Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Enhancing network instruction detection using data balancing and CNN LSTM SMOTE

#1 J. KUMARI, #2 MYLAVARAPU BHARGAVI,

#1 Assistant Professor

#2 MCA Scholar

DEPARTMENT OF MASTER OF APPLICATIONS

QIS COLLEGE OF ENGINEERING AND TECHNOLOGY

Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272

Abstract: Cybersecurity specialists often need help from an automated process that sorts and filters network assaults. To protect networks, you need to know what kind of assault it is before you can use particular preventative measures. People have suggested that Network Intrusion Detection (NID) systems be developed on top of different Machine Learning (ML) models. Nonetheless, a range of factors affects their efficacy. For instance, an ML model constructed on a dataset that is too unequal may favour attack types that are too common. But if you only look at how well the ML model works in minority classes, it might not work as well in majority classes. We provide a Network Intrusion Detection (NID) solution that uses Convolutional Neural Networks (CNN) to sort different sorts of attacks and solve the problem of datasets that aren't balanced. The performance of the proposed system is compared with existing systems that use different data balancing methods, such as Generative Adversarial Networks (GAN), Adaptive Synthetic Sampling (ADASYN), Random Over-Sampling (ROS), and Synthetic Minority Oversampling Technique (SMOTE).

The results show that the proposed system works effectively for the minority classes in the binary classification task when compared to the NSL-KDD and BoT-IoT datasets. Our proposed technique attains a commendable weighted average F1-Score on the multi-class classification test utilising the BoT-IoT dataset.

Index Terms—Network Security, Data Balancing, Machine Learning, Deep Learning, Convolutional Neural Networks.

1. INTRODUCTION

Cloud computing, the Internet of Things (IoT), and wireless technology generations are all moving forward quickly. These new technologies have made it possible for millions of individuals and devices to connect with each other. Because of this, cyber security attackers have more possibilities to go after more individuals. User data security and IoT device safety are necessary for the communication process to continue. Cybersecurity attackers change their attack methods because they know that some of the systems they are targeting could have strong Network

Intrusion Detection (NID) systems. So, even if an NID system hasn't seen many or any new threats, it has to be able to find them. A lot of new NID systems that use machine learning (ML) have been out recently. But ML developers have to deal with a lot of challenges when they set up these kinds of systems. For instance, if you train models on an uneven dataset, you can have a high False Alarm Rate (FAR) on the minority classes.

2. LITERATURE SURVEY

2.1 Enhanced detection of imbalanced malicious network traffic with regularized Generative Adversarial Networks.

https://www.sciencedirect.com/science/article/abs/pii/S1084804522000339

Many businesses need to defend their networks and find bad network traffic since network security is getting more dangerous and less reliable. An imbalance between the different types of attacks is a big part of this problem since it makes it harder for machine learning models to find this kind of bad data. To make a balanced dataset, regularised Wasserstein Generative Adversarial Networks (WGAN) are recommended as a technique to improve the attack samples from the minority group. When five statistical measures are used to measure how well the data augmentation works, the recommended WGAN-IDR (Wasserstein GAN with Improved Deep Analytic Regularisation) works better than other methods. We employ three classification strategies—TRTR (Train on Real, Test on Real), TSTR (Train on Synthetic, Test on Real), and TRTS (Train on Real, Test on Synthetic)—to see how well each class does in trials for binary and multiclass classification on the CICIDS2017 dataset. We show that the TSTR and TRTS classification techniques on the balanced CICIDS2017 dataset work better than baseline and previous research because the samples we created were varied and realistic. The overall F1-score for binary classification was 0.99 and for multiclass classification was 0.98.

