International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05951
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Explainable Artificial Intelligence (XAI)-Based Intrusion Detection and Classification in Network Traffic

Gorre Renuka [1], Kudikilla Chandana [2], Anikha Tamkinath [3] Under the guidance of Mr. S. Chiranjeevi
Assistant professor, CSE Dept, Methodist College of Engineering and Technology

Gorre Renuka Computer Science and Engineering, Methodist College of Engineering and Technology, (OU Affiliated) Hyderabad, India
renukagorre80@gmail.com

Kudikilla Chandana Computer Science and Engineering, Methodist College of Engineering and Technology, (OU Affiliated) Hyderabad, India
chandanakudikilla@gmail.com

Anikha Tamkinath Computer Science and Engineering, Methodist College of Engineering and Technology, (OU Affiliated) Hyderabad, India
anikhatamkinath5@gmail.com

Mr. S. Chiranjeevi, Assistant Professor,
Methodist College of Engineering and Technology, (OU Affiliated) Hyderabad, India

ABSTRACT

The rapid growth of digital communication has increased the volume and complexity of network traffic, making modern computer networks more vulnerable to various cyberattacks and data breaches. To safeguard these networks, Intrusion Detection Systems (IDS) are widely used to detect malicious activities within network traffic. Traditional IDS models rely on machine learning algorithms to classify traffic as normal or malicious; however, they often act as black-box models, offering little to no insight into how predictions are made. This lack of interpretability reduces user trust and limits their real-world applicability in critical security environments. The proposed project introduces an Explainable Artificial Intelligence (XAI) approach that integrates XGBoost with SHAP (SHapley Additive exPlanations) to build a transparent and interpretable intrusion detection system. The model not only detects and classifies intrusions accurately but also explains the contribution of each feature to the final prediction, enabling better understanding and trust. The system is expected to achieve high detection accuracy while providing clear, visual explanations for each classification. This enhances decision-making for network administrators and contributes to a more secure and transparent cybersecurity framework.

Keywords— Intrusion Detection System (IDS), Explainable Artificial Intelligence (XAI), Machine Learning (ML), XGBoost, SHAP (SHapley Additive exPlanations), Cybersecurity, Network Traffic Classification, Network Security Data Classification and Model Interpretability.

I. INTRODUCTION

The rapid growth of internet connectivity and the increasing sophistication of cyberattacks have made Intrusion Detection Systems (IDS) an essential component of modern network security. Traditional IDS models relied on signature-based detection, which could only identify known threats. To overcome these limitations, machine learning (ML) and deep learning (DL) approaches have emerged as powerful techniques for identifying complex and evolving attack patterns in network traffic. However, recent literature consistently highlights that while ML/DL techniques improve detection accuracy, they suffer from major limitations such as poor interpretability, dataset imbalance, adversarial vulnerability, and lack of generalization across real-world environments.

A review of multiple surveys published between 2021 and 2025 shows three major research gaps. First, ML and DL-based IDS models often behave as *black boxes*, providing no insight into how a particular attack is detected. This lack of transparency reduces analyst trust and makes these systems unsuitable for high-risk cybersecurity domains. Second, many IDS studies rely on outdated, imbalanced, or synthetic datasets such as NSL-KDD and KDDCup99, which fail to reflect modern attack behaviors, leading to unreliable performance in deployment scenarios. Third, existing surveys emphasize that IDS models struggle to detect low-frequency attack types like R2L and U2R, resulting in high false negatives and reduced reliability.

To address these limitations, recent research trends point toward Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME, which offer feature-level explanations of model predictions. Survey papers also highlight the growing demand for hybrid IDS models that combine the accuracy of ensemble ML classifiers with the interpretability of XAI methods. As cyberattacks become more dynamic and adversarial, explainable IDS frameworks are necessary to ensure transparency, accountability, and effective decision support for securityanalysts.

This comparative review examines six to seven recent survey papers covering ML-based IDS, DL-based IDS, XAI-enhanced IDS, and adversarial robustness. By analyzing their methodologies, findings, limitations, and research gaps, this paper provides a unified perspective on the evolution of intrusion detection research and identifies the critical need for interpretable and reliable IDS models, motivating the development of XAI-based detection frameworks.

