

AI Based Network Intrusion Detection System

Subtitle as needed (*paper subtitle*)

MANNE JITHIN, RAKESH T A²

*Student , Information and Science , CMR Institute of Technology ,
Bengaluru , INDIA*

mannejithin01@gmail.com, rakeshta386@gmail.com

Abstract—The increasing use of digital technologies has created new challenges in protecting information and maintaining system security. Various cyberattacks, including Denial of Service, Probe, Remote to Local, User to Root, and other malicious activities, can affect systems and compromise sensitive data. Traditional security systems mainly identify threats by comparing network activities with previously known attack patterns. However, these methods may not be effective in detecting new or unfamiliar threats.

In this project, we developed an AI Based Network Intrusion Detection System using machine learning and deep learning techniques. The system uses network intrusion datasets such as NSL KDD, UNSW NB15, CICIDS2017, and IoT 23. Before analyzing the data, it is examined, organized, and converted into a suitable format for further processing.

Different techniques are applied to handle the imbalance in the number of samples available for each category. An autoencoder is used to learn meaningful representations of the input data. A transformer based model is used for classification, while CNN and DNN models are also employed for comparison.

The proposed system supports both binary and multiclass classification. Binary classification determines whether the network traffic is normal or malicious, whereas multiclass classification identifies the specific type of attack. SHAP is also used to explain the model predictions and understand which features influence the classification results. A web based dashboard is developed to display the predictions and model results.

The main objective of this project is to develop a complete AI based intrusion detection system that can process network data effectively, identify different types of cyberattacks, and provide clear and understandable results.

Keywords—Network Intrusion Detection System, Artificial Intelligence, Machine Learning, Deep Learning, Autoencoder, Transformer, Explainable Artificial Intelligence, Cybersecurity.

Introduction

Digital technologies support many everyday activities, including communication, education, business, and financial services. As people increasingly depend on digital systems, protecting these systems has become an important concern. Security threats may involve unauthorized access, service interruptions, information theft, and other harmful activities. Identifying suspicious network activity is therefore an important step in maintaining network security.

Traditional intrusion detection systems mainly depend on established rules and previously identified threats. These systems can recognize familiar attack patterns, but they may face difficulties in detecting new or modified attacks. Machine

learning based approaches provide an alternative method for identifying potential threats. Instead of relying entirely on fixed rules, trained models learn patterns from network data and use them to make predictions. Deep learning techniques can also help identify complex patterns in network traffic.

However, network intrusion datasets present certain challenges. One major problem is the unequal distribution of samples across different attack categories. Some attack classes contain a large number of records, while others have only a limited number of examples. This imbalance may affect the ability of a model to identify less represented attack categories accurately.

Another challenge is the large number of features present in network traffic data. Some features may provide limited information, whereas others may contain important details required for detecting malicious activities. Therefore, preparing the data and learning useful representations are important steps in developing an effective intrusion detection system.

To address these challenges, our project combines data preprocessing, class balancing, WGAN GP, Autoencoder, Transformer, CNN, DNN, and SHAP within a single system. WGAN GP is used to generate additional samples for attack categories with fewer records. An Autoencoder is applied to learn a compact representation of the input data. The Transformer model is used for classification, while CNN and DNN models provide additional approaches for comparison. SHAP is incorporated to explain the model predictions and identify the features that influence the classification results.

The proposed system supports two types of classification. Binary classification determines whether the network traffic represents normal or suspicious activity. Multiclass classification identifies the specific category of the detected attack. Through these approaches, the system aims to improve the identification of malicious network activities and provide results that are easier to understand.

I. PROPOSED SYSTEM

A. Overview of the Proposed System

The proposed AI Based Network Intrusion Detection System is designed to identify malicious activities in network traffic using machine learning and deep learning techniques. The system processes network intrusion datasets through several stages, including data preprocessing, class balancing, feature representation, classification, and prediction explanation.

B. Data Preprocessing

The collected network data is examined and prepared before being used for model training. This process includes handling missing values, removing duplicate records, converting categorical values into numerical representations, and scaling numerical features. These steps help prepare the data for further analysis.

C. Class Imbalance Handling

Network intrusion datasets may contain different numbers of samples for different attack categories. To address this issue, under sampling, over sampling, and WGAN GP techniques are applied. These techniques help improve the representation of underrepresented attack categories.

D. Feature Representation and Classification

An Autoencoder is used to learn a compact representation of the prepared network data. The extracted information is then passed to the classification models. The Transformer model is used for classification, while CNN and DNN models are used for comparison.

