



Explainable Artificial Intelligence In Intrusion Detection Systems

Nirgude Divya S.¹ & Prof. Boraste N. N.²

^{1,2}Department of Computer Science MVP Samaj's K. K. Wagh Art's, Science & Commerce College, Pimpalgaon (Baswant), Nashik, Maharashtra, India.

Corresponding Author – Nirgude Divya S.

DOI - 10.5281/zenodo.18874029

Abstract:

With the rapid expansion of computer networks, cloud infrastructures, and Internet of Things (IoT) environments, the number and complexity of cyber-attacks have increased significantly. Intrusion Detection Systems (IDS) play a crucial role in identifying malicious activities and protecting network resources. Traditional IDS techniques, including signature-based and anomaly-based systems, face challenges such as high false-positive rates, poor adaptability to new attacks, and limited interpretability.

Recently, Artificial Intelligence (AI) and Machine Learning (ML) techniques have been widely adopted in IDS due to their high detection accuracy and ability to analyze large volumes of network data. However, most AI-based IDS models operate as “black boxes,” making it difficult for security analysts to understand how and why a particular decision is made. This lack of transparency reduces trust, limits practical deployment, and creates challenges in regulatory compliance.

Explainable Artificial Intelligence (XAI) addresses these issues by providing human-interpretable explanations for AI decisions. This research focuses on the integration of XAI techniques into IDS to enhance transparency, trust, and decision-making capability while maintaining strong detection performance. The proposed approach combines machine learning-based intrusion detection with explainability methods such as feature importance and rule-based explanations. Experimental evaluation demonstrates that XAI-enabled IDS improves security analysis, accountability, and auditability without significantly compromising performance.

Keywords: Intrusion Detection System, Explainable AI, Cyber Security, Machine Learning, Network Security, Transparency.

Introduction:

Cybersecurity has become one of the most critical concerns in modern digital environments due to the rapid growth of networked systems, cloud computing, and IoT devices. Organizations rely heavily on computer networks for data storage, communication, and business operations, making them attractive targets for cyber-attacks such as malware, denial-of-service (DoS), phishing, and unauthorized access.

Intrusion Detection Systems (IDS) are designed to monitor network traffic and system activities to detect malicious behavior. Traditional

IDS approaches are mainly categorized into signature-based and anomaly-based systems. Signature-based IDS can only detect known attacks, while anomaly-based IDS often generate a high number of false alarms.

To overcome these limitations, machine learning and deep learning techniques have been widely adopted in IDS. AI-based IDS can automatically learn attack patterns from data and detect unknown threats with high accuracy. However, these systems suffer from a major drawback—the lack of interpretability. Most AI models, especially deep learning models, function

as black boxes, providing predictions without explanations.

Explainable Artificial Intelligence (XAI) aims to make AI models transparent by explaining their decisions in a human-understandable manner. In the context of IDS, XAI helps security analysts understand why a network activity is classified as normal or malicious. This enhances trust, improves incident response, and supports compliance with security regulations.

This research investigates the role of Explainable AI in Intrusion Detection Systems. It proposes an XAI-enabled IDS framework that combines detection accuracy with transparency, enabling better security analysis and decision-making.

Literature Review:

Traditional IDS techniques rely on predefined rules or statistical thresholds, which are limited in detecting sophisticated and zero-day attacks. Machine learning-based IDS models such as Decision Trees, Support Vector Machines (SVM), Random Forest, and Neural Networks have shown improved detection accuracy.

Recent studies highlight that deep learning models such as CNNs, RNNs, and LSTM networks achieve high performance in IDS but lack interpretability. This creates challenges in understanding attack behavior and reduces user trust.

Explainable AI techniques such as:

- Feature importance analysis
- Rule-based explanations
- Model-agnostic methods like LIME and SHAP
- have been proposed to improve transparency in AI systems. Researchers have shown that XAI-based IDS allows analysts to identify key features responsible for intrusion detection decisions, reducing false positives and

improving response strategies.

However, existing research indicates a trade-off between explainability and performance. Therefore, designing an IDS that balances accuracy, interpretability, and scalability remains an open research challenge.

Research Methodology:

This research follows a systematic and structured methodology to study the role of Explainable Artificial Intelligence (XAI) in Intrusion Detection Systems (IDS). The methodology focuses on understanding existing intrusion detection techniques, designing an AI-based IDS model, integrating explainability mechanisms, and evaluating the system using standard datasets and performance metrics.

The objective of this methodology is not only to achieve high intrusion detection accuracy but also to ensure transparency, interpretability, and trust in the decision-making process of AI models. The research methodology combines descriptive analysis, experimental design, and comparative evaluation to provide a comprehensive understanding of XAI-enabled IDS.