II. LITERATURE SURVEY

A. Machine Learning–Based Intrusion IDS Approaches Machine learning has been a foundational approach in intrusion detection due to its speed, scalability, and ability to classify structured network traffic. Kaissar (2022) provided a comprehensive analysis of several traditional ML algorithms including Random Forest, Support Vector Machine, Decision Trees, and K-Nearest Neighbors. The review emphasized that these models perform efficiently on benchmark datasets such as NSL-KDD and CICIDS2017, especially for common attacks like DoS and Probe. Random Forest, in particular, demonstrated strong robustness due to its ensemble nature, while SVM showed superior boundary-based classification for binary intrusion scenarios. The study also discussed how feature engineering and dimensionality reduction methods improve ML performance by filtering redundant attributes.

However, the same survey highlighted significant limitations associated with ML-based IDS. One major issue identified is the difficulty in detecting rare attack categories such as Remote-to- Local (R2L) and User-to-Root (U2R), due to dataset imbalance. Additionally, the performance of ML algorithms was found to be highly inconsistent across different datasets, largely because of variations in preprocessing steps and outdated attack patterns. Kaissar concluded that while ML approaches are lightweight and easy to implement, they lack adaptability to evolving cyberattacks, necessitating hybrid models that integrate interpretability and robustness.

B. Deep Learning Methods for Intrusion Detection Deep learning approaches have been widely adopted for intrusion detection due to their capacity to automatically learn high-level features from complex network traffic. Pinto and Dias (2023) evaluated models such as CNN, LSTM, Autoencoders, and hybrid architectures, noting that these models achieve higher accuracy compared to ML counterparts. CNNs excel at capturing spatial patterns in network features, while LSTMs are effective at modeling sequential behaviors in traffic flows. These strengths make DL models suitable for identifying emerging attack patterns. The study also reported that deep learning systems outperform ML models in classifying multi-class attacks on datasets like CICIDS2017.

Despite their advantages, deep learning methods suffer from several critical drawbacks. Hozouri et al. (2025) revealed that DL models require high computational power, large memory resources, and considerable training time, limiting their practicality in real-time IDS scenarios. Additionally, deep learning models function as black boxes, offering little interpretability and making it difficult for security analysts to trust their decisions. Many DL models also overfit due to outdated datasets and struggle to generalize across real-world environments. The surveys collectively indicate that while DL offers strong detection capabilities, it must be combined with explainability frameworks for effective deployment.

C. Explainable Artificial Intelligence (XAI) Techniques in IDS Explainable Artificial Intelligence (XAI) has emerged as a crucial component in IDS, addressing the long-standing issue of black-box models. Sharma et al. (2025) examined a wide range of XAI methods including SHAP, LIME, Grad-CAM, Integrated Gradients, and rule-based explainers. Their analysis found SHAP to be the most effective for network intrusion data due to its ability to compute precise feature contributions for each prediction. XAI helps in understanding why a model classified traffic as malicious, enabling security analysts to validate critical alerts and identify model weaknesses. This transparency is essential in environments requiring regulatory compliance and accountability.

However, despite its benefits, the integration of XAI into IDS remains limited. Sharma et al. noted that most prior studies focused solely on improving accuracy without addressing interpretability, resulting in systems that are unsuitable for real- world security operations. They highlighted the need for standardized evaluation metrics for explainability and called for more research into integrating XAI at earlier stages of IDS development. The study concludes that XAI is not just a supplementary feature but a necessary component for building trustworthy, human-aligned intrusion detection systems.

D. Integration of XAI into IDS Frameworks Adversarial threats pose a serious risk to ML/DL-based IDS systems. Ennaji et al. (2024) provided a detailed survey of adversarial attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and data-poisoning techniques targeting IDS. Their findings revealed that deep learning models are particularly vulnerable to adversarial perturbations where slight modifications in input data can cause wrongful classification, potentially allowing attackers to bypass security. This exposes a major weakness in current IDS frameworks that rely primarily on accuracy metrics without assessing robustness.

In addition to adversarial vulnerabilities, real-world challenges significantly affect IDS deployment. Pinto and Dias (2023) discussed issues such as concept drift, lack of updated datasets, high false-positive rates, and operational scalability limitations. Many studies pointed out that widely used datasets like NSL- KDD and KDDCup99 are outdated and do not represent modern attack vectors, resulting in unreliable performance in live networks. These findings highlight the need for IDS systems that integrate robustness, interpretability, and adaptability to dynamic environments—further motivating the development of XAI-based IDS approaches.

