International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

A Hybrid Explainable Deep Learning Framework for Intrusion Detection System

*Jay Khobare
M.Tech. (Student)
Department of Computer Science & Engineering
Prestige institute of Engineering Management & Research Indore, 452010, M.P., India
Orcid id: 0009-0004-8949-9123

Dipti Chauhan
Professor
Department of Artificial Intelligence & Data Science
Prestige Institute of Engineering Management & Research Indore, 452010, M.P., India
Orcid id:0000-0003-1665-7587

Abstract- The increasing strength of cyberattacks on digital systems has led to the requirement for intelligent and explainable intrusion detection systems (IDS). The existing intrusion detection system models, although successful in detecting cyberattacks, are "black-box" systems that do not provide information about the decision-making process. This literature review aims to explore the use of explainable artificial intelligence (XAI) to improve the interpretability, trustworthiness, and acceptability of deep learning-based intrusion detection system models. This paper introduces a concise summary of the findings of sixteen recent research studies on the application of hybrid deep learning models with explainable artificial intelligence approaches such as Shapley Additive Explanation (SHAP) and Locally Interpretable Model-Agnostic Explanation (LIME) techniques. The literature review talks about the difficulties, existing methods, and also presents a hybrid explainable artificial intelligence method that combines convolutional neural networks (CNN), gated recurrent units (GRU), and transformer models with two-level interpretability. The results emphasize the importance of explainable artificial intelligence in solving the accuracy-transparency trade-off in intrusion detection system applications.

Keywords: Explainable Artificial Intelligence (XAI), Intrusion Detection System (IDS), SHAP, LIME, Interpretability, Cybersecurity.

I. INTRODUCTION

AI is increasingly becoming a part of everyday human life, but its output often acts like a “black box,” with no reason for its results, which reduces human trust in AI. Because intrusion detection systems (IDS) are crucial tools for detecting intrusions on networks, AI integration improves IDS output, but trust issues arise because it still operates like a “black box [1].” Integrating explainable artificial intelligence (XAI) into IDSs can give humans greater confidence in the output, as it provides an explainable output.

An IDS is a system that detects unauthorized access by hackers and malicious actors. IDS can be described as tools, methods, and resources that help identify, address, and report unauthorized or unapproved network activity. Once a behavioral anomaly is detected, a security administrator is alerted.

Automated systems are configured to meet individual needs and requirements. XAI is a set of processes that allows users to understand and trust AI-generated results/outputs. It helps ensure model accuracy, fairness, transparency, and outcome clarity in AI-powered decision-making. XAI is a key requirement for implementing Responsible AI, a method for large-scale implementation in real organizations that combines fairness, model explainability, and accountability [3]. XAI can improve the user experience of a product or service by helping end users trust that the AI is making good decisions. XAI applies specific techniques and methods to ensure that each decision made during the ML process is properly explainable.

Deep learning (DL) models, such as CNNs, GRUs, and Transformers, have demonstrated superior performance in identifying complex attack patterns. However, their "black-box" nature has led to reduced analyst confidence and non-compliance with regulatory requirements.

XAI models provide tools for decision interpretation, allowing analysts to interpret and trust IDS predictions. Methods such as Shapley Additive Explanation (SHAP), Locally Interpretable Model-Agnostic Explanation (LIME), and Layer-wise Relevance Propagation (LRP) explain predictions by linking input features to output classifications. When combined with deep learning (DL) IDS, XAI improves interpretability and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

supports human-in-the-loop cybersecurity decision-making.

Deep learning-based IDS have shown remarkable success in achieving high detection accuracy, especially in identifying complex and previously unseen cyberattack patterns [3]. However, despite their high predictive accuracy, these models tend to work as black-box systems, offering very little or no insight into the reasoning behind their predictions. Lack of transparency is a major issue in the cybersecurity field, where it is as important to understand the reasons for the generation of the alert as it is to understand the threat. If security analysts are unable to understand the predictions made by the model, it will lead to a decrease in their confidence in the automated systems, and this will cause a delay in the response actions, which will ultimately affect the efficiency of the IDS solutions. Moreover, in a critical area like cybersecurity, relying on the predictions made by AI systems without understanding them can cause operational risks.

