INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 972-977

DATA BALANCING AND CNN BASED NETWORK INTRUSION DETECTION SYSTEM

M. Sharmila Devi[1], K. Tejeswari[2], V. Sreelekha[3], K. Mahek Sultana[4], S. Alfiya[5]

[1]Assistant Professor in Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

[2,3,4,5]Student, Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.

DOI: https://www.doi.org/10.58257/IJPREMS33302

ABSTRACT

The help of an automated process that filters and classifies network intrusions is often needed by cyber-security professionals. The classification of the attack type is essential for applying specific preventive measures to secure networks. Numerous Machine Learning (ML) models have been proposed as the foundation for Network Intrusion Detection (NID) systems. Yet, their efficacy varies based on many factors. For instance, an ML model trained on a highly unbalanced dataset may be biased towards over-represented attack types. On the other hand, focusing solely on the ML model's performance in minority classes can have a negative impact on its performance in the majority classes. We proposes a Network Intrusion Detection (NID) system that addresses the issue of imbalanced datasets and uses Convolutional Neural Networks (CNN) to classify different attack types. The performance of the proposed system is compared to other systems that use different techniques such as Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) for data balancing. The NSL-KDD and BoT-IoT datasets are used for benchmarking, and the results show that the proposed system performs well in the minority classes on the binary classification task. Our proposed system scores a good weighted average F1-Score on the multi-class classification task using the Bot-Iot dataset. etc.

Keywords: CNN, LSTM, ADASYN, SMOTE, ROS.

1. INTRODUCTION

Cloud computing and Internet of Things (IoT) technologies, and the generations of wireless technologies are advancing expeditiously. With the help of these advanced technologies, millions of users and devices are interconnected. This creates more opportunities for cyber-security attackers to target more victims. Securing users' information and protecting the IoT devices is crucial for the continuation of the communication process. Knowing that some of the targeted systems may have a strong Network Intrusion Detection (NID) system, cyber-security attackers use reformed attack methods. Therefore, a well performing NID system must be able to distinguish new attacks even if it has not seen any or many of them. Many machine learning (ML) based NID systems have been introduced recently. However, while implementing such systems, ML engineers have to address several issues. For instance, fitting models on an imbalanced dataset may result in a high False Alarm Rate (FAR) on the minority classes.

2. LITERATURE REVIEW

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

Due to the emerging network security vulnerabilities and threats, securing the network and identifying malicious network traffic is crucial for various organizations. One critical aspect of this problem is an imbalance among different attack classes, which degrades the learning performance of machine learning models for detecting such malicious traffic. In this work, regularized Wasserstein Generative Adversarial Networks (WGAN) are proposed for augmenting the minority attack samples to obtain a balanced dataset. The data augmentation performance is evaluated statistically with five statistical measures, and it is shown that the proposed WGAN-IDR (Wasserstein GAN with Improved Deep Analytic Regularization) performs better than other augmentation methods. Experiments for binary as well as multiclass classification are conducted on the CICIDS2017 dataset to evaluate the per-class performance using 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). Using WGAN-IDR, we show that the TSTR and TRTS classification strategies on the balanced CICIDS2017 dataset outperform baseline and existing works due to diverse and realistic generated samples, with the overall F1-score of 0.99 for binary classification and 0.98 for multiclass classification.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 972-977

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

Network intrusion detection systems play an important role in protecting the network from attacks. However, Existing network intrusion data is imbalanced, which makes it difficult to accurately detect minority attacks, and the training and detection time of deep neural network detection systems is relatively long. According to these problems, this paper proposes a network intrusion detection system based on adaptive synthetic (ADASYN) oversampling technology and LightGBM. First, we normalize and one-hot encode the original data through data preprocessing to avoid the impact of the maximum or minimum value on the overall characteristics. Second, we increase the minority samples by ADASYN oversampling technology to solve the problem of the low detection rate of minority attacks due to the imbalance of the training data. Finally,the LightGBM ensemble learning model is used to further reduce the time complexity of the system while ensuring the accuracy of detection. Through experimental verification on the NSL-KDD, UNSW-NB15 and CICIDS2017 data sets, the results show that the detection rate of minority samples can be improved after ADASYN oversampling, thereby improving the overall accuracy rate. The accuracy of the proposed algorithm is up to 92.57%, 89.56% and 99.91% respectively in the three test sets, and it consumes less time in the training and detection process, which is superior to other existing methods.