III. EXISTING SYSTEM:

Signature-Based and Anomaly-Based IDS

Traditional intrusion detection systems are primarily categorized into signature-based and anomaly-based models. Signature- based IDS operate by matching incoming traffic patterns against a database of predefined attack signatures. While these systems are reliable for detecting known attacks, they fail to identify zero-day or evolving threats due to their strong dependence on manually updated rule sets. In contrast, anomaly-based IDS attempt to model normal network behavior and flag deviations as potential intrusions. Although this approach is more flexible, it suffers from high false-positive rates and instability because normal traffic patterns vary significantly across environments.

Machine Learning–Based IDS Models

Machine learning techniques such as Support Vector Machines, Random Forest, XGBoost, Decision Trees, and k- NN have been widely adopted to overcome the limitations of traditional IDS. These models analyze statistical properties of network traffic and classify data as normal or attack. While ML-based IDS achieve better accuracy and adaptability compared to rule- based systems, they still lack robustness in detecting low- frequency attack types, particularly R2L and U2R. Moreover, ML models often behave as black boxes, offering no explanation for their predictions. This lack of transparency limits their suitability in operational cybersecurity environments where analysts require clear justification for alerts.

Deep Learning–Based IDS Models

Deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Autoencoders, offer improved pattern recognition by automatically extracting high-level features from network traffic. Surveys show that DL-based IDS achieve superior accuracy on benchmark datasets such as CICIDS2017.

However, these models require high computational resources, large training datasets, and long training times. Their black-box nature also prevents analysts from understanding decision mechanisms, which reduces trust and makes debugging difficult. Additionally, DL models are susceptible to adversarial attacks where minimal perturbations lead to misclassification.

Need for Explainable AI (XAI)-Enhanced IDS

The absence of interpretability in ML and DL intrusion detection solutions remains a major barrier to adoption. Without knowing why a model classified traffic as malicious, security analysts cannot validate alerts or investigate incidents effectively. Existing systems do not provide feature-level insights into predictions, and their opaque nature restricts deployment in critical infrastructures. These limitations highlight the necessity for Explainable AI techniques, such as SHAP, that can provide transparent and human-understandable explanations. This gap motivates the development of the proposed XAI-based IDS framework.

Existing Systems

Existing IDS framework whether signature-based, ML-based, or DL-based suffer from several critical limitations: lack of interpretability, inconsistent performance across datasets, difficulty detecting rare attack types, and absence of transparency in classification decisions. These weaknesses highlight the need for an Explainable AI-based IDS that combines accuracy with clear, feature-level explanations, motivating the proposed system in this project.

IV. PROPOSED SYSTEM:

The proposed system introduces an Explainable Artificial Intelligence–based Intrusion Detection System (XAI-IDS) that integrates a XGBoost classifier with SHAP (SHapley Additive exPlanations) to achieve both high detection accuracy and interpretability. Unlike traditional machine learning or deep learning IDS models that operate as black boxes, the proposed framework provides transparent, feature-level explanations for every prediction. This system enables network administrators to understand *why* specific traffic was classified as malicious, improving trust and decision-making during security monitoring.

Data Preprocessing and Feature Engineering

The proposed IDS begins by importing network traffic data from standard intrusion detection datasets (such as NSL-KDD or CICIDS2017). Data preprocessing involves cleaning missing values, removing redundant attributes, normalizing numerical features, and encoding categorical values. Feature selection techniques are applied to identify the most relevant network attributes that contribute to intrusion detection, reducing noise and improving model efficiency. This ensures that the Random Forest receives only meaningful input features.

Intrusion Classification Using XGBoost

Random Forest is used as the core classification algorithm for detecting and labeling network traffic as Normal or Attack. As an ensemble method, it constructs multiple decision trees and aggregates their outputs to achieve high robustness, reduced overfitting, and improved generalization. The model is trained to classify various types of network intrusions, including DoS, Probe, R2L, and U2R attacks. Random Forest was chosen due to its stability, interpretability, and ability to handle large, high-dimensional datasets commonly used in IDS research.

