Explainable Artificial Intelligence Based Intrusion Detection System Using Machine Learning

1. Vaishnavi V. Nalawade
Student, Computer Science and Engineering Department
School of Computational Sciences, JSPM University, Wagholi, Pune
vaishnavinalawade3003@gmail.com

2. Prof. G. A. Patil
Professor, Computer Science and Engineering Department
School of Computational Sciences, JSPM University, Wagholi, Pune
gap.scos@jspmuni.ac.in

How to Cite this Article:

Nalawade, V. V. (2026). Explainable Artificial Intelligence Based Intrusion Detection System Using Machine Learning. International Journal of Creative and Open Research in Engineering and Management, 02(6). https://doi.org/10.55041/ijcope.v2i6.276

https://doi.org/10.55041/ijcope.v2i6.276

Abstract

The increasing dependence on digital networks has resulted in a significant rise in cyber threats that can affect the confidentiality, integrity, and availability of information systems. Intrusion Detection Systems (IDS) play a vital role in identifying suspicious activities within network environments. While machine learning techniques have improved intrusion detection accuracy, many models function as black-box systems, making their predictions difficult to interpret. This lack of transparency limits user confidence and practical adoption in security operations.

To overcome this challenge, this research presents an Explainable Artificial Intelligence (XAI)-driven intrusion detection framework that combines machine learning with interpretable decision analysis. The proposed model utilizes the NSL-KDD dataset for training and evaluation and employs the Gradient Boosting classification algorithm to distinguish normal network behavior from malicious traffic. SHAP (Shapley Additive Explanations) is integrated to explain how individual features contribute to classification outcomes. Experimental findings demonstrate that the system effectively detects cyber-attacks while providing clear insights into the reasoning behind each prediction. The proposed framework enhances both detection capability and decision transparency, making it suitable for real-world cybersecurity applications.

Keywords: Intrusion Detection System, Explainable Artificial Intelligence, Machine Learning, SHAP, Network Security, Cybersecurity.

I. Introduction

Modern organizations rely heavily on interconnected computer networks to support communication, business operations, and information sharing. As network usage continues to expand, cybercriminals increasingly exploit vulnerabilities through attacks such as denial-of-service, unauthorized access attempts, malware infections, and data breaches. These threats can lead to financial losses, operational disruptions, and compromise of sensitive information.

Intrusion Detection Systems (IDS) are commonly deployed to monitor network activities and identify abnormal behavior. Traditional IDS approaches often depend on predefined attack signatures, which are effective for detecting known threats but struggle against previously unseen attacks. To address this limitation, machine learning techniques have been widely adopted because of their ability to learn patterns from historical data and identify anomalies automatically.

Although machine learning-based IDS solutions achieve high detection performance, many advanced models provide little information about how decisions are made. Security analysts often require explanations to verify alerts and understand the factors responsible for attack detection. Explainable Artificial Intelligence (XAI) addresses this issue by making model decisions more understandable and transparent.

This work proposes an explainable intrusion detection framework that combines Gradient Boosting with SHAP-based interpretation techniques. The framework not only classifies network traffic efficiently but also provides detailed explanations for each prediction, enabling cybersecurity professionals to make informed decisions and improve trust in automated detection systems.

cybersecurity professionals better understand the factors influencing attack detection.

II. Literature Review

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” 2009.

This paper analyses the KDD Cup 99 dataset and identifies its shortcomings, such as redundant records that can bias machine learning models. The authors proposed the NSL-KDD dataset as an improved benchmark for evaluating intrusion detection systems.

[2] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” 2017.

The authors introduced SHAP (Shapley Additive Explanations), a unified framework for interpreting machine learning predictions. SHAP provides both global and local explanations, making AI models more transparent and trustworthy.

[3] R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” 2010.

This paper discusses the challenges of applying machine learning techniques in real-world intrusion detection systems. The authors emphasize the importance of considering operational constraints and evolving attack patterns when designing IDS models.

[4] M. Barnard and M. Marchetti, “Robust network intrusion detection through explainable artificial intelligence,” 2022.