Research Design:

1. Research Approach

- The research adopts a combined descriptive and experimental research approach.
- The descriptive approach is used to study existing literature related to:
 - Traditional Intrusion Detection Systems
 - Machine Learning and Deep Learning-based IDS
 - Explainable Artificial Intelligence techniques
- This approach helps in identifying limitations of existing IDS models, particularly the lack of interpretability and transparency.

The experimental approach is used to:

- [illegible]

- Secure handling of network datasets
- Restricted access to IDS dashboards
- Protection against data poisoning and

adversarial attacks

- Integrity checks to ensure datasets are not manipulated

The design ensures that the IDS itself does not become a vulnerability.

6.Evaluation Design:

The evaluation of the proposed XAI-IDS system is conducted using multiple performance and security parameters:

- Detection accuracy and false positive rate
- Explainability effectiveness
- System response time
- Scalability under increasing traffic loads

The evaluation results are compared with traditional IDS systems to highlight improvements.

7.Tools and Technologies:

The following tools and technologies are used:

- Programming Language: Python
- ML Libraries: Scikit-learn, TensorFlow / PyTorch
- Explainability Tools: SHAP, LIME
- Datasets: KDD Cup 99, UNSW-NB15
- Network Tools: Wireshark

Data Sources and Selection Criteria:

Data Sources:

1. Primary Data Sources:

Primary data is collected directly through experimentation:

- Network traffic datasets used for training and testing IDS models
- IDS classification outputs
- Explainability reports generated by XAI techniques
- Performance metrics such as accuracy, precision, recall, and latency

This data is essential for evaluating the effectiveness of the proposed system.

2.Secondary Data Sources:

Secondary data is collected from existing sources:

- IEEE research papers and journals
- Conference proceedings on IDS and XAI
- Technical documentation of IDS datasets
- Cybersecurity reports and standards

Secondary data supports theoretical understanding and comparison.

3. Data Selection Criteria:

Relevance:

- Data must be related to intrusion detection and network security
- Features should reflect real attack behavior

4.Reliability and Credibility:

- Standard benchmark datasets are preferred
- Peer-reviewed sources are used

5.Timeliness

- Recent datasets representing modern attacks are selected

6.Accuracy and Consistency

- Data must be correctly labeled
- Inconsistent records are removed

7.Ethical Considerations

- No personal or sensitive data is used
- Data usage complies with ethical guidelines

Purpose of Data Selection:

The purpose of selecting appropriate data is to:

- Ensure valid experimental results
- Accurately evaluate IDS performance
- Compare traditional IDS with XAI-based IDS

Experimental Results and Discussion:

1. Experimental Setup:

The proposed XAI-IDS system was implemented in a controlled environment using benchmark datasets. Machine learning models were trained on labeled data and tested on unseen

traffic samples. Explainability techniques were applied to interpret IDS decisions.

Experimental Results:

1.Security Results: The system demonstrated strong resistance to intrusion attempts. Unauthorized access patterns were accurately detected, and XAI explanations clearly highlighted features responsible for detection.

Result: The system demonstrated strong resistance to intrusions and ensured reliable and secure network monitoring.

2.Performance Results: The IDS achieved high accuracy with acceptable processing time. Explainability introduced minimal computational overhead.

Result: The IDS maintained efficient performance while providing transparent and interpretable security decisions.

3.Scalability Results: As network traffic increased, the system maintained stable detection performance up to a certain threshold.

Result: The XAI-based IDS proved scalable and suitable for small to medium-scale network environments.

4.Access Control Results: Access to IDS explanations was restricted to authorized analysts, improving system security.

Result: The integration of Explainable AI significantly improved trust, transparency, and usability of the IDS.

Discussion:The results indicate that XAI enhances IDS effectiveness by making AI decisions interpretable. Security analysts can understand attack behavior, reducing false alarms and improving response strategies. However, scalability and computational overhead remain challenges.

Summary of Findings:

- Improved intrusion detection accuracy
- Enhanced transparency and interpretability

- Better trust and auditability
- Acceptable performance overhead

Results:

The experimental evaluation of the proposed Explainable Artificial Intelligence based Intrusion Detection System (XAI-IDS) demonstrates significant improvements in intrusion detection accuracy, transparency, and security management when compared to traditional IDS approaches. The results are analyzed across multiple dimensions, including security effectiveness, performance efficiency, scalability, access control, and explainability.

1. Improved Intrusion Detection Accuracy: The XAI-based IDS achieved higher detection accuracy for both known and unknown attack patterns. Machine learning models successfully identified anomalies in network traffic, while explainable mechanisms provided clear reasoning behind each detection decision. This reduced the occurrence of false positives, which is a common issue in conventional anomaly-based IDS.

2. Enhanced Data Security: The system effectively protected network data by detecting unauthorized access attempts and malicious activities in real time. Explainable AI helped security analysts understand which features contributed to intrusion detection, enabling faster and more informed response actions.