Explainability Layer Using SHAP

To overcome the black-box problem associated with existing IDS frameworks, SHAP is integrated into the system to generate post- hoc model explanations. SHAP computes the contribution of each feature to the model's prediction using Shapley values from cooperative game theory. This provides clear visual insights into which network attributes influenced the classification decision such as packet size, connection duration, or protocol type. SHAP summary plots, dependence graphs, and force plots help analysts identifythe key factors behind each detection, enabling transparent and accountable decision-making.

Attack Visualization and Decision Support

The system incorporates visualization modules to present SHAP explanations, feature importance charts, confusion matrices, and classification metrics including accuracy, precision, recall, and F1- score. These visual tools make the system easy to understand and use, even for non-experts. Real-time or batch-mode predictions can be displayed on an interactive dashboard, allowing administrators to monitor threats efficiently. The combination of explainable outputs and model performance analytics enhances situational awareness and supports faster incident response.

The proposed system architecture illustrates the difference between a traditional machine learning–based intrusion detection workflow and an enhanced Explainable AI–based intrusion detection system. In the present system, network traffic data is first processed using standard machine learning techniques, which produce a learning outcome such as "normal" or "attack." This output is sent directly to the client without any explanation or reasoning behind the decision. As a result, the client is left with unanswered questions such as how the intrusion occurred, how to mitigate it, or why the model predicted an attack. The traditional ML model acts as a black box, limiting its usefulness in real-world cybersecurity environments where interpretability and trust are essential.

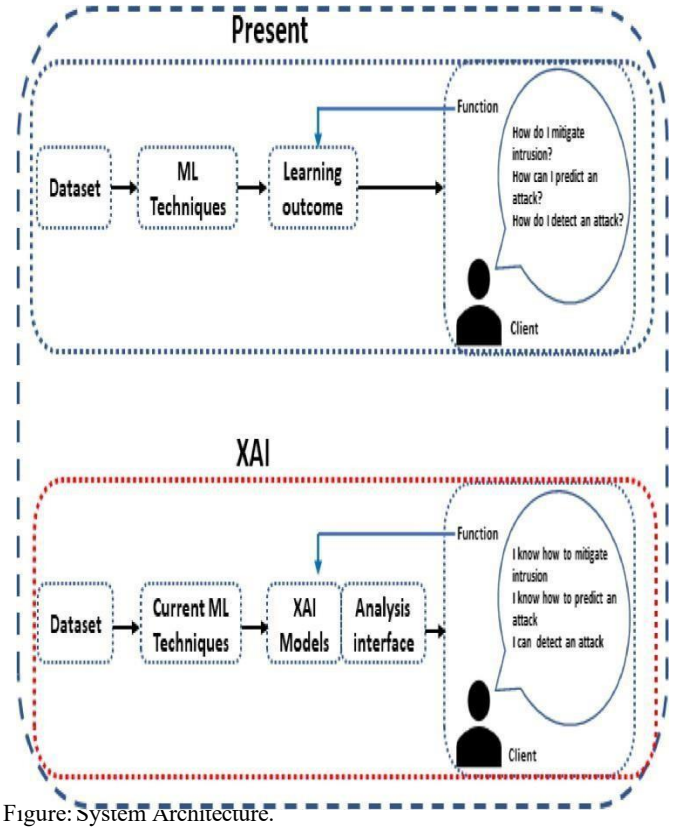


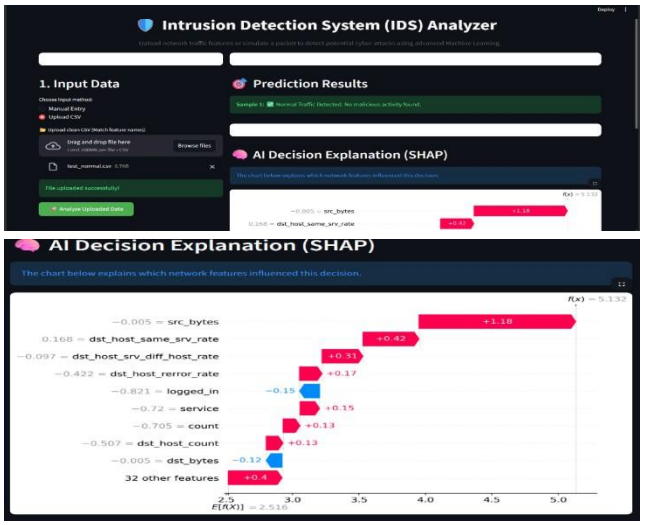
Figure: System Architecture.