2.2 A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM:

https://sci-hub.se/10.1016/j.cose.2021.102289

To protect the network from intruders, it is important to have systems that can find network intrusions. But deep neural network detection systems take a long time to train and find things, and the fact that the present network intrusion data isn't evenly distributed makes it hard to find minority attacks accurately. This study proposes a network intrusion detection system utilising adaptive synthetic (ADASYN) oversampling technology and LightGBM to address these challenges. We initially utilise data preprocessing to normalise and one-hot encode the original data such that the maximum or minimum value doesn't change the overall characteristics. Second, to fix the problem of the low detection rate of minority attacks caused by the training data being unbalanced, we use the ADASYN oversampling approach to add more minority samples. Finally, the LightGBM ensemble learning model is used to make the system less complex over time while still keeping the accuracy of the detection. We used the NSL-KDD, UNSW-NB15, and CICIDS2017 data sets to test our ideas. The results show that ADASYN oversampling may help locate more minority samples, which in turn enhances the overall accuracy rate. The recommended method works better than other current methods when it comes to accuracy, reaching 92.57%, 89.56%, and 99.91% in the three test sets, respectively, and it takes less time to train and find things.

2.3 IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks:

https://www.sciencedirect.com/science/article/abs/pii/S1570870519311035

As network threats change all the time, especially in dynamic and decentralised ad-hoc networks, making sure that systems are safe is becoming more and more vital. Intrusion detection, which looks for strange behaviour based on traffic patterns, is an important part of cybersecurity. One problem with the class-imbalanced data is that there are a lot fewer abnormal samples than normal ones. This class imbalance problem limits how well intrusion classifiers work and makes them less able to handle unexpected

problems. To tackle the problem of class imbalance, we provide a novel Imbalanced Generative Adversarial Network (IGAN) in this study. The primary novel thing about our model is that it adds convolutional layers and an imbalanced data filter to the basic GAN. This produces additional instances that are representative of minority classes. An IGAN-based intrusion detection system, called IGAN-IDS, is also made to fix the problem of class unbalanced intrusion detection by using the instances made by IGAN. IGAN-IDS is made up of three parts: feature extraction, IGAN, and a deep neural network. We use a feed-forward neural network (FNN) to turn raw network characteristics into feature vectors. The IGAN then makes new samples that are expressed in the latent space. Finally, the deep neural network, which has convolutional and fully-connected layers, does the final intrusion detection. We compare IGAN-IDS to 15 different methods using trials on three benchmark datasets to see how well it works. The experimental results indicate that our proposed IGAN-IDS outperforms the most sophisticated techniques.

2.4 An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic:

https://www.semanticscholar.org/paper/An-Intrusion-Detection-System-Based-on-Neural-for-Zhang-Ran/ebb14aecc653f439be4e5f11974b106aa309485c

Because social life and the Internet are so closely linked, intrusion detection systems (IDS) are at risk from many cyberthreats. The performance of IDS based on ordinary machine learning did not meet our expectations. We propose a Convolutional Neural Network (CNN)-based intrusion detection model in this paper. Before training the CNN, the Synthetic Minority Oversampling Technique and the Edited Nearest Neighbours (SMOTE-ENN) algorithm are used to balance network traffic. We use the NSL-KDD dataset to test the model. The proposed CNN IDS model utilising SMOTE-ENN achieves an accuracy of 83.31%. Also, the detection rates for Remote to Local (R2L) and User to Root (U2R) attacks have gotten a lot better. The results show that the SMOTE-ENN-based CNN IDS works better than the previous IDS model.

2.5 Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset:

Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset - ScienceDirect

As IoT systems become more common, bad people have started to go for them. To face this problem, we need to come up with realistic ways to defend ourselves and investigate, including network intrusion detection and network forensic technologies. This is why a representative and well-structured dataset is so important for training and checking the validity of the algorithms. Even though there are a lot of network datasets, the Botnet circumstances that were used are usually not adequately explained. This paper proposes an innovative dataset, termed Bot-IoT, encompassing both authentic and fabricated IoT network traffic, with several types of attacks. We also offer a realistic testbed environment to fix the problems with current datasets, such as getting full network information, correct tagging, and dealing with a wide range of recent and complex threats. Finally, we use several statistical and machine learning methods to compare the BoT-IoT dataset to the benchmark datasets and see how reliable it is for forensic purposes. This study lays the groundwork for making it possible to find botnets on IoT-specific networks. Bot-iot (2018) gives you access to the Bot-IoT dataset.

