SMART INTRUSION DETECTION USING CNN & DATA BALANCING

[1]I.shalini,[2]M.vijay kumar,[3]M.kalyan sai,[4]B.venkatesh,[5]sk.jaffar sadik6 s.penchala sagar
[1]Associate Professor, [2,3,4,5,6]Student,
[1,2,3,4,5,6] Computer Science Engineering Department (Data Science),
[1,2,3,4,5,6]Geethanjali Institute of Science and Technology, Nellore, India

Abstract:
Cyber-security professionals often require automated systems to effectively detect and classify network intrusions. Identifying the type of attack is crucial for applying appropriate preventive measures to secure networks.

Several Machine Learning (ML) models have been proposed for Network Intrusion Detection (NID) systems. However, their performance depends on various factors. For example, when an ML model is trained on a highly imbalanced dataset, it may become biased toward frequently occurring attack types. Conversely, focusing only on improving minority class detection may negatively affect the performance on majority classes.

To address these challenges, this work proposes a Network Intrusion Detection (NID) system that handles imbalanced datasets using advanced data balancing techniques and employs Convolutional Neural Networks (CNN) for accurate attack classification. The proposed system is evaluated against other techniques such as Random Over-Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN).

The system is tested using benchmark datasets including NSL-KDD and BoT-IoT. Experimental results demonstrate that the proposed model achieves strong performance, particularly in detecting minority class attacks in binary classification tasks.
Additionally, the system obtains a high weighted average F1-score in multi-class classification using the BoT-IoT dataset.

I. INTRODUCTION

In the modern digital era, network security has become a critical concern for organizations and individuals due to the rapid growth of cyber-attacks. Cyber-security professionals require efficient and automated systems to monitor, detect, and classify network intrusions in real-time. However, traditional intrusion detection systems often struggle to accurately identify different types of attacks, especially when dealing with large and complex datasets.

One of the major challenges in Network Intrusion Detection (NID) systems is the imbalance in datasets, where some attack types are over-represented while others are under-represented. This imbalance can lead to biased Machine Learning (ML) models that perform poorly in detecting minority class attacks, which are often the most critical threats. As a result, existing systems may fail to provide reliable and accurate detection across all categories of intrusions.

To address these challenges, a smart intrusion detection system is required that can effectively handle imbalanced data and improve classification performance. This project proposes a **Smart**

Intrusion Detection System using Convolutional Neural Networks (CNN) and Data Balancing techniques. The system utilizes advanced methods such as Random Over-Sampling (ROS),

Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN) to balance the dataset and enhance detection accuracy.

Additionally, the system provides intelligent and automated analysis of network traffic by generating meaningful insights that help cyber-security professionals take preventive actions. The overall architecture is designed to be scalable, efficient, and easy to deploy. The backend of the proposed system is built using Python-based technologies, enabling seamless integration with Machine Learning models and ensuring maintainability for real-world applications.

By leveraging deep learning and data balancing approaches, the proposed system aims to provide improved performance in both majority and minority classes, ensuring better security and reliability in network environments.

II. LITERATURE REVIEW

[1] Studies on Machine Learning-based Network Intrusion Detection Systems (NIDS) highlight the effectiveness of algorithms such as Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) in detecting cyber-attacks. These models analyze network traffic patterns to classify normal and malicious activities. However, many of these approaches suffer from reduced accuracy when handling large-scale and complex datasets.

[2] Research on Deep Learning techniques, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), demonstrates improved performance in intrusion detection tasks. CNN models are capable of extracting spatial features from network data, leading to better classification results. However, these models require large amounts of balanced data for optimal performance.

[3] Several studies focus on data imbalance problems in intrusion detection datasets such as NSL-KDD and BoT-IoT. Techniques like Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) have been proposed to address this issue. While these methods improve minority class detection, they may sometimes introduce noise or overfitting.

[4] Research on Generative Adversarial Networks (GAN) shows their capability in generating synthetic data samples to balance datasets. GAN-based approaches have shown better performance compared to traditional oversampling techniques. However, training GAN models is complex and computationally expensive.

[5] Studies on hybrid models combining Machine Learning and Deep Learning techniques indicate improved detection accuracy and robustness. These systems leverage the strengths of multiple algorithms but often increase system complexity and require careful tuning.

[6] Explainable AI (XAI) techniques such as LIME and SHAP have been applied in intrusion detection systems to improve transparency and interpretability of model predictions. These methods help security analysts understand why a particular traffic pattern is classified as an attack.