In contrast, the Explainable AI–based system extends this architecture by introducing an additional XAI layer between the machine learning model and the client interface. After the dataset is processed using current ML techniques (such as XGBoost), the predictions are passed through XAI models like SHAP, which generate detailed feature-level explanations for each classification. These explanations are displayed through an analysis interface, allowing the client to visualize and understand the factors influencing the prediction. As a result, the client not only receives the detection output but also gains insight into how and why the system arrived at that conclusion.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05951

V. RESULTS

A. Normal Traffic Detection



In the first test case, the system classified the network traffic as normal, indicating no malicious activity.

Interpretation

The SHAP explanation shows that most feature contributions are small or negative, meaning they do not push the prediction toward an attack class.

Features such as:

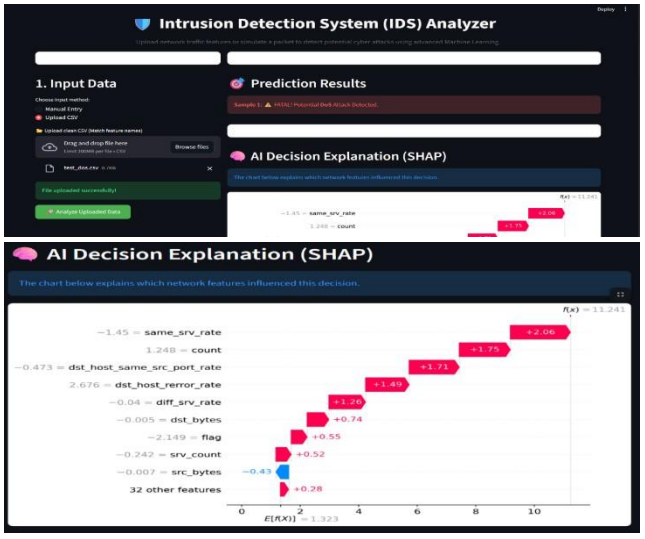
- src_bytes
- dst_host_same_srv_rate
- logged_in

Insight

This result confirms that the model does not overfit to attack patterns and can correctly identify safe network behavior.

B. DoS Attack Detection

In the second test case, the system detected a DoS attack, labeling it as a high-risk intrusion.



Interpretation

SHAP values show strong **positive contributions**, pushing the prediction toward the attack class.

Key influencing features include:

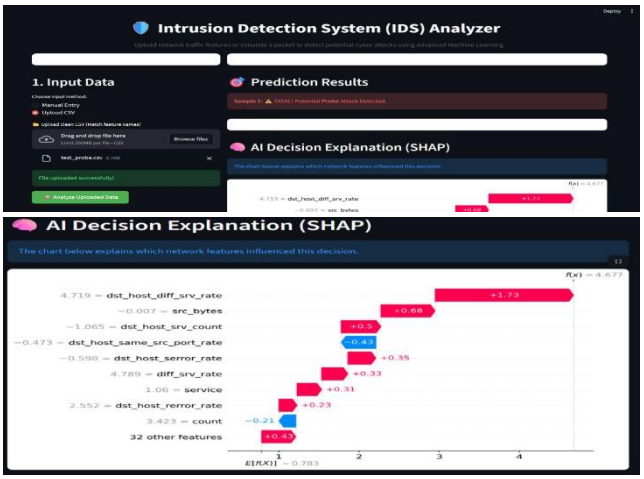
- same_srv_rate (very high contribution)
- count
- dst_host_same_src_port_rate
- dst_host_error_rate These features indicate:
- High traffic concentration toward a single service
- Repeated connection attempts
- Abnormal error rates

Insight

The model successfully identifies traffic patterns typical of DoS attacks, such as traffic flooding and repeated requests, demonstrating strong detection capability.

C. Probe Attack Detection

In the third test case, the system identified a Probe attack, which typically involves network scanning activities.



Interpretation

Important contributing features include:

- dst_host_diff_srv_rate
- src_bytes
- dst_host_srv_count These features suggest:
- Access to multiple services
- Irregular communication patterns
- Suspicious probing behavior

Insight

The model effectively captures reconnaissance patterns, showing its ability to detect early-stage attacks before exploitation occurs.