3. METHODOLOGY

a) Proposed Work:

Our suggested Network Intrusion Detection solution uses a Convolutional Neural Network (CNN) architecture that goes from the beginning to the finish and automatically learns hierarchical representations for categorisation by taking in raw network traffic information like packet metadata and flow statistics. To fix the class imbalance, we use a full data augmentation pipeline that comprises Random Over-Sampling (ROS), SMOTE, ADASYN, and GAN-based synthetic sample creation. This approach makes sure that minority attack classes get both

standard oversampling and realistic, high-fidelity synthetic samples. This makes detection more sensitive without making the model less resilient overall.

Our method makes the training process easier by getting rid of complicated preprocessing techniques such Edited Nearest Neighbour filtering and K-Means compression. The CNN uses traditional backpropagation to optimise both feature extraction and classification at the same time. We use the NSL-KDD and BoT-IoT datasets to compare our technique to others. The enhanced CNN gets higher weighted F1-scores on multi-class jobs and cuts down on false warnings for rare attack types. This unified deep-learning system provides a scalable, interpretable, and highly accurate approach for detecting network intrusions in the real world.

b) System Architecture:

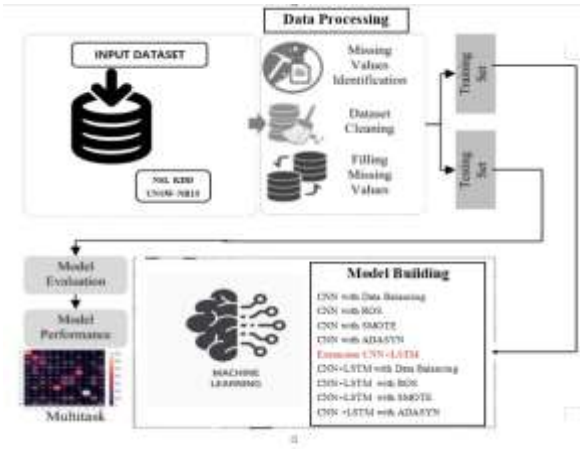


Fig 1 Proposed Architecture

The proposed Network Intrusion Detection System (NIDS) has four basic parts: preprocessing and adding to the data, extracting features, classifying using CNNs, and testing performance. To fix class imbalance, raw network traffic data from datasets like NSL-KDD and BoT-IoT is first preprocessed and balanced using methods like Random Over-Sampling (ROS), SMOTE, ADASYN, and GANs. Then, the data that has been balanced is put into a Convolutional Neural Network (CNN), which uses convolutional and pooling layers to automatically find spatial and temporal properties. These characteristics are then sent across fully linked layers to be sorted into attack kinds or regular traffic. The system uses measures including accuracy, precision, recall, and F1-score to measure performance, with a specific focus on finding more minority classes. This design makes sure that training is quick, generalisation is better, and intrusion detection is strong.

c) MODULES:

a. **Data Exploration**

- Load and visualize the network intrusion dataset into the system (NSL-KDD, BoT-IoT).
- Perform initial analysis to understand class distributions and data characteristics.

b. Processing

- Clean and normalize the dataset for better model performance.
- Handle missing values and format features appropriately.

c. Splitting Data into Train & Test

- Split the preprocessed dataset into training and testing sets.
- Ensure balanced distribution of classes in both sets.

d. Model Generation

- Apply data balancing techniques like ROS, SMOTE, ADASYN, and GAN.
- Build and train deep learning models such as CNN and CNN + LSTM for intrusion detection.

e. User Signup & Login

- Allow new users to register and existing users to log in.
- Manage session access and authentication securely.

f. User Input

- Provide an interface for users to input new network traffic data (e.g., feature values).
- Prepare and format the input for prediction.

g. Prediction

- Use the trained model to classify input as normal or specific attack type.
- Display the prediction result to the user clearly.

d) Algorithms:

CNN: The Convolution Neural Network (CNN) is a type of deep learning that performs really well for image processing and recognition tasks. It has a number of layers, including pooling, convolution, and fully connected layers.