3. Data Integrity Assurance: Integrity of the monitoring data and IDS outputs was maintained throughout the experimental process. The system ensured that classification results were consistent and verifiable. Any abnormal modification in traffic patterns was immediately identified, preventing silent attacks and data manipulation.

4. Improved Transparency and Auditability: One of the most important results of the proposed system is enhanced transparency.

Unlike black-box IDS models, XAI-IDS provides interpretable explanations for each alert. These explanations serve as an audit trail that can be reviewed by administrators for compliance, forensic investigation, and reporting purposes.

5. Efficient Performance: Although explainability adds an additional processing layer, experimental results show that the overall system performance remains efficient. The latency introduced by XAI techniques was minimal and acceptable for intrusion detection tasks, making the system suitable for practical deployment.

6. Scalability Performance: The system maintained stable performance under moderate network traffic loads. As the number of data instances increased, detection accuracy remained consistent. However, slight performance degradation was observed at very high traffic volumes due to computational overhead.

7. Secure Access Control: Access to intrusion alerts and explanation dashboards was restricted to authorized users only. This ensured that sensitive security information was not exposed to unauthorized individuals, further strengthening the system's security posture.

Overall Result:

The overall results confirm that integrating Explainable Artificial Intelligence into Intrusion Detection Systems significantly enhances the effectiveness, usability, and trustworthiness of network security solutions. The proposed XAI-IDS successfully combines strong detection capabilities with interpretability, enabling security analysts to understand, verify, and justify IDS decisions.

The system provides a balanced trade-off between accuracy and transparency, making it suitable for real-world cybersecurity

environments where accountability and compliance are essential. Compared to traditional IDS models, the XAI-based approach offers better visibility into attack behavior, improved incident response, and greater confidence in automated security decisions.

Overall, the study proves that Explainable AI is not only a supportive feature but a critical requirement for next-generation intelligent intrusion detection systems.

Discussion:

The findings of this research highlight the importance of explainability in modern intrusion detection systems. Traditional AI-based IDS models, although accurate, fail to provide insight into their decision-making processes. This lack of transparency limits trust and adoption in critical environments such as financial systems, healthcare networks, and government infrastructures.

The integration of Explainable AI addresses this challenge by enabling human-understandable interpretations of intrusion detection decisions. Security analysts can identify why a particular activity is classified as malicious, which features influenced the decision, and how confident the model is. This improves situational awareness and supports better decision-making.

Experimental results confirm that XAI does not significantly compromise detection performance. Instead, it adds value by reducing false alarms and improving response efficiency. However, the discussion also reveals challenges related to scalability, computational overhead, and dataset limitations.

Overall, the discussion confirms that Explainable AI enhances the practicality and reliability of IDS, making them more aligned with real-world cybersecurity requirements where

transparency, accountability, and trust are essential.

Conclusion:

This research concludes that Explainable Artificial Intelligence plays a crucial role in enhancing Intrusion Detection Systems by combining intelligent detection with transparency and interpretability. The proposed XAI-based IDS successfully detects malicious activities while providing meaningful explanations for its decisions.

The system improves data security, ensures integrity, enhances transparency, and supports auditability, which are critical factors for modern cybersecurity solutions. Although challenges related to scalability and performance exist, the benefits of explainability outweigh these limitations.

In conclusion, Explainable AI-enabled IDS represents a significant advancement over traditional black-box security models. It offers a reliable, trustworthy, and effective solution for protecting modern network environments and sets a strong foundation for future research and development in intelligent cybersecurity systems.

Future Work:

Despite the promising results achieved by the proposed Explainable Artificial Intelligence based Intrusion Detection System (XAI-IDS), several areas remain open for further research and improvement. Future work can focus on enhancing scalability, real-time performance, robustness, and applicability of XAI-enabled IDS in complex and dynamic environments.

1. Scalability Enhancement for Large-Scale Networks: Future research can focus on improving the scalability of XAI-based IDS to support large-scale and high-speed networks. As network size and traffic volume increase, both intrusion detection and explanation

generation require additional computational resources. Advanced optimization techniques such as distributed learning, parallel processing, and lightweight explainability models can be explored to reduce processing overhead. Scalable architectures will allow XAI-IDS to be deployed effectively in enterprise networks, cloud infrastructures, and Internet Service Provider (ISP) environments.

2. Real-Time Intrusion Detection with Explainability:

Although the current system demonstrates acceptable performance, real-time deployment remains a challenge due to the additional computation required for generating explanations. Future work can focus on designing real-time XAI frameworks that generate explanations instantly without affecting detection speed. Techniques such as approximate explanations, streaming-based analysis, and incremental learning models can be explored to achieve real-time intrusion detection while maintaining interpretability.