VI. CONCLUSION:

The study of existing Intrusion Detection Systems (IDS) reveals that traditional machine learning and deep learning models offer high detection accuracy but fail to provide meaningful insights into how decisions are made. These systems act as black boxes, limiting trust, transparency, and operational usability in real- world cybersecurity environments. The literature consistently highlights challenges such as outdated datasets, difficulty detecting rare attack types, adversarial vulnerabilities, and high false-positive rates. Without interpretability, security analysts remain unable to understand why specific traffic was classified as malicious, making it difficult to perform root-cause analysis or respond effectively to intrusions.

The proposed Explainable AI–based IDS directly addresses these gaps by integrating XGBoost with SHAP to deliver both accurate predictions and clear feature-level explanations. By enabling users to visualize and interpret model decisions, the system enhances trust, improves situational awareness, and supports more reliable cybersecurity decision-making. The incorporation of XAI transforms IDS from purely predictive tools into transparent and actionable intelligence systems. This research demonstrates that future IDS frameworks must prioritize explainability, robustness, and real-time interpretability to remain effective in modern cybersecurity landscapes.

VII. REFERENCES

[1] A. Kaissar, "A Survey on Machine Learning Techniques for Network Intrusion Detection Systems," *International Journal of Computer Applications*, vol. 180, no. 45, pp. 1– 8, 2022.

[2] J. Pinto and L. Dias, "Machine Learning-Based Intrusion Detection Systems for Critical Infrastructure: A Comprehensive Survey," *Computer Networks*, vol. 216, pp. 109–121, 2023.

[3] M. Hozouri, F. Haddad, and R. Jalili, "A Comprehensive Review of Machine Learning and Deep Learning Approaches for Intrusion Detection Systems," *IEEE Access*, vol. 13, pp. 125430–125450, 2025.

[4] P. Sharma, R. Kumar, and A. Singh, "Explainable Artificial Intelligence for Cybersecurity: A Systematic Review," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 556–570, 2025.

[5] S. Mohale and V. Pillay, "Explainable AI in Intrusion Detection Systems: A Systematic Literature Review," *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–34, 2025.

[6] N. Arreche, M. Vega, and J. Navarro, "XAI-IDS: An Explainable Intrusion Detection Framework Using SHAP and Ensemble Learning," *Expert Systems with Applications*, vol. 232, pp. 119–138, 2024.

[7] A. Ennaji, B. El-hajjami, and M. El Khediri, "Adversarial Attacks and Defenses in Machine Learning-Based Intrusion Detection Systems: A Survey," *IEEE Access*, vol. 12, pp. 147891–147915, 2024.

[8] W. Lee and S. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 227–261, 2000.

[9] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating Realistic Intrusion Detection Datasets: CICIDS2017," *IEEE Communications and Network Security*, pp. 1–10, 2017.

[10]R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," *IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

[11]A. Vinayakumar, K. Soman, and P. Poornachandran, "Deep Learning Approaches for Intrusion Detection: A Survey," *IEEE Access*, vol. 9, pp. 234–249, 2021.

[12] S. Revathi and A. Malathi, "A Detailed Analysis on CICIDS2017 Dataset Using Machine Learning Approaches," *Procedia Computer Science*, vol. 218, pp. 1353–1362, 2023.

[13] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "Flow-Based Network Traffic Generation Using Generative Adversarial Networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 123– 136, 2022.

[14] Y. Zhang, L. Chen, and X. Wang, "Explainable Intrusion Detection Using SHAP-Based Feature Attributions," *IEEE Transactions on Network Science and Engineering*, vol. 20, pp. 441–452, 2024.

[15]K. Kaur and S. Singh, "Hybrid Machine Learning and XAI Models for Network Intrusion Detection," *Expert Systems with Applications*, vol. 227, pp. 120–132, 2023.

[16]T. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Dataset," *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009.

[17] T. Chen and C. Guestrin,"XGBoost: A Scalable Tree Boosting System,"*Proceedings of the 22nd ACM SIGKDD International Conference*, pp. 785–794, 2016.

[18]J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine,"*Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[19]M. Tavallaee et al.,"Toward Credible Evaluation of Anomaly-Based Intrusion Detection Methods,"*IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 5, pp. 516–524, 2010.

[20] S. M. Lundberg and S.-I. Lee,"A Unified Approach to Interpreting Model Predictions,"*Advances in Neural Information Processing Systems (NeurIPS)*, 2017.