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

As the Internet integrates with social life closely, various cyber threats pose a huge challenge to Intrusion Detection Systems (IDS). The performance of IDS based on traditional machine learning did not meet our expectations. In this paper, we propose an intrusion detection model based on Convolutional Neural Network (CNN). Before CNN training, Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN) algorithm is applied to balance the network traffic. We use NSL-KDD dataset to evaluate the model. The proposed SMOTE-ENN-based CNN IDS model achieves an accuracy of 83.31%. Furthermore, the detection rates of User to Root (U2R) and Remote to Local (R2L) attacks are significantly improved. Results show that SMOTE-ENN-based CNN IDS outperforms the previous IDS model.

3. METHODOLOGY

In this section, the methodology was adopted in order to predict Network Intrusion. More specifically in section A, the dataset information is described. And Section B consists of evaluation metrics.

A. Dataset Information

The NSL-KDD dataset contains a variety of network traffic data, including normal and attack instances across different types of attacks like DoS, R2L, U2R, and probing. There are 41 features here are the main features of the NSL-KDD dataset with brief description :

- Duration: Length of the connection in seconds.
- Protocol Type: Network protocol used (e.g., TCP, UDP).
- Service: Type of service (e.g., http, ftp, telnet).
- Flag: Status of the connection (e.g., SF for normal, S0 for invalid).
- Source Bytes: Number of bytes sent from source to destination.
- Destination Bytes: Number of bytes sent from destination to source.
- Count: Number of connections to the same host as the current connection in the past two seconds.
- Same Service: Indicates if the service is the same as the current connection.
- Diff Service: Indicates if the service is different from the current connection.
- Logged In: Indicates if a user is logged in.
- Root Shell: Indicates if root shell access was obtained.
- Is Guest Login: Indicates if the login is as a guest.
- Failed Login: Number of failed login attempts.
- Attack Type: Type of attack (e.g., DoS, R2L, U2R).
- Label: Binary label indicating normal or attack.

The Bot-IoT dataset focuses on Internet of Things (IoT) devices and their network traffic. There are 46 features here are the main features of the Bot-IoT dataset with brief description :

- Duration: Time duration of the flow.
- Protocol: Network protocol used (e.g., TCP, UDP).
- Source Port: Source port number.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 972-977

- Destination Port: Destination port number.
- Total Packets: Total number of packets in the flow.
- Total Bytes: Total number of bytes in the flow.
- Source Bytes: Number of bytes sent from source to destination.
- Destination Bytes: Number of bytes sent from destination to source.
- Min Packet Length: Minimum length of packets in the flow.
- Max Packet Length: Maximum length of packets in the flow.
- Mean Packet Length: Mean length of packets in the flow.
- Packet Length Variance: Variance of packet lengths in the flow.
- Flow Bytes/s: Bytes per second in the flow.
- Flow Packets/s: Packets per second in the flow.
- Label: Binary label indicating normal or botnet traffic.

B. Evaluation Metrics

- Accuracy: It is the percentage of examples correctly classified.
- Recall: It is the percentage of actual positives that were correctly classified.

$$\text{accuracy} = \frac{Number\ of\ correct\ predictions}{Number\ of\ total\ values}$$

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

$$\text{recall} = \frac{TP}{TP+FN}$$

Precision: is the percentage of predicted positives that were correctly classified.

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{F1-score} = 2\frac{precision.recall}{precision+recall}$$

F1-score: a combination of recall and precision to get a single measure, which falls between these two metrics

4. IMPLEMENTATION AND ANALYSIS

In this section, the implementation details are mentioned to detect Network Intrusion. It contains the model selection, and the analysis that has done, and its accuracy is shown.