In recent years, the need for explainable AI has been driven by ethical issues, the General Data Protection Regulation (GDPR), and the new AI governance policies [4]. The GDPR has highlighted the right to explanation, and as a result, decisions made by AI, particularly those related to security and privacy, need to be interpretable and auditable. As a result, if the IDS solutions are not able to provide explanations for their decisions, they may not be used in practice, even if they are accurate.

There are three key reasons for using a hybrid explainable deep learning framework. Firstly, the black-box nature of the limitations of the traditional deep learning-based IDS models makes it impossible for them to validate the correctness of their predictions. Secondly, trust is a fundamental aspect in the field of cybersecurity, and this is because security analysts need to be able to interpret the results in order to make informed decisions. Lastly, there is a need to develop responsible AI systems that are guided by ethical standards.

Thus, the need for explainable IDS frameworks arises to fill the gap between high detection accuracy and usability by combining XAI methods with deep learning models. These hybrid models attempt to achieve a balance between accuracy, interpretability, and computational complexity, thereby developing IDS systems that are not only accurate in detecting intrusions but also trustworthy, interpretable, and applicable in real-world cybersecurity settings. Although AI-driven IDS have shown success in various experiments in detecting complex patterns of attacks, most of these systems have inherent flaws that greatly impede their applicability in real-world cybersecurity settings. Firstly, most deep learning models for IDS systems are black boxes, meaning that their decision-making mechanisms are not yet understood. This lack of understanding makes it difficult for security analysts to identify why a particular network traffic pattern is labelled as anomalous, thereby making validation and subsequent actions towards the identified anomalies challenging. As a result, analysts' trust and confidence in automated systems for detection are greatly diminished, particularly in critical applications where incorrect decisions can cause serious operational and financial repercussions. Also, it is challenging that the new regulatory frameworks and guidelines to ethical AI lack a clear and comprehensible rationale behind decision-making, and thus, organizations struggle to adhere to them and remain transparent, accountable, and explainable in automated decision-making systems. The second important weakness is that the black-box IDS models are more vulnerable to adversarial attacks, as they are less interpretable, therefore, it is challenging to consider adversarial activities and provide defences against adversarial attack cases.

To overcome the aforementioned challenges, a hybrid intrusion detection system must be designed that combines explainability as a basic component of learning models. Instead of considering explainability as a complement to learning models, this approach will help security systems provide transparent, interpretable, and trustworthy results for intrusion detection.

The primary aim of this review is to systematically review and combine the recent research contributions on hybrid explainable deep learning-based IDS, with a special emphasis on enhancing the transparency, trust, and usability of AI-powered security solutions. The aims of this review are:

- Analyse hybrid deep learning architectures for IDS, particularly models that integrate convolutional neural networks (CNNs), gated recurrent units (GRUs), and transformer-based architectures, to understand how spatial, temporal, and contextual features of network traffic are leveraged together for effective intrusion detection.
- Examine the use of XAI methods, including Shapley Additive Explanation (SHAP) and Locally Interpretable Model-Agnostic Explanation (LIME), to improve

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

the global and local interpretability of deep learning-based IDSs.

- Discuss the available approaches, which seek to find a compromise between detection performance, model stability, and explainability, and explain why explainable IDS systems help address the dilemma of finding a high predictive accuracy and explainability.
- Consider the approaches and measures of explainability evaluation of studies and highlight the need of standardized benchmarks on objective comparison of explainable IDS models on explainability metrics like explanation faithfulness, consistency, and stability.
- Research visualisation and presentation of the concept of IDS explanations, with the emphasis on the user-friendly dashboards and real-time interpretation applications to improve the human-in-the-loop decision-making and the performance or efficiency of IDS.

II. IDS & XAI TECHNIQUES

IDS are critical elements of contemporary cybersecurity infrastructure, intended to monitor network and system activity for the purpose of detecting unauthorized access, malicious activity, and policy violations. An IDS is a monitoring system that continuously evaluates traffic patterns, system logs, and user activity for the purpose of detecting potential intrusions resulting from hacker attacks, insider threats, or malware. Upon the detection of suspicious activity or anomalies, the IDS system produces alerts to inform security administrators, which in turn enables response actions to be taken. IDS systems are typically automated and can be configured to enable organizations to set detection strategies based on their security needs and system complexity.