[7] Research on real-time intrusion detection systems highlights the importance of efficient data processing and low-latency detection. Many systems struggle to maintain high accuracy while ensuring real-time performance in dynamic network environments.

Despite these advancements, a significant research gap exists: most existing systems fail to effectively handle imbalanced datasets while maintaining high accuracy across both majority and minority classes. The proposed system addresses this issue by integrating CNN with advanced data balancing techniques to achieve improved performance in network intrusion detection.

.

III. PROPOSED METHODOLOGY

A. SYSTEM ARCHITECTURE

The proposed system follows a modular and layered architecture for efficient network intrusion detection. The input layer collects network traffic data from datasets such as NSL-KDD and BoT-IoT. The preprocessing layer cleans and transforms the raw data by handling missing values, normalization, and feature selection.

The data balancing layer applies techniques such as Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN) to handle imbalanced datasets.

The core AI/ML layer uses Convolutional Neural Networks (CNN) to extract features and classify network traffic into normal and attack categories. The output layer displays the classification results and performance metrics such as accuracy, precision, recall, and F1-score.

B. KEY MODULES

1. **Data Collection Module:** This module collects benchmark datasets such as NSL-KDD and BoT-IoT, which contain various types of network traffic data including normal and attack records.
2. **Data Preprocessing Module:** This module performs data cleaning, normalization, encoding of categorical features, and feature selection to prepare the dataset for model training.
3. **Data Balancing Module:** To handle imbalanced data, techniques such as Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN) are applied to improve minority class representation.
4. **CNN-Based Classification Module:** This module uses Convolutional Neural Networks (CNN) to learn patterns from the dataset and classify network traffic into different attack categories or normal traffic.
5. **Model Training & Evaluation Module:** The system trains the CNN model using balanced datasets and evaluates its performance using metrics such as accuracy, precision, recall, and F1-score for both binary and multi-class classification.
6. **Detection & Prediction Module:** This module predicts whether incoming network traffic is normal or malicious and identifies the type of attack in real-time or test scenarios.
7. **Visualization & Result Analysis Module:** Graphs and charts are generated using visualization tools to analyze model performance and compare results across different data balancing techniques.

A. TOOLS & TECHNOLOGIES

Category	Tools / Technologies
Programming Language	**Python, JavaScript**
IDE	**Jupyter Notebook, VS Code**
ML Libraries	**Scikit-learn, Pandas, NumPy**
Deep Learning	**TensorFlow, PyTorch**
NLP	**HuggingFace Transformers, NLTK, SpaCy**
Web Scraping	**BeautifulSoup, Scrapy**
Backend Framework	**Flask (RESTful API)**
Frontend	**React.js / Streamlit**
Visualization	**Matplotlib, Seaborn, Plotly**

B. SYSTEM WORKFLOW

[2] The workflow begins with the collection of network traffic data from benchmark datasets such as NSL-KDD and BoT-IoT. The collected data is then passed to the preprocessing module, where data cleaning, normalization, and feature selection are performed to prepare the dataset for further processing.

[3] After preprocessing, the data balancing module is applied to handle class imbalance using techniques such as Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN). This step ensures that both majority and minority classes are properly represented.

[4] The balanced dataset is then fed into the Convolutional Neural Network (CNN) model, which extracts important features and performs classification of network traffic into normal or different attack categories.

[5] Once the model is trained, it is used for prediction, where new or unseen network data is classified in real-time or test scenarios. The system then evaluates the model performance using metrics such as accuracy, precision, recall, and F1-score.

[6] Finally, the results are visualized using graphs and charts to analyze the effectiveness of different data balancing techniques and the overall performance of the intrusion detection system.

III. RESULTS & DISCUSSION

A. SYSTEM PERFORMANCE ANALYSIS

The proposed intrusion detection system was evaluated using benchmark datasets such as NSL-KDD and BoT-IoT. The Convolutional Neural Network (CNN) model achieved high classification accuracy in detecting both normal and malicious network traffic. The application of data balancing techniques such as Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Generative Adversarial Networks (GAN) significantly improved the detection performance for minority class attacks.

The system demonstrated strong performance in terms of evaluation metrics such as accuracy, precision, recall, and F1-score. Particularly, the model achieved better recall values for minority classes, which are critical in identifying rare but dangerous cyber-attacks. The system also showed consistent results in both binary and multi-class classification tasks.