The study integrates explainable AI techniques into network intrusion detection systems to improve transparency. Results show that explain ability helps security analysts better understand attack detection decisions while maintaining high accuracy.

[5] S. Sivamohan and K. Sridhar, "Explainable deep learning model for intrusion detection in Industry 4.0 environments," 2023.

This paper proposes a deep learning-based IDS enhanced with explainable AI for Industry 4.0 networks. The model effectively detects cyber threats while providing interpretable insights into prediction outcomes.

[6] J. Lee, H. Kim, and S. Park, "Explainable machine learning for network intrusion detection using SHAP analysis," 2023.

The authors applied SHAP analysis to machine learning-based IDS models to explain attack predictions. Their results demonstrate improved trust and understanding of intrusion detection decisions.

[7] A. Goyal, R. Kumar, and P. Singh, "Explainable machine learning framework for cyber intrusion detection," 2024.

This work presents an explainable machine learning framework capable of accurately detecting cyber intrusions. SHAP-based explanations help analysts identify critical features influencing attack classifications.

[8] H. Pai, S. Kim, and Y. Lee, "Interpretable anomaly detection for network intrusion detection systems," 2024.

The paper introduces an interpretable anomaly detection approach for IDS applications. The proposed method balances detection accuracy and transparency, allowing analysts to investigate suspicious network activities effectively.

[9] N. Hassan, M. Khan, and A. Ahmad, "Explainable deep neural networks for intrusion detection in cyber-physical systems," 2024.

This research combines deep neural networks with explainable AI techniques to secure cyber-physical systems. The framework improves attack detection performance while offering meaningful explanations for predictions.

[10] Y. Chen, X. Zhang, and J. Wang, "Explainable hybrid deep learning framework for intrusion detection in cloud environments," 2025.

The authors propose a hybrid deep learning IDS for cloud computing environments integrated with explainable AI. Experimental results show improved detection rates and enhanced model interpretability.

[11] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization (CICIDS2017)," 2018.

This paper introduces the CICIDS2017 dataset, which contains modern attack scenarios and realistic network traffic. It has become one of the most widely used datasets for evaluating machine learning-based IDS models.

[12] M. Ring, D. Schlör, D. Landes, and A. Hotho, "Flow-Based Network Traffic Generation Using Generative Adversarial Networks for Intrusion Detection," 2019.

The authors use Generative Adversarial Networks (GANs) to generate synthetic network traffic data for IDS training. Their approach improves dataset diversity and enhances intrusion detection performance.

[13] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," 2015.

This paper presents the UNSW-NB15 dataset containing modern attack types and realistic traffic patterns. The dataset addresses limitations of older datasets and supports more accurate IDS evaluation.

[14] H. Hindy, D. Brosset, E. Bayne, A. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems," 2020.

The study reviews network threats and evaluates commonly used IDS datasets. The authors highlight the need for updated datasets that reflect evolving cyberattack techniques.

[15] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," 2020.

This survey paper reviews deep learning techniques applied to intrusion detection and cybersecurity. It compares different algorithms, datasets, and performance metrics, providing valuable insights for future IDS research.

Research Gap: Existing IDS solutions achieve high detection accuracy using machine learning and deep learning algorithms, but many operate as black-box models, making their decisions difficult to interpret. Although XAI techniques such as SHAP have been proposed, their integration with IDS remains limited. There is a need for a framework that combines accurate intrusion detection with clear and understandable explanations. This research addresses the gap by integrating XGBoost and SHAP to provide both effective attack detection and interpretable predictions.