E. Classification and Explainability

The system supports binary and multiclass classification. Binary classification identifies whether the network traffic is normal or malicious, while multiclass classification identifies the specific attack category. SHAP is used to explain the model predictions and identify the features that influence the classification results.

F. Web Application

A web based dashboard is developed to display the classification results and model predictions. It provides an interface through which users can interact with the intrusion detection system and understand the detected network activities.

II. SYSTEM PREPARATION AND TECHNICAL DETAILS

Before developing the proposed AI Based Network Intrusion Detection System, the network intrusion datasets are collected and prepared for analysis. The data is examined to identify missing values, duplicate records, and inconsistencies. The required preprocessing techniques are applied to convert the data into a suitable format for machine learning and deep learning models.

The system uses different machine learning and deep learning techniques to identify network intrusions. The datasets are prepared before training, and the classification models are evaluated using the prepared data. The following sections describe the abbreviations and important technical concepts used in the proposed system.

A. Abbreviations and Acronyms

The proposed system uses several abbreviations related to network security, machine learning, and deep learning. Their full forms are provided below.

- NIDS — Network Intrusion Detection System

- AI — Artificial Intelligence
- ML — Machine Learning
- DL — Deep Learning
- CNN — Convolutional Neural Network
- DNN — Deep Neural Network
- SHAP — SHapley Additive exPlanations
- WGAN-GP — Wasserstein Generative Adversarial Network with Gradient Penalty
- DoS — Denial of Service
- IoT — Internet of Things

These abbreviations are used throughout the paper to describe the proposed system, its components, and the techniques applied for intrusion detection.

B. Dataset Preparation

The proposed system uses NSL-KDD, UNSW-NB15, CICIDS2017, and IoT-23 datasets. The collected data is cleaned and organized before being used for model training and evaluation. Categorical attributes are converted into numerical values, duplicate records are removed, and numerical features are scaled where required.

C. Model Preparation

The prepared datasets are processed using different machine learning and deep learning techniques. WGAN-GP is used to generate additional samples for underrepresented attack categories. An Autoencoder is applied to learn a compact representation of the input data. The resulting features are then used by the Transformer, CNN, and DNN models for classification.

D. Classification and Prediction Explanation

The system performs binary and multiclass classification to identify network attacks. Binary classification distinguishes normal traffic from malicious traffic, while multiclass classification identifies the specific attack category. SHAP is used to explain the predictions by identifying the features that influence the model output.

The classification results and related information are displayed through a web based dashboard, allowing users to understand the detected network activities.

III. USING THE TEMPLATE

The proposed AI-Based Network Intrusion Detection System is designed to analyze network traffic and identify potentially harmful activities using machine learning and deep learning techniques. The system uses the NSL-KDD, UNSW-NB15, CICIDS2017, and IoT-23 datasets to support intrusion detection across different types of network traffic.

The input data is first preprocessed to remove unnecessary records, handle missing values, convert categorical attributes into numerical form, and prepare the features for model training. Since the datasets contain an unequal number of samples across different classes, under-sampling, over-

sampling, and WGAN-GP are applied to address class imbalance.

An Autoencoder is then used to obtain a compact representation of the prepared data. The extracted features are analyzed using Transformer, CNN, and DNN models. The system performs binary classification to distinguish normal traffic from attack traffic and multiclass classification to identify the specific attack category.

SHAP is incorporated to explain the model predictions by showing the contribution of individual input features. The results are presented through a web-based application, allowing users to view the classification output and understand the detected network activity.

A. Authors and Affiliations

The authors of this research work belong to the same institution. Both authors are affiliated with the Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India.

1) Authors with a Single Affiliation

a) Manne Jithin

Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India.

b) Rakesh T A

Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India.

B. Identify the Headings

The headings in this research paper are organized to present the proposed Network Intrusion Detection System in a clear and structured manner. The paper uses component headings to identify major sections, such as the Abstract, Acknowledgment, and References, while text headings are used to organize the technical content of the proposed system.

The main sections of the paper are represented using uppercase Roman numerals. Subsections are arranged under their respective main sections using alphabetic labels. Further subdivisions are introduced only when required to explain the content in greater detail.

The heading structure is organized as follows:

1) Main Section Headings

Main sections are used to represent the major topics of the research paper. These include:

- Introduction
- Overview of Proposed Work
- System Flowchart
- Methodology
- Implementation
- Results and Discussion

2) Subsection Headings

Subsection headings are used to divide the main sections into smaller topics. For example, the Methodology section may

include subsections describing dataset preprocessing, data balancing, model development, and explainable AI techniques.