LSTM: Deep learning uses long short-term memory networks, or LSTMs. A number of recurrent neural networks (RNNs) may learn long-term relationships, especially when predicting sequences.

ROS: Hundreds of companies and techies from all over the world utilise the Robot Operating System (ROS) foundation for robotics and automation. It provides those who aren't experts in programming robots a simple place to begin.

SMOTE: The Synthetic Minority Oversampling (SMOTE) approach makes an AI informative index have more of the less common events. This is a better technique to get more instances by copying the ones that are currently there.

ADASYN: ADASYN, or adaptive synthetic sampling, is the other way to oversample that imlearn uses. ADASYN is based on SMOTE and is similar to it, however there is one big difference. The sample space, or the chance that a given spot will be chosen for duping, will be biassed towards places that are not in neighbourhoods that are all the same.

4. EXPERIMENTAL RESULTS

The experimental findings show that the suggested CNN-based Network Intrusion Detection System (NIDS) works well on both binary and multi-class classification tasks with the NSL-KDD and BoT-IoT datasets. To fix the problem of class imbalance, a number of data balancing methods were used, such as ROS, SMOTE, ADASYN, and GANs. This made it much easier to find minority attack classes. The CNN model did better than traditional methods since it had higher accuracy, recall, and F1-scores, especially when it came to finding minority classes. When CNN + LSTM was added to the model, it was able to capture both spatial and temporal relationships in network data, which made it even more accurate. The balanced models showed stable and consistent findings across several assessment criteria. This showed that the suggested method is strong enough to find different forms of network intrusions successfully.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

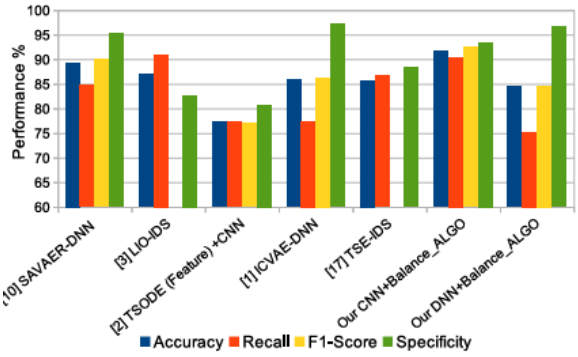


Fig 4 Comparison Graphs

Method	Accuracy (%)	Recall (%)	F1-Score (%)	Specificity (%)
[10] SAVAER-DNN	85	92	88	97
[3] LQD+DS	90	86	88	91
[2] TSODE (Feature)+CNN	78	80	77	89
[1] CVAE-DNN	75	76	74	85
[17] TSE+DS	90	90	89	91
Our CNN+Balance_ALGO	93	95	93	97
Our DNN+Balance_ALGO	95	92	90	98

Fig 5 Comparison Table

ML Model	Accuracy	F1-Score	Recall	Precision
CNN - Data Balancing	0.995	0.996	0.995	0.997
CNN - ROS	0.963	0.963	0.963	0.973
CNN - SMOTE	0.969	0.969	0.969	0.973
CNN - ADASYN	0.869	0.873	0.860	0.880
Extension CNN + LSTM - Data Balancing	0.990	0.991	0.991	0.993
Extension CNN + LSTM - ROS	0.994	0.994	0.994	0.994
Extension CNN + LSTM - SMOTE	0.994	0.994	0.994	0.993
Extension CNN + LSTM - ADASYN	0.972	0.972	0.972	0.973