3. Integration with Deep Learning Models:

Future research can explore the integration of Explainable AI with advanced deep learning architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. Deep learning models are highly effective in detecting complex attack patterns, but they lack transparency. Applying advanced explainability techniques to deep learning-based IDS can improve both detection accuracy and interpretability, making them suitable for complex cyber-attack scenarios.

4. Adaptive and Self-Learning IDS Models:

Cyber-attacks continuously evolve, and static IDS models may become ineffective over time. Future work can focus on adaptive and self-learning XAI-IDS models that automatically update detection rules and explanations based on new attack patterns. Online learning and

reinforcement learning techniques can be integrated to allow the IDS to adapt dynamically to changing network behavior while maintaining clear and consistent explanations.

5. Explainability for Security Analysts and Non-Technical Users: While XAI provides explanations, further research can improve how these explanations are presented to users. Future systems can focus on developing user-friendly visualization dashboards that present explanations in simple and intuitive formats. Graphical representations, heatmaps, and rule-based summaries can help both technical and non-technical users understand intrusion alerts more effectively, improving incident response and decision-making.

6. Integration with IoT and Edge Computing Environments: With the rapid growth of IoT devices and edge computing, future research can extend XAI-IDS to resource-constrained environments. IoT networks are highly vulnerable to attacks but have limited processing capabilities. Lightweight machine learning models combined with efficient explainability techniques can be developed to provide secure and interpretable intrusion detection at the network edge.

7. Privacy-Preserving Explainable IDS: Future work can focus on integrating privacy-preserving mechanisms into XAI-IDS. While explanations improve transparency, they may also expose sensitive information if not carefully designed. Techniques such as differential privacy, federated learning, and secure multi-party computation can be explored to ensure that explanations do not leak confidential network or user information.

8. Cross-Dataset and Cross-Domain Evaluation: Most IDS models are trained and tested on specific datasets. Future research can focus on evaluating XAI-IDS across multiple

datasets and network environments to improve generalization. Cross-domain evaluation will help assess how well the system performs in different organizational contexts such as healthcare, finance, and smart cities.

9. Automated Policy and Compliance Support: Future XAI-IDS systems can incorporate automated compliance and policy enforcement mechanisms. Explainable alerts can be linked with organizational security policies and regulatory requirements. This will help organizations demonstrate compliance with cybersecurity standards and regulations by providing clear justifications for security decisions.

10. Hybrid Security Frameworks: Future research can explore hybrid frameworks that integrate XAI-based IDS with other security solutions such as firewalls, intrusion prevention systems (IPS), and Security Information and Event Management (SIEM) systems. Such integration can create a comprehensive and intelligent security ecosystem that combines detection, prevention, explanation, and response capabilities.

Limitations:

Despite the advantages of the proposed XAI-based Intrusion Detection System, several limitations were identified during the research and experimental evaluation:

1. Scalability Issues: As network size and traffic volume increase, the computational complexity of both intrusion detection and explainability mechanisms also increases. This can impact real-time detection in large-scale networks.

2. Performance Overhead: Explainable AI techniques require additional computation to generate interpretations. Although the overhead is moderate, it may affect system responsiveness in high-speed or real-time

environments.

- 3. Dataset Dependency:** The performance of the IDS heavily depends on the quality and diversity of the training dataset. Limited or outdated datasets may reduce detection accuracy and explanation reliability.
- 4. Complexity of Model Interpretation:** While XAI improves transparency, some explanations may still be difficult for non-technical users to fully understand. Proper training of security personnel is required.
- 5. Limited Generalization:** Models trained on specific datasets may not generalize well to all network environments or emerging attack types without retraining or adaptation.
- 6. Resource Consumption:** The system requires sufficient computational resources, especially when deploying complex machine learning models with explainability modules.
- 7. Real-Time Deployment Challenges:** Applying XAI-IDS in real-time production environments requires optimization to handle continuous high-volume data streams efficiently.

References:

1. D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
2. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.
3. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of the ACM SIGKDD Conference*, pp. 1135–1144, 2016.
4. 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.
5. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," *Military Communications and Information Systems Conference*, 2015.
6. W. Wang, Y. Sheng, J. Wang et al., "HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection," *IEEE Access*, vol. 6, pp. 1792–1806, 2018.
7. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The Great Time Series Classification Bake Off," *Data Mining and Knowledge Discovery*, vol. 31, pp. 606–660, 2017.
8. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
9. H. Hindy, D. Brosset, E. Bayne et al., "A Taxonomy of Network Threats and the Effect of Explainable AI on Intrusion Detection Systems," *IEEE Access*, vol. 8, pp. 104650–104675, 2020.
10. NIST, "Guide to Intrusion Detection and Prevention Systems (IDPS)," *NIST Special Publication 800-94*, National Institute of Standards and Technology, 2012.