B. EFFECTIVENESS OF DATA BALANCING

The inclusion of data balancing techniques played a crucial role in improving the overall performance of the intrusion detection system. Compared to models trained on imbalanced datasets, the balanced dataset-based models showed significant improvement in detecting minority class attacks.

Among the techniques used, SMOTE and ADASYN provided better generalization, while GAN-based approaches generated more realistic synthetic samples, leading to improved classification accuracy. The combination of CNN with data balancing ensured that the model performs well across all classes without bias.

.

A. COMPARISON WITH EXISTING SYSTEMS

Feature	Traditional Portals	SkillBridge
Personalized Recommendations	No	Yes (AI-driven)
Unified Resource Access	No	Yes

B. LIMITATIONS

The current intrusion detection system has certain limitations. The performance of the model is highly dependent on the quality and size of the dataset used for training. If the dataset does not represent real-world network traffic accurately, the model's effectiveness may be reduced.

Data balancing techniques such as SMOTE, ADASYN, and GAN may introduce synthetic samples that can sometimes lead to overfitting or noise in the dataset. Additionally, training deep learning models like CNN requires high computational resources and time.

The system is currently tested on benchmark datasets such as NSL-KDD and BoT-IoT and may require further validation in real-time network environments. Deployment in large-scale networks would require additional infrastructure and optimization for handling real-time traffic efficiently.

C. DISCUSSION

The results confirm that the proposed system effectively addresses the major challenge of imbalanced datasets in network intrusion detection. By integrating Convolutional Neural Networks (CNN) with advanced data balancing techniques, the system significantly improves the detection of both majority and minority class attacks.

The model demonstrates better performance compared to traditional Machine Learning approaches, particularly in identifying rare and critical cyber-attacks. The combination of deep learning and data preprocessing techniques ensures improved accuracy, reliability, and robustness of the system.

Overall, the proposed system provides an efficient and scalable solution for enhancing network security and can be further extended for real-time intrusion detection applications.

V. CONCLUSION

1. **Improved Intrusion Detection Accuracy:** The proposed system effectively utilizes Convolutional Neural Networks (CNN) to achieve high accuracy in detecting network intrusions across both binary and multi-class classification tasks.
2. **Handling Imbalanced Datasets:** The integration of data balancing techniques such as Random Over-Sampling (ROS), SMOTE, ADASYN, and GAN significantly improves the detection performance for minority class attacks.
3. **Enhanced Detection of Rare Attacks:** The system successfully identifies low- frequency but critical cyber-attacks, which are often missed by traditional intrusion detection systems.
4. **Efficient Feature Learning:** CNN-based models automatically extract important features from network traffic data, reducing the need for manual feature engineering.
5. **Robust Performance:** The system demonstrates consistent performance across different datasets such as NSL-KDD and BoT-IoT, ensuring reliability and generalization.
6. **Scalability and Practical Usage:** The proposed model can be extended for real-time intrusion detection systems with proper infrastructure and optimization.
7. **Scope for Future Enhancements:** Future work may include real-time deployment, integration with cloud-based systems, improvement of GAN models for better data generation, and incorporation of Explainable AI techniques for better interpretability of model decisions.

REFERENCES

[1] [1] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A., "A Detailed Analysis of the NSL-KDD Dataset," Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1– 6.

[2] [2] Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B., "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," Future Generation Computer Systems, 2019, 100, pp. 779–796.

[3] [3] LeCun, Y., Bengio, Y., & Hinton, G., "Deep Learning," Nature, 2015, 521(7553), pp. 436–444.

[4] [4] Goodfellow, I., Bengio, Y., & Courville, A., "Deep Learning," MIT Press, 2016.

[5] [5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 2002, 16, pp. 321–357.

[6] [6] He, H., Bai, Y., Garcia, E. A., & Li, S., "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 2008, pp. 1322– 1328.

[7] [7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y., "Generative Adversarial Networks," Advances in Neural Information Processing Systems (NeurIPS), 2014.

[8] [8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011, 12, pp. 2825–2830.

[9] [9] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Zheng, X., "TensorFlow: A System for Large-Scale Machine Learning," Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2016, pp. 265–283.

[10] [10] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems (NeurIPS), 2019, 32, pp. 8024–8035.

[11] [11] Ribeiro, M. T., Singh, S., & Guestrin, C., "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[12] [12] McKinney, W., "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference (SciPy), 2010, pp. 56–61.

[13] [13] Richardson, L., & Ruby, S., "RESTful Web Services: Web Services for the Real World," O'Reilly Media, 2007.