3) Formatting of Headings

The headings are formatted according to the IJAMRED template hierarchy. Main section headings use uppercase Roman numerals, while subsection headings use capital letters. Further subdivisions use numbered or alphabetic formats as specified by the template. This structure helps maintain consistency and improves the readability of the research paper.

C. Figures and Tables

1) *Positioning Figures and Tables:* The proposed Enhanced AI-Based Network Intrusion Detection System uses figures and tables to present the system's performance and detection results in a clear and understandable manner. The dashboard provides visual information about network traffic, detected attacks, model performance, and the features that influence the classification results.

2) *The figures are placed close to the relevant discussions in the paper. Figure captions are provided below the figures, while table titles are placed above the tables. Each figure is referred to in the text using the abbreviation "Fig." followed by its number.*

3) *The main dashboard displays important performance measures, including accuracy, weighted F1-score, false positive rate, and true positive rate. It also provides a live threat feed containing details such as traffic type, source IP address, model name, confidence score, and detection status.*

TABLE I. PERFORMANCE METRICS OF THE PROPOSED NIDS

Model Performance	Performance Value
Overall Accuracy	92.7%
Weighted F1-Score	0.915
False Positive Rate	3.3%
True Positive Rate	92.8%

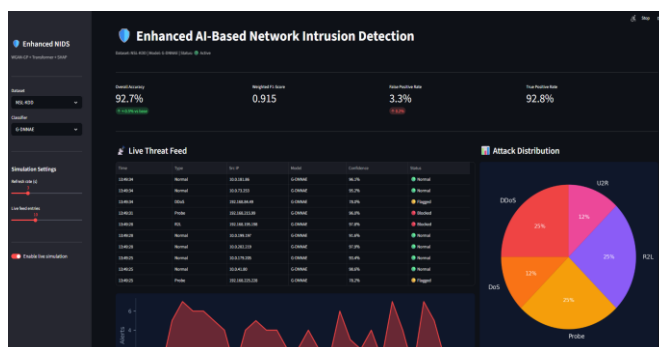


Fig. 1. Enhanced NIDS dashboard showing performance metrics and live threat monitoring.

The system also presents the distribution of detected attack

categories using a pie chart. The chart includes attack types such as DDoS, DoS, Probe, R2L, and U2R. An alert monitoring graph is used to display changes in the number of detected alerts over time, helping users observe variations in suspicious network activity.



Fig. 2. Attack distribution and alert monitoring in the proposed NIDS

The explainability component uses SHAP to show how an individual network traffic features influence the model of prediction. The visualization highlights the contribution of features such as source bytes, destination bytes, service, duration, flag, land, and protocol. This helps users understand which features support or oppose an attack prediction.

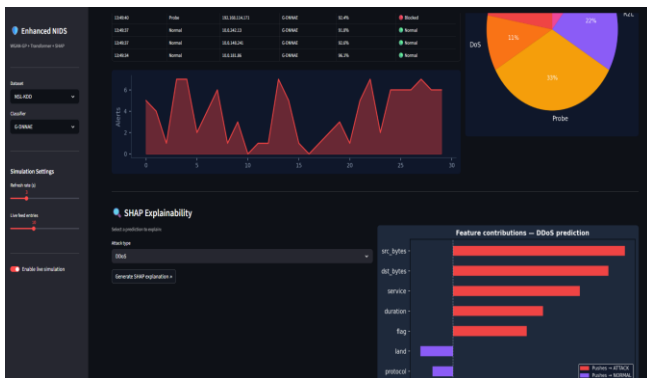


Fig. 3. SHAP-based feature contributions for a DDoS prediction.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their project guide, Dr. Manikandan H., for his valuable guidance, encouragement, and support throughout the development of this research work. The authors are also thankful to the faculty members of the Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, for their support and for providing the necessary facilities to carry out the project. Finally, the authors extend their appreciation to everyone who contributed directly or indirectly to the successful completion of this work.

REFERENCES

- [1] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, and D. Hong, "An enhanced AI-based network intrusion detection system using generative adversarial networks," *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2330–2345, Feb. 2023, doi: 10.1109/IIOT.2022.3211346.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [5] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [6] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.
- [7] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proceedings of the Military Communications and Information Systems Conference*, 2015.
- [8] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proceedings of the International Conference on Information Systems Security and Privacy*, 2018.
- [9] S. Garcia, A. Parmisano, and M. J. Ertürk, "IoT-23: A labeled dataset with malicious and benign IoT network traffic," *Stratosphere Laboratory*, 2020.
- [10] A. Arreche et al., "Explainable intrusion detection" [full title and publication details to be confirmed], 2024.