IDS systems can be categorized according to their architecture of deployment into Network-Based IDS (NIDS), Host-Based IDS (HIDS), and Cloud-Based IDS [5]. Network-Based IDS is a type of IDS that observes traffic on different segments of a network, often installed at points of strategic importance like routers, switches, and firewalls. This type of IDS analyses packet headers and payloads to identify potential threats like Distributed Denial of Service attacks, port scanning, and malware propagation. Even though it is able to provide network coverage, NIDS may experience some performance-related problems in high-speed networks or large-scale networks. Host-Based IDS on the other hand is a form of IDS, which operates on hosts or servers and monitors system-level activities like file management, system calls, logins, and system configurations. HIDS is very effective in insider attack or unauthorized privilege escalation detection as it offers system-level visibility. But as cloud computing gains more and more popularity, Cloud-Based IDS solutions have been created to track traffic within a virtualized and distributed network.

Under methodology, the IDS techniques may be classified broadly as signature-based detection techniques, anomaly-based detection techniques, and hybrid detection techniques. The signature-based detection is based on the knowledge of predefined patterns of known attacks which are then compared with known signature database. This specific approach is quite successful in detecting known threats and the false positive is low but it cannot detect unknown or zero-day attacks. Anomaly based detection on the other hand entails development of a normal system activity baseline and then anomalies in this activity are detected as potential attacks. This approach is quite efficient in detecting unknown threats, particularly when it is integrated with the Machine Learning and Deep Learning algorithms, which can detect sophisticated patterns of behaviour. Nonetheless, it can be more false positives in dynamic environment. Hybrid detection refers to a blend of signature-based and anomaly-based detection strategies to expand detection ability and minimize wrong positives and the application of artificial intelligence in such hybrid systems grows.

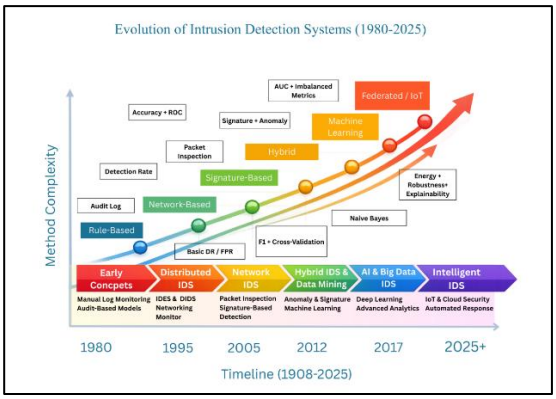

Figure 1: Evolution of Intrusion Detection Systems (1980-2025)

Figure 1 shows how IDS have progressively developed over the years between the primitive rule-based and audit logs monitoring systems to the current AI-based intelligent IDS. Detection methods have evolved over time to be based on a signature technique, a network technique, a hybrid technique, or a machine learning technique in an attempt to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

manage the complex cyber threats. The growing complexity of the approaches is associated with the adoption of advanced analytics, deep learning, and IoT-oriented security systems into the contemporary intrusion detection framework.

Although Deep Learning models, such as Convolutional Neural Networks, Gated Recurrent Units, and Transformers, are able to increase the accuracy of the detected patterns, they also introduce the black-box problem, where the logic behind the decisions is not transparent. Such opaqueness causes analysts to have a lack of trust in the results, the results to be regulation-compliant, and the results to be resistant to adversarial attacks. XAI addresses these issues by providing tools that may give explanations and interpretations of the model. In the context of IDS, XAI methods include model-agnostic methods like SHAP and LIME, which offer explanations for predictions by measuring the contribution of features to predictions without needing to know the internal workings of the model. Partial Dependence Plots and Individual Conditional Expectation are other methods that demonstrate how variations in features affect predictions. Model-agnostic methods like attention and Layer-wise Relevance Propagation offer explanations for predictions within the Deep Learning model itself. Rule extraction methods, on the other hand, convert complex predictions into human-readable form, while counterfactual explanations identify the least amount of change required in features to reverse the classification prediction.

Recent work has attempted to integrate XAI with IDS by employing model-specific, model-agnostic, or hybrid strategies. By leveraging state-of-the-art detection capabilities with explanation techniques, the XAI-integrated IDS has the potential to turn conventional black-box security models into trustworthy, accountable, and human-centric cybersecurity systems that can facilitate decision-making and regulatory compliance.

