif. Intell., vol. 56,2000, pp. 111–117.

6. Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, pp. 436–444, 2015.

7. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, ‘‘Deep learning for visual understanding: A review,’’Neurocomputing, vol. 187, pp. 27–48, Apr. 2016.

8. T. Young, D. Hazarika, S. Poria, and E. Cambria, ‘‘Recent trends in deep learning based natural languageprocessing [review article],’’ IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, Aug. 2018.

9. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, ‘‘A deep learning approach to network intrusion detection,’’ IEEETrans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.

10. D. A. Cieslak, N. V. Chawla, and A. Striegel, ‘‘Combating imbalance in network intrusion datasets,’’ in Proc.IEEE Int. Conf. Granular Comput., May 2006, pp. 732–737.

11. B. B. Zarpelo, R. S Miani, C. T. Kawakami, and S. C. de Alvarenga, ‘‘A survey of intrusion detection in Internet ofThings,’’ J. Netw. Comput. Appl., vol. 84, pp. 25–37, Apr. 2017.

12. B. Mukherjee, L. T. Heberlein, and K. N. Levitt, ‘‘Network intrusion detection,’’ IEEE Network., vol. 8, no. 3, pp.26–41, May 1994.

13. S. Kishorwagh, V. K. Pachghare, and S. R. Kolhe, ‘‘Survey on intrusion detection system using machine learningtechniques,’’ Int. J. Control Automat., vol. 78, no. 16, pp. 30–37, Sep. 2013.

14. N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, ‘‘Survey on SDN based network intrusion detection systemusing machine learning approaches,’’ Peer-to-Peer Netw. Appl., vol. 12, no. 2, pp. 493–501, Mar. 2019.

15. W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, ‘‘A new intrusion detection system based on KNN classification algorithmin wireless sensor network,’’ J. Electron. Comput. Eng., vol. 2014, pp. 1–8, Jun. 2014.

16. Subhani Shaik and Dr. Uppu Ravibabu, "Detection and Classification of Power Quality Disturbances Using Curvelet Transform and Support Vector Machines", in the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016) at S.A Engineering College, Chennai, India on 25th - 26th, February 2016.

17. J. Lavanya, M. Ramesh, J. Sravan Kumar, G. Rajaramesh and Subhani Shaik,” Hate Speech Detection Using Decision Tree Algorithm”, Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page66-75, June-2023.

18. Mr. Sujan Reddy, Ms. Renu Sri and Subhani Shaik,” Sentimental Analysis using Logistic Regression”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.