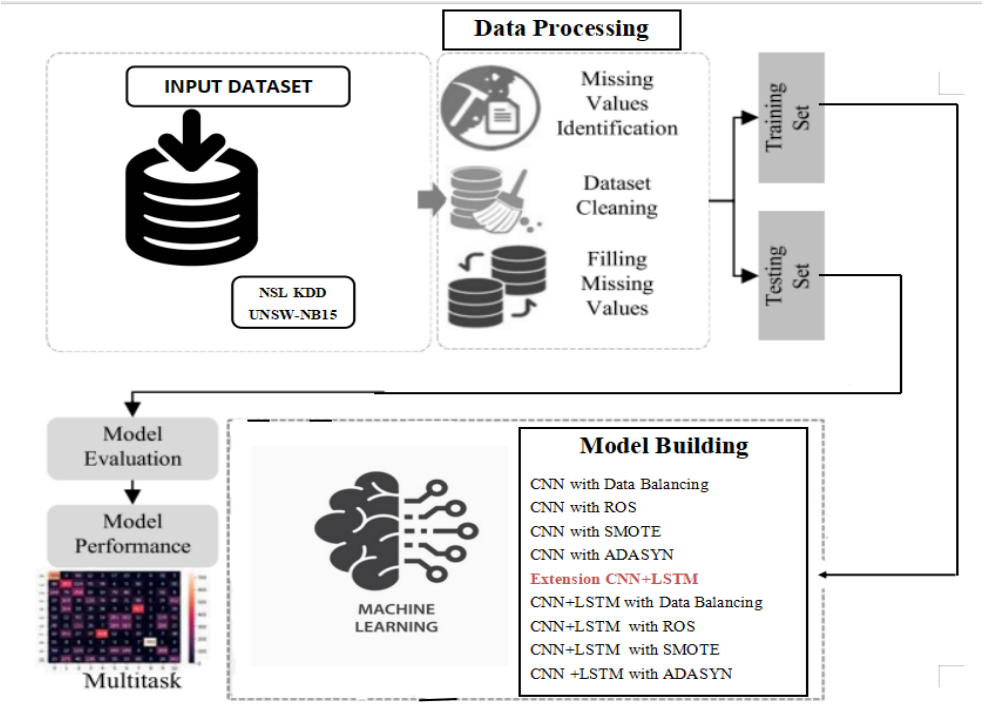

Figure1 : System Architecture

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 972-977

Model - 1: CNN with Data Balancing:

A Convolutional Neural Network (CNN) is a type of deep learning model designed for processing structured grid-like data, such as images. It works by using convolutional layers to extract features from input data, followed by pooling layers to reduce dimensionality and increase computational efficiency. To improve accuracy we are balancing the data. CNNs employ activation functions like ReLU to introduce non-linearity, and they use techniques like dropout to prevent overfitting. The final layers typically include fully connected layers for classification or regression tasks, making CNNs highly effective for tasks like image recognition, object detection, and image segmentation.

Model-2: CNN with SMOTE:

A CNN with SMOTE integrates synthetic data generation with convolutional neural networks to tackle class imbalance. SMOTE augments the training set by creating synthetic samples for minority classes, enhancing model performance by providing more balanced data representation. This approach helps CNNs learn from a diverse range of examples, particularly in situations with imbalanced class distributions, leading to improved accuracy and robustness in classification tasks.

Model-3: CNN with ROS:

A CNN with ROS (Random Over-sampling with Replacement) combines deep learning with data resampling to handle class imbalance. ROS randomly duplicates samples from minority classes, increasing their representation in the training set. In a CNN with ROS, the network architecture remains standard, but the training data is modified to have a more balanced class distribution. This approach helps CNNs learn from a broader range of examples, improving their ability to classify minority classes accurately.

Model-4: CNN with AdaSyn:

A CNN with ADASYN (Adaptive Synthetic Sampling) integrates deep learning with a data resampling technique designed for imbalanced datasets. ADASYN generates synthetic samples for minority classes based on their difficulty of classification, focusing more on challenging instances. In a CNN with ADASYN, the network architecture is conventional, but the training data is augmented using ADASYN to create a more balanced dataset. This approach improves the CNN's ability to learn from complex and underrepresented examples, enhancing its performance in classifying minority classes accurately.