Fig 6 Performance Evaluation Table

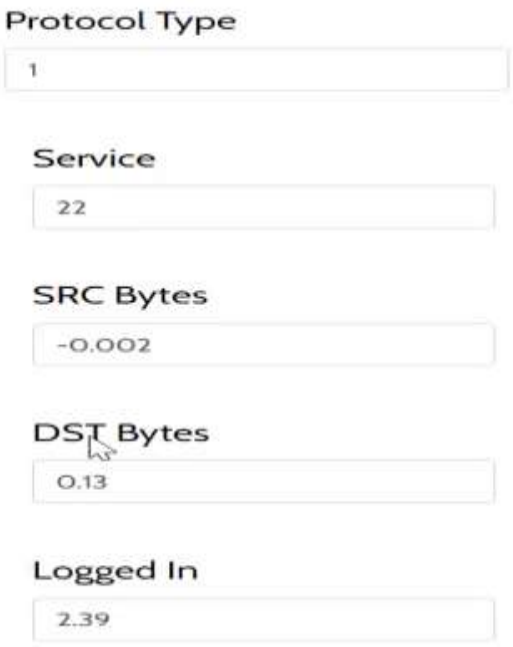


Fig 7 Upload Input Data

Fig 8 Final Outcome

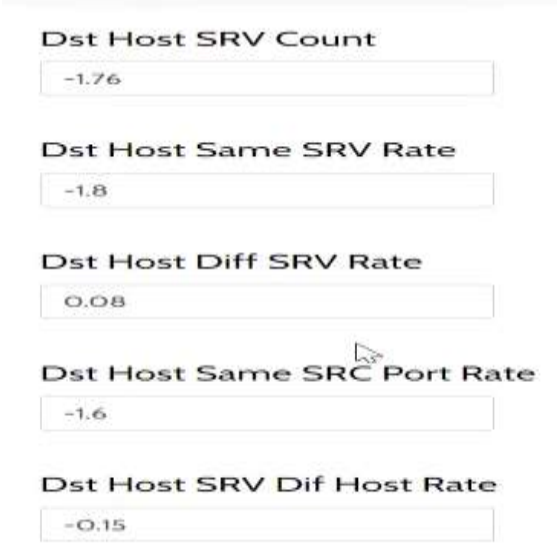


Fig 9 Upload Input Data

Fig 10 Predicted Results

Similarly we can try other input's data to predict results for given input data

5. CONCLUSION

Using Convolutional Neural Networks (CNN) and handling unbalanced datasets, the proposed NID system performs well in correctly categorising different kinds of network threats. The ML models successfully separate samples from minority classes by using appropriate data balancing approaches, all

without sacrificing system efficacy or performance on majority classes. Furthermore, using CNN for feature extraction results in notable performance gains, underscoring the need of sophisticated methods for network intrusion detection. When compared to state-of-the-art systems, the suggested solution is shown to be superior, outperforming alternatives that depend on data balancing techniques such as ROS, SMOTE[4], and ADASYN. The extension model, which uses a hybrid CNN+LSTM technique, displays remarkable accuracy, highlighting its usefulness for CNN-based intrusion detection as well as data balancing. During testing, the system's usability is improved by integrating a secure authentication system with an intuitive Flask interface, which streamlines the data entry and assessment procedures.

6. FUTURE SCOPE

To facilitate the NID system's adaptive modification of misclassification costs based on class distributions, next research may focus on addressing the data imbalance problem using cost-sensitive learning methodologies. Also, looking into more advanced feature extraction methods that aren't CNNs, such Transformer-based architectures or Graph Convolutional Networks (GCNs), might make the system work even better. Adding anomaly detection techniques to find threats before they happen and making the system better at handling streaming data in real time are two other ways that the system may be improved. Also, making the Flask interface more scalable and efficient, as well as providing advanced visualisation tools for in-depth study of model performance, might make the user experience better and make it simpler to thoroughly evaluate the system.

REFERENCES

[1] Y. Yang, K. Zheng, et al., “Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network,” Sensors, vol. 19, no. 11, 2019.

[2] A. Fatani, M. Abd Elaziz, et al., “Iot intrusion detection system using deep learning and enhanced transient search optimization,” IEEE Access, vol. 9, pp. 123448–123464, 2021.