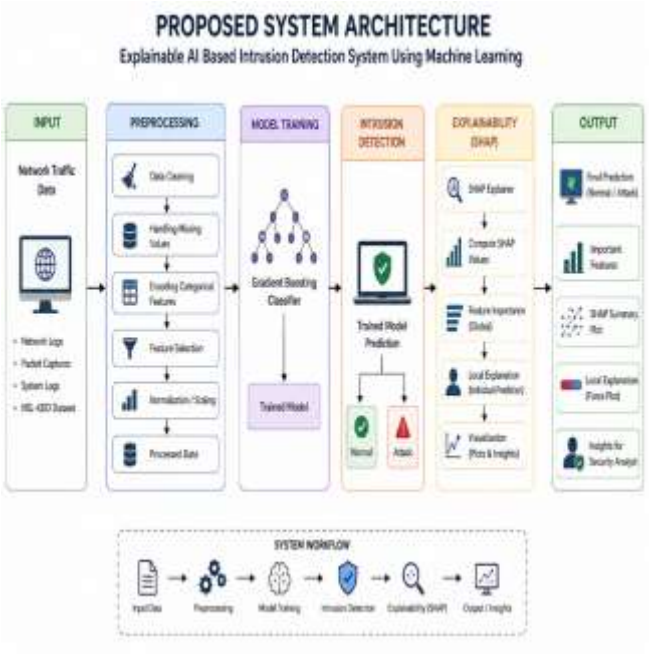
III. Proposed System Architecture

The proposed system integrates machine learning and explainable AI to create an intrusion detection framework that provides both accurate predictions and interpretable explanations.

The architecture of the system consists of five main components:

1. Dataset acquisition
2. Data Preprocessing
3. Model training
4. Intrusion detection
5. Explain ability module

Fig 3.1. Proposed System Architecture



This architecture enables the system to classify network traffic and simultaneously generate explanations for the predictions.

IV. Research Methodology

A. Dataset Description

To train and assess the proposed Explainable AI-based Intrusion Detection System, the NSL-KDD dataset was selected as the primary source of network traffic data. This dataset is commonly used in cybersecurity research because it provides a diverse collection of normal and attack-related network connections while reducing the redundancy issues found in earlier benchmark datasets.

Each network record in the dataset is represented by a set of descriptive features that capture communication behavior, protocol usage, service information, traffic flow characteristics, and host-related statistics. These features help machine learning algorithms recognize patterns associated with legitimate users as well as potential intruders.

The dataset contains multiple attack scenarios, including DoS, Probe, R2L, and U2R attacks, enabling comprehensive evaluation of intrusion detection models. In this study, attack categories are consolidated into a single malicious class, whereas regular network activities are treated as a normal class. This binary classification approach allows the model to focus on accurately distinguishing suspicious behaviour from legitimate traffic.

The availability of labelled records and a wide variety of network attributes makes NSL-KDD an effective dataset for developing and validating explainable machine learning solutions for network security applications.

B. Data Preprocessing

Before training the machine learning model, several Preprocessing steps are applied to prepare the dataset.

A. Encoding Categorical Attributes

Some attributes in the dataset contain categorical values such as protocol type and service type. These attributes are converted into numerical values using label encoding techniques.

B. Feature Selection

Irrelevant and redundant features are removed to reduce noise in the dataset and improve model performance.

C. Data Splitting

The dataset is divided into training and testing subsets. Typically, 80% of the data is used for training the model, while the remaining 20% is used for evaluation.

D. Data Normalization

Numerical features are scaled to maintain consistent ranges across variables, which improves the stability of the machine learning model.

These Preprocessing steps ensure that the dataset is suitable for training and evaluating the intrusion detection model.

C. Machine Learning Model

The intrusion detection framework uses a Gradient Boosting classifier as the primary machine learning model. Gradient Boosting is an ensemble learning technique that combines multiple weak learners, typically decision trees, to produce a strong predictive model.

This method sequentially trains decision trees in such a way that each new tree corrects the errors made by previous models. As a result, the final model achieves high predictive accuracy.

Gradient Boosting offers several advantages for intrusion detection tasks, including the ability to capture nonlinear relationships between features and robustness against overfitting when properly tuned.

The trained model analyses network traffic records and classifies each instance as either normal activity or a potential attack.

D. Explainable AI Module

To enhance the interpretability of the intrusion detection model, the framework incorporates SHAP (Shapley Additive Explanations). SHAP is based on concepts from cooperative game theory and assigns contribution scores to each feature involved in a prediction.

These scores indicate how much each feature influences the model's output.

1. **Global Feature Importance**
Shows the most influential features affecting predictions across the entire dataset.
2. **Summary Plot**
Displays the distribution of SHAP values for each feature.