Model-5: CNN-LSTM with Data Balancing:

A CNN-LSTM is a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This fusion leverages CNNs' ability to extract spatial features from data like images or sequences, followed by LSTM's capability to capture temporal dependencies and long-range dependencies in sequential data. CNNs process the input data and extract relevant features, which are then fed into LSTM layers for sequential modeling, making CNN-LSTM models effective for tasks like time series prediction, video analysis, and natural language processing. To improve accuracy we are balancing the data.

Model-6 CNN-LSTM with SMOTE:

Combines CNN's spatial feature extraction with LSTM's sequential modeling, augmented by SMOTE's synthetic minority data generation. Enhances learning from imbalanced datasets, improving accuracy in tasks like time series prediction and sequential classification.

Model-7 CNN-LSTM with ROS:

Integrates CNN's feature extraction with LSTM's sequential modeling, enhanced by ROS's random duplication of minority class samples. Improves model robustness in handling class imbalance, benefiting applications such as time series analysis and sequential pattern recognition.

Model-8 CNN-LSTM with ADASYN:

Fuses CNN's spatial feature learning with LSTM's sequential processing, utilizing ADASYN's adaptive synthetic sampling for minority classes. Enhances model generalization and performance, especially in scenarios with challenging and imbalanced data distributions, such as anomaly detection and event prediction.

5. RESULTS

Analysis of data is done by using deep Learning Models with different data balancing techniques and recurrent models which helps us to choose the best model in order to predict Network Intrusion. We have used different Models such as CNN with data balancing, CNN with SMOTE, CNN with ROS, CNN with ADASYN, CNN-LSTM with data balancing, CNN-LSTM with SMOTE, CNN-LSTM with ROS and CNN-LSTM with ADASYN.

In the bar plot, the accuracy of each model has been plotted.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 972-977

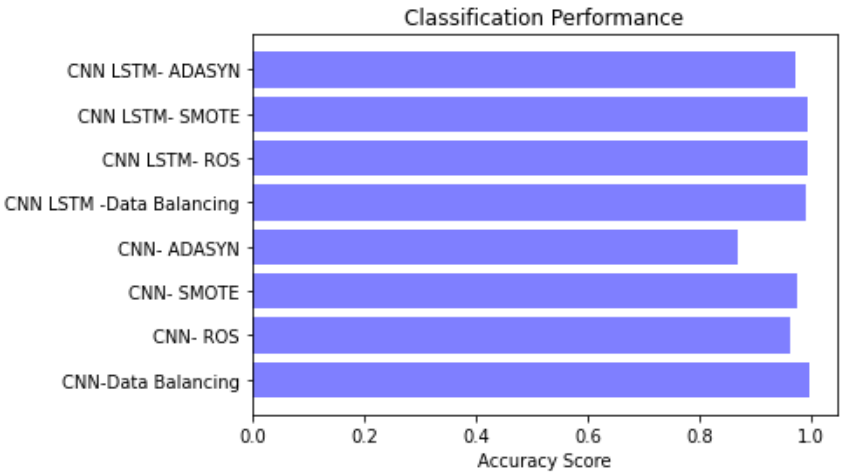

Figure 2: Accuracy depicted on NSL-KDD dataset

- The Accuracy score for testing data using CNN-Data Balancing is 99.5%.
- The Accuracy score for testing data using CNN- ROS is 96.3%.
- The Accuracy score for testing data using CNN- SMOTE is 97.3%.
- The Accuracy score for testing data using CNN- ADASYN is 86.9%.
- The Accuracy score for testing data using CNN LSTM -Data Balancing is99.0%.
- The Accuracy score for testing data using is CNN LSTM- ROS 99.4%.
- The Accuracy score for testing data using CNN LSTM- SMOTE is 99.4%.
- The Accuracy score for testing data using CNN LSTM-ADASYN is 97.2%.