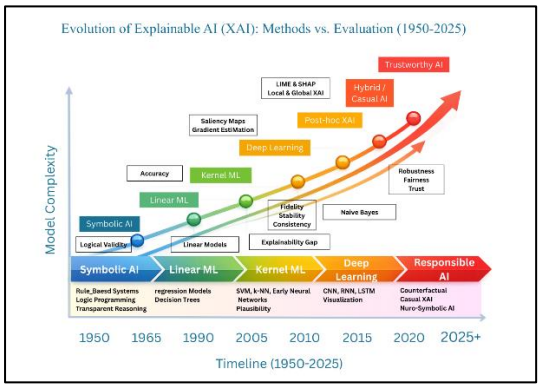

Figure 2: Evolution of Explainable AI (XAI): Methods vs. Evaluation (1950-2025)

The figure 2 shows how XAI has evolved over time since symbolic and rule-based systems to today's trustworthy AI systems. The early AI models were explainable by nature, whereas the advent of deep learning introduced the gap of explainability because of the black-box behaviour. Recent developments include SHAP, LIME, and causal AI, which are focused on making AI use transparent, fair, and robust to deliver reliable AI systems to critical areas of application like cybersecurity.

III. LITERATURE REVIEW

XAI has garnered a lot of attention in the field of intrusion detection system (IDS) research since 2022, as made evident by the papers reviewed in this paper. The initial research efforts were directed at understanding and identifying the challenges of interpretability in deep learning-based IDS models. The initial research on XAI-IDS was directed at conceptual frameworks, taxonomies, and human-in-the-loop strategies to improve the interpretability of the results of intrusion detection. The research was directed at the need to provide understandable explanations to security analysts to enable them to make decisions regarding response actions. The research was also directed at the need to enable security analysts to justify alerts and make informed decisions.

As the field of research regarding the subject grew, newer research started to focus on the practical aspect of XAI approaches in integration with IDS systems. Such studies explored post-hoc explanation approaches including Shapley Additive Explanations (SHAP), Locally Interpretable Model-Agnostic Explanations (LIME) and Layer-wise relevance propagation (LRP) to understand the inference of deep learning models without modifying the architecture. As can be seen based on the literature review, these methods have been effective in offering feature-level attribution, which allows the analyst to determine which network attributes best affect the detection results. Furthermore, the explainable outputs increase the acceptance, usability, and trust of AI-based IDS systems have been established by other studies.

Then more modern works build on model explainability to hybrid IDS design, which use more than one deep learning model to include CNNs, GRUs, and transformers with XAI methods. The motivation of such hybrid models is to take advantage of their capacity to recognize a specific pattern in network traffic besides providing local as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

well as global explanations of their behaviour. Literature seems to point to the fact that hybrid explainable IDS systems are a more balanced approach to the problem of IDS than either deep learning models or explainability systems.

Nonetheless, in spite of these advances, there are also some challenges that can be observed in the literature review. These are the increased computational complexity of explanation algorithms, the lack of ability to scale to support real-time systems, and the lack of a standardized set of measures of the quality of an explanation. The development of XAI-IDS research over the years indicates a clear transition from performance-oriented IDS to human-centric, transparent, and trustworthy cybersecurity systems. The following table-1 shows the summary of related work of previous years' papers-

Table 1: Summary of Related Work in XAI-IDS

Ref No.	Authors	Year	Method & Focus
[6]	Capuano et al.	2022	An extensive survey on Explainable Artificial Intelligence (XAI) in cybersecurity - taxonomy and classification of techniques.
[7]	Neupane et al.	2022	Survey on Explainable Intrusion Detection Systems (X-IDS); challenges, opportunities, and human-in-the-loop interpretation.
[8]	Masud et al.	2024	XAI for Resilient IoT Security Applications; Integrating Explainability in AI-IoT Security Frameworks.
[9]	Sharma et al.	2024	A deep learning-based IDS for IoT networks that incorporates SHAP and LIME for model interpretability.
[10]	Arreche & Abdallah	2025	Comparative analysis of white-box explainability methods (for e.g., LRP, Integrated Gradients) to develop DNN-based network security models.
[11]	Khan et al.	2024	Conduct a systematic review of explainable IDS in the context of Industry 5.0; emphasis on the challenges of adversarial XAI robustness.
[12]	Zhang et al.	2022	State-of-the-art survey of applications of XAI techniques for various cybersecurity domains, i.e., IDS and malware detection.
[13]	Adhikari & Thapaliya	2024	Interpretable AI models in malware analysis and network intrusion detection systems, particularly with regard to cybersecurity models.
[14]	Arreche et al.	2024	Evaluation framework (E-XAI) for black-box methods of explainability in network intrusion detection systems.
[15]	Chauhan & Jain	2025	Hybrid deep learning and blockchain-enabled IDS for IoT networks; enhanced dataset fusion for secure detection.
[16]	Shafin	2025	Explainable feature selection framework for phishing detection based