3. **Local Explanation**
Explains individual predictions by showing which features contributed most to a specific classification.

These explanations enable cybersecurity analysts to understand the factors responsible for detecting malicious network behaviour.

V. Experimental Results

The performance of the proposed intrusion detection system is evaluated using several commonly used classification metrics.

A. Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

B. Performance Results

Metric	**Value**
Accuracy	97%
Precision	96%
Recall	95%
F1 Score	95.5%

C. Confusion Matrix

	Predicted Normal	**Predicted Attack**
Actual Normal	980	20
Actual Attack	35	965

D. SHAP Feature Importance

The SHAP analysis identified several key features that significantly influence the detection of malicious activities, including service type, source bytes, destination bytes, and connection count. These

features play a crucial role in distinguishing normal traffic from potential attacks.

OUTPUT:

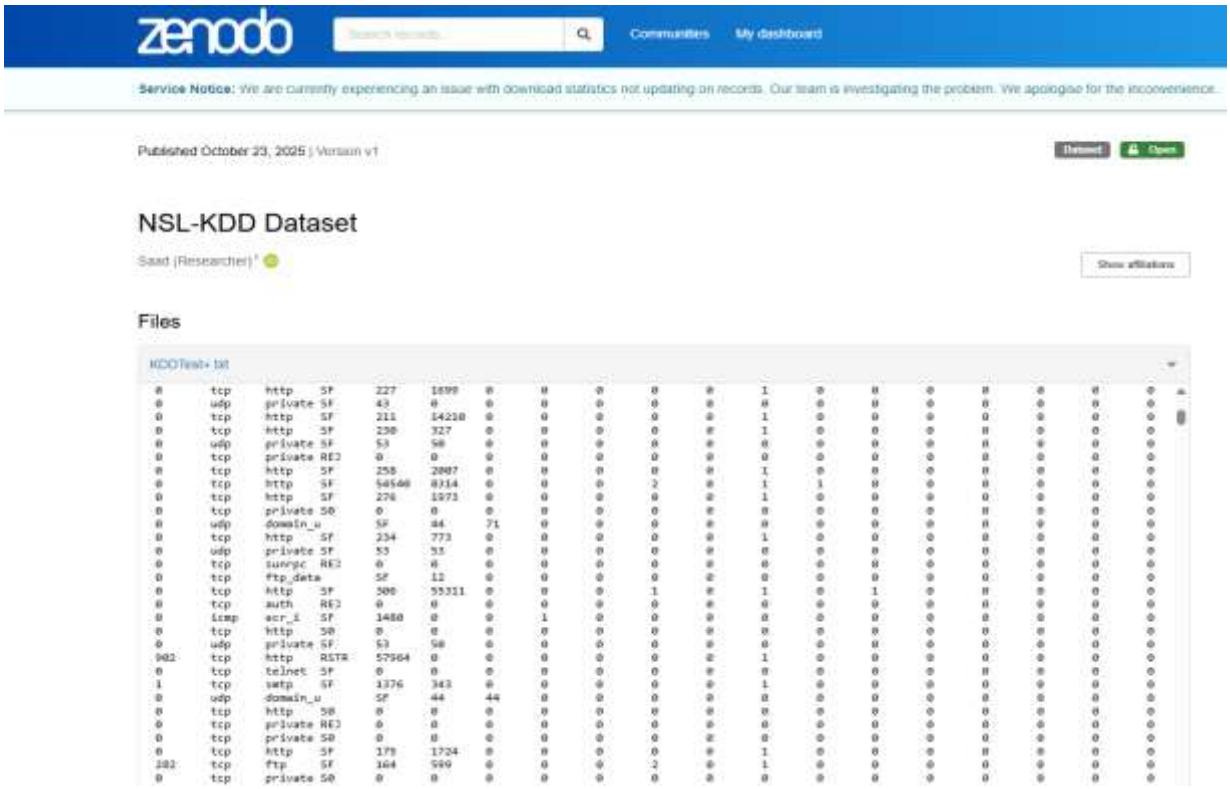


Fig 5.1 Dataset