Table 1: Performance Comparison of All Algorithms Implemented

	ML Model	Accuracy	f1_score	Recall	Precision
0	CNN-Data Balancing	0.995	0.996	0.995	0.997
1	CNN- ROS	0.963	0.963	0.963	0.963
2	CNN- SMOTE	0.973	0.973	0.973	0.973
3	CNN- ADASYN	0.869	0.873	0.869	0.88
4	CNN LSTM -Data Balancing	0.99	0.991	0.99	0.993
5	CNN LSTM- ROS	0.994	0.994	0.994	0.994
6	CNN LSTM- SMOTE	0.994	0.994	0.994	0.994
7	CNN LSTM- ADASYN	0.972	0.972	0.972	0.973

6. CONCLUSION

We proposed NID system that addresses the issue of imbalanced datasets and uses Convolutional Neural Networks (CNN) to classify different attack types performs well in the minority classes while maintaining an good recall on the binary classification task.

The proposed system is compared to other systems that use different techniques for data balancing, and the results show that the proposed system outperforms them. Future work can focus on improving the proposed system's performance by using more advanced techniques for data balancing and feature extraction.

7. REFERENCES

[1] Devi, M. S., Mahammad, F. S., Bhavana, D., Sukanya, D., Thanusha, T. S., Chandrakala, M., & Swathi, P. V. (2022).” Machine Learning Based Classification And Clustering Analysis Of Efficiency Of Exercise Against Covid-19 Infection.” Journal Of Algebraic Statistics, 13(3), 112-117.

[2] Devi, M. M. S., & Gangadhar, M. Y. (2012).” A Comparative Study Of Classification Algorithm For Printed Telugu Character Recognition.” International Journal Of Electronics Communication And Computer Engineering, 3(3), 633-641.

[3] Devi, M. S., Meghana, A. I., Susmitha, M., Mounika, G., Vineela, G., & Padmavathi, M. Missing Child Identification System Using Deep Learning.

[4] Kumar, M. S., Harika, A., Sushama, C., & Neelima, P. (2022). Automated Extraction Of Non-Functional Requirements From Text Files: A Supervised Learning Approach. Handbook Of Intelligent Computing And Optimization For Sustainable Development, 149-170.

[5] Devi, M. S., Poojitha, M., Sucharitha, R., Keerthi, K., Manideepika, P., & Vasudha, C. Extracting And Analyzing Features In Natural Language Processing For Deep Learning With English Language.

[6] B.Krishna Naga Deepthi, Dr.M.V.Subramanyam," Analysis And Optimization Of Power And Area Of Domino Full Adder And Its Applications", Iosr Journal Of Electronics And Communication Engineering, Vol.10,No.3,Pp.55-63,2015.

[7] Y.Murali Mohan Babu, Dr.M.V.Subramanyam,M.N. Giri Prasad," A New Approach For Sar Image Denoising", International Journal Of Electrical And Computer Engineering, Vol.5,No.5,Pp.984-991,2015. (Scopus Indexed)

[8] Ch.Nagaraju, Dr.Anil Kumar Sharma, Dr.M.V.Subramanyam," A Review On Ber Performance Analysis And Papr Mitigation In Mimo Ofdm Systems", International Journal Of Engineering Technology And Computer Research, Vol.3,No.3,Pp.237-238, June, 2015.

[9] D.Lakshmaiah, Dr.M.Subramanyam, Dr.K.Satya Prasad," Design Of Low Power 4- Bit Cmos Braun Multiplier Based On Threshold Voltage Techniques", Global Journal Of Research In Engineering, Vol.14(9),Pp.1125-1131,2014.

[10] R Sumalatha, Dr.M.Subramanyam, "Image Denoising Using Spatial Adaptive Mask Filter", Ieee International Conference On Electrical, Electronics, Signals, Communication &Amp; Optimization (Eesco-2015), Organized Byvignans Institute Of Information Technology, Vishakapatnam, 24 Th To 26th January 2015. (Scopus Indexed)

[11] P.Balamurali Krishna, Dr.M.V.Subramanyam, Dr.K.Satya Prasad, "Hybrid Genetic Optimization To Mitigate Starvation In Wireless Mesh Networks", Indian Journal Of Science And Technology,Vol.8,No.23,2015. (Scopus Indexed)