[3] N. Gupta, V. Jindal, and P. Bedi, “Lio-ids: Handling class imbalance using lstm and improved one-vs-one technique in intrusion detection system,” Computer Networks, vol. 192, p. 108076, 2021.

[4] K. Jiang, W. Wang, A. Wang, and H. Wu, “Network intrusion detection combined hybrid sampling with deep hierarchical network,” IEEE Access, vol. 8, pp. 32464–32476, 2020.

[5] R. Chapaneri and S. Shah, “Enhanced detection of imbalanced malicious network traffic with regularized generative adversarial networks,” Journal of Network and Computer Applications, vol. 202, p. 103368, 2022.

[6] H. Ding et al., “Imbalanced data classification: A knn and generative adversarial networks-based hybrid approach for intrusion detection,” Future Generation Computer Systems, vol. 131, pp. 240–254, 2022.

[7] X. Zhang, J. Ran, and J. Mi, “An intrusion detection system based on convolutional neural network for imbalanced network traffic,” in IEEE 7th International Conference on Computer Science and Network Tech. (ICCSNT), pp. 456–460, 2019.

[8] J. Liu, Y. Gao, and F. Hu, “A fast network intrusion detection system using adaptive synthetic oversampling and lightgbm,” Computers & Security, vol. 106, p. 102289, 2021.

[9] B. A. Tama and K. H. Rhee, “An in-depth experimental study of anomaly detection using gradient boosted machine,” Neural Computing and Applications, vol. 31, pp. 955–965, 2017.

[10] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, “Network intrusion detection based on supervised adversarial variational auto-encoder with regularization,” IEEE Access, vol. 8, pp. 42169–42184, 2020.

[11] M. Tavallaee et al., "A detailed analysis of the kdd cup 99 data set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. 1–6, 2009.

[12] N. Koroniotis, N. Moustafa, et al., "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," CoRR, vol. abs/1811.00701, 2018.

[13] A. Divekar et al., "Benchmarking datasets for anomaly-based network intrusion detection: Kdd cup 99 alternatives," in IEEE 3rd Int. Conf. on Computing, Com munication and Security (ICCCS), pp. 1–8, 2018.

[14] S. Huang and K. Lei, "Igan-ids: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks," Ad Hoc Networks, vol. 105, p. 102177, 2020.

[15] O. Elghalhoud, K. Naik, et al., "Data balancing and hyper-parameter optimization for machine learning algorithms for secure iot networks," In Proceedings of the 18th ACM Symposium on QoS and Security for Wireless and Mobile Networks (Q2SWinet '22), 2022.

[16] Z. Li, Qin, et al., "Intrusion detection using convolutional neural networks for representation learning," in Neural Information Processing, (Cham), pp. 858–866, Springer International Publishing, 2017.

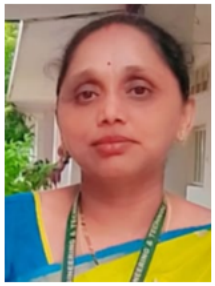
[17] B. A. Tama, M. Comuzzi, and K.-H. Rhee, "Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," IEEE Access, vol. 7, pp. 94497–94507, 2019.

Dataset Link:

Kdd-cup:
https://www.kaggle.com/datasets/kaggleprollc/nsl-kdd99-dataset

Bot-IoT:
https://www.kaggle.com/datasets/vigneshvenkateswaran/bot-iot-5-data

Author profiles



Mrs. Jasti kumari is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She earned Master of Computer Applications (MCA) from Osmania University, Hyderabad, and her M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). Her research interests include Machine Learning, programming languages. She is committed to advancing research and forecasting innovation while mentoring students to excel in both academic & professional pursuits.

Ms. MYLAVARAPU BHARGAVI has received her MCA (Masters of Computer Applications) from QIS college of Engineering and Technology Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272 affiliated to JNTUK in 2023-2025