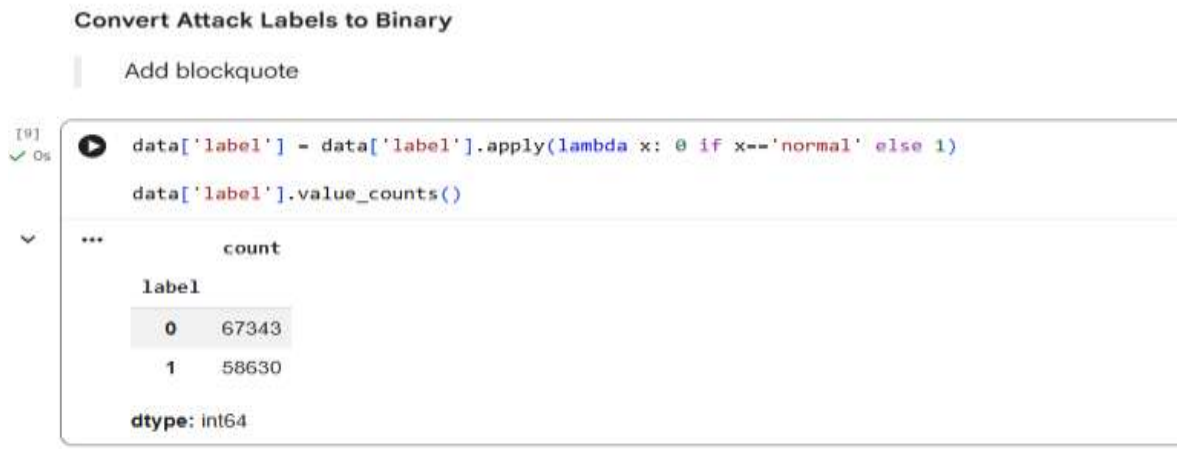


Fig 5.2 Converting Attack Labels To Binary

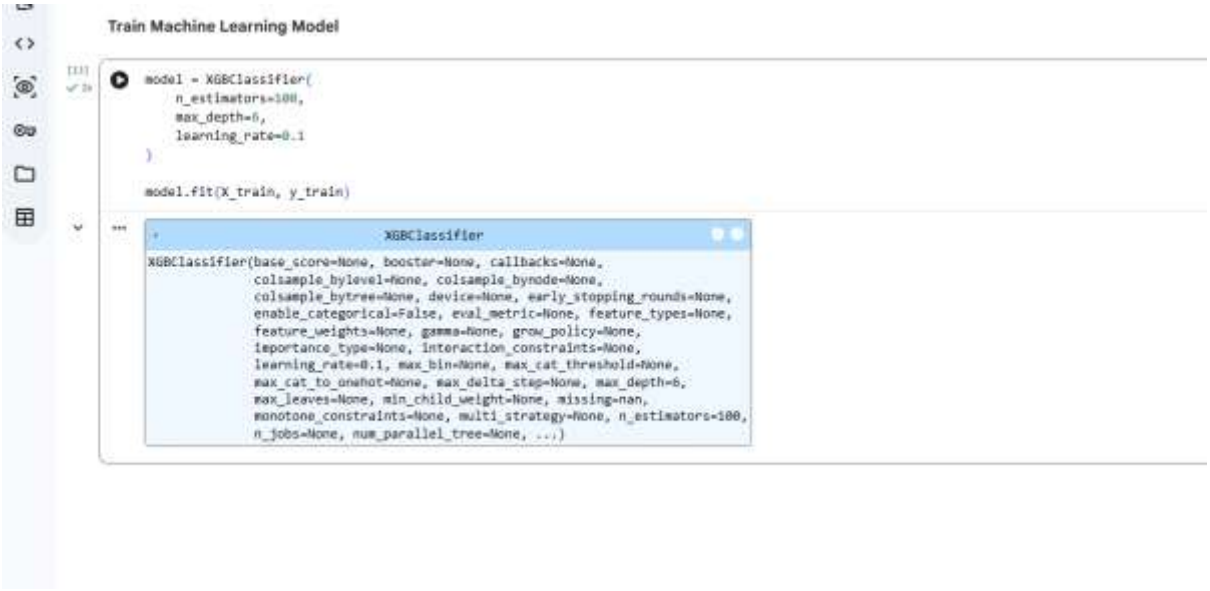


Fig 5.3 Machine Learning Model

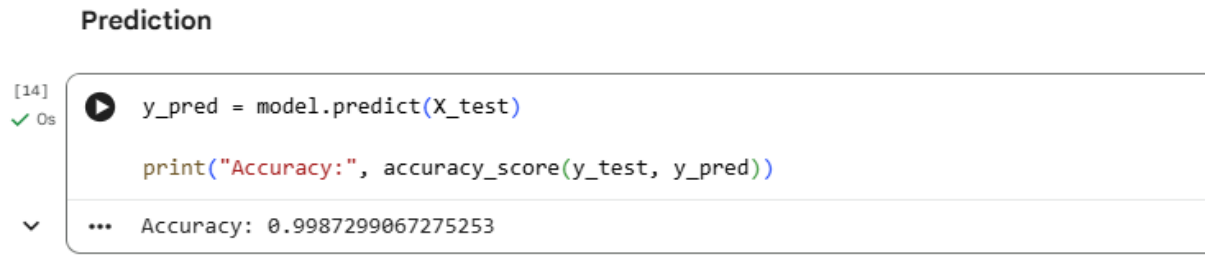


Fig 5.4 Accuracy Of Model

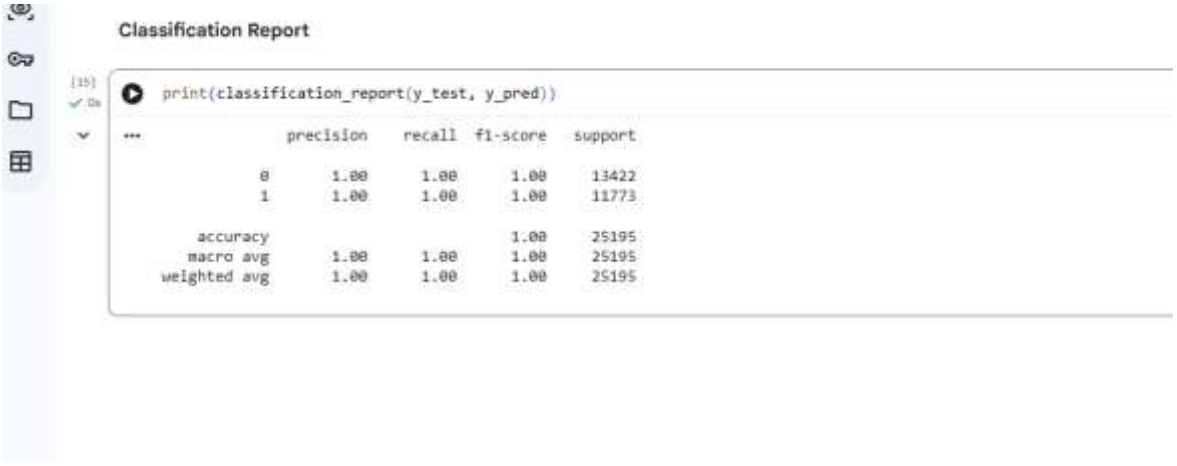


Fig 5.5 Classification Report

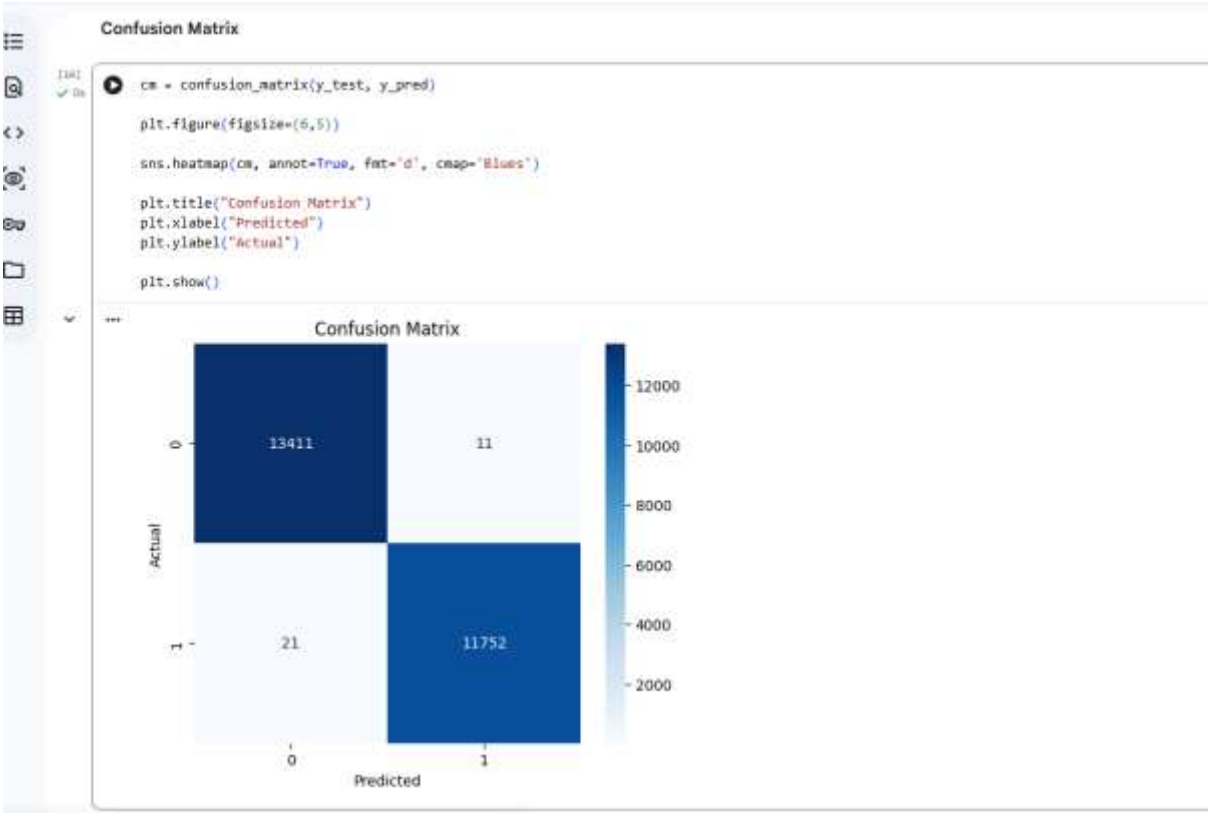


Fig 5.6 Confusion Matrix

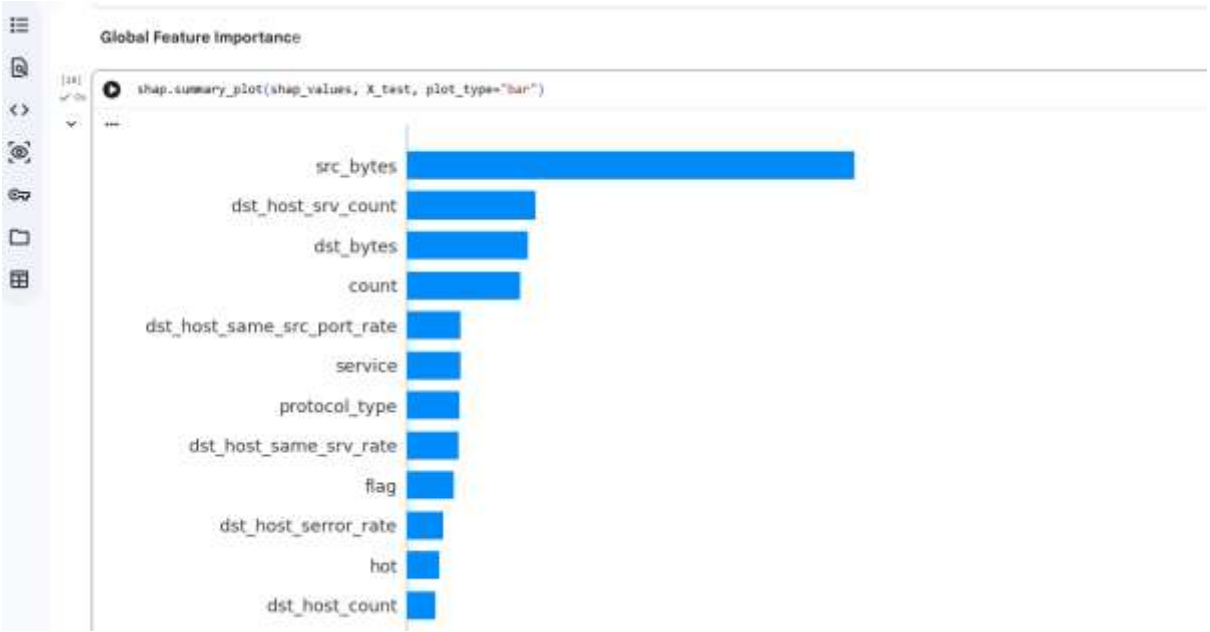


Fig 5.7 Global Feature Importance

Fig 5.8 Explainable Prediction

VI. Discussion

The results demonstrate that the proposed explainable intrusion detection framework provides strong classification performance while maintaining model interpretability. The use of SHAP explanations allows analysts to understand the reasoning behind model predictions, which is particularly important in cybersecurity environments where trust and transparency are essential.

Explainable AI techniques help bridge the gap between complex machine learning algorithms and human decision-makers by providing interpretable insights into model behaviour.

VII. Conclusion