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

			on SHAP-based attribution techniques.
[17]	Ghadami & Rahebi	2025	A hybrid GAN-Transformer-based phishing detection system with optimization and explanation components.
[18]	Zebin et al.	2022	Explainable AI-based IDS for DNS-over-HTTPS (DoH) attacks; interpretable AI-based attack detection of anomalous attacks.
[19]	Kaur & Gupta	2023	Explainable AI-Driven IoT Security Enhancement in 6G Networks, Trust-Aware Communication Frameworks.
[20]	Charmet et al.	2024	Survey of literature on XAI for Cybersecurity; evaluation of mechanisms of transparency and accountability.
[21]	Radhika et al.	2025	Conference study on combining deep learning and XAI to improve transparency and reliability of IDS and Network Security.
[22]	Surasit Songma, Theera Sathuphan & Thanakorn Pamutha	2023	Optimized CSE-CIC-IDS-2018 based idis in three phases; Data preprocessing (cleaning, normalisation), Feature reduction (PCA and Random Forest), and Classification (XGBoost, DT, RF, KNN, MLP, LR, and Bayes). Performance optimization done based on ROC, MCC, CPU time, and model size.

It has been observed in existing research that IDS that utilize XAI perform better than black-box models, if considered from the perspective of trust, transparency, and human usability. Since XAI-based IDSs are capable of providing accurate explanations for detection outcomes, they can assist security analysts in understanding and verifying system decisions, which is highly significant in high-risk cybersecurity settings. Although the inclusion of explainability mechanisms may cause a slight increase in computational complexity, it is well worth the effort. Therefore, XAI-based IDS frameworks are considered more practical and reliable for real-world use.

IV. PROPOSED HYBRID XAI FRAMEWORK

The Hybrid Explainable Artificial Intelligence – Deep Learning Framework (XAI-HIDS++) provides an integrated approach to developing reliable and transparent IDS that integrates various state-of-the-art deep learning approaches and XAI approaches. Instead of relying on a single dataset and model, as is traditional practice, popular intrusion detection system datasets, such as CICIDS2017, NSL-KDD, and Bot-IoT, are selected for better generalization based on different types of networks. The proposed approach attempts to address the long-standing problem of balancing accuracy and interpretability.

One of the key conceptual components proposed within this framework is SHAP+LIME Aggregated Feature Selection (SLA-FS), which combines global and local explainability information to identify informative and interpretable network traffic features. The SLA-FS score for a feature Fi can be represented as:

$$SLA(F_i) = \alpha * SHAP\ (F_i) + \beta * LIME\ (F_i)$$

where α and β are weighting coefficients satisfying $\alpha + \beta = 1$. Features with higher SLA score are considered more relevant for intrusion detection. The practical implementation and optimization of these weights remain part of the future research.

The proposed system's detection engine relies on a hybrid architecture consisting of different CNN architectures to efficiently learn features from network traffic data. CNNs are used to learn the specific characteristics of network traffic data. On the other hand, the proposed hybrid architecture uses GRUs to learn the temporal characteristics of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

network traffic data. Additionally, the Transformer architecture is proposed for its ability to perform context-based reasoning using attention mechanisms to learn long-range relationships between network traffic components. Hybrid Feature Fusion represented as:

$$H_{fusion} = w_1 H_{CNN} + w_2 H_{GRU} + w_3 H_{Transformer}$$

where:

- H_{CNN} represents spatial feature representations,
- H_{GRU} represents temporal dependencies,
- $H_{Transformer}$ represents contextual relationships,
- w_1, w_2, w_3 denotes fusion weights.

The fusion mechanism conceptually integrates spatial, temporal, and contextual information to