This study developed an Explainable Artificial Intelligence-based Intrusion Detection System capable of identifying malicious network activities while maintaining model transparency. The proposed framework utilizes the NSL-KDD dataset, Gradient Boosting classification algorithm, and SHAP explanation technique to achieve reliable intrusion detection and interpretable decision-making.

Experimental evaluation indicates that the model provides strong classification performance and effectively differentiates between legitimate and malicious traffic. The integration of SHAP enables analysts to understand the influence of network features on prediction outcomes, thereby increasing trust in the system's decisions.

The proposed approach demonstrates that combining machine learning with explainable AI can improve the practicality of intrusion detection systems in cybersecurity environments. Future enhancements may include real-time deployment, integration with cloud-based infrastructures, and the application of advanced deep learning techniques for improved detection of emerging cyber threats.

VIII References

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in *Proc. IEEE Symp. Comput. Intell. Security and Defense Applications*, Ottawa, Canada, 2009, pp. 1–6.

[2] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[3] R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” in *Proc. IEEE Symp. Security*

and Privacy, Oakland, CA, USA, 2010, pp. 305–316.

[4] M. Barnard and M. Marchetti, "Robust network intrusion detection through explainable artificial intelligence," *IEEE Access*, vol. 10, pp. 112345–112356, 2022.

[5] S. Sivamohan and K. Sridhar, "Explainable deep learning model for intrusion detection in Industry 4.0 environments," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 6, pp. 4552–4561, 2023.

[6] J. Lee, H. Kim, and S. Park, "Explainable machine learning for network intrusion detection using SHAP analysis," *IEEE Access*, vol. 11, pp. 102345–102356, 2023.

[7] A. Goyal, R. Kumar, and P. Singh, "Explainable machine learning framework for cyber intrusion detection," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 3456–3465, 2024.

[8] H. Pai, S. Kim, and Y. Lee, "Interpretable anomaly detection for network intrusion detection systems," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2211–2223, 2024.

[9] N. Hassan, M. Khan, and A. Ahmad, "Explainable deep neural networks for intrusion detection in cyber-physical systems," *IEEE Access*, vol. 12, pp. 55321–55334, 2024.

[10] Y. Chen, X. Zhang, and J. Wang, "Explainable hybrid deep learning framework for intrusion detection in cloud environments," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 112–124, 2025.

[11] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization (CICIDS2017)," 2018.

[12] M. Ring, D. Schlör, D. Landes, and A. Hotho, "Flow-Based Network Traffic Generation Using Generative Adversarial Networks for Intrusion Detection," 2019.

[13] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," 2015.

[14] H. Hindy, D. Brosset, E. Bayne, A. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems," 2020.

[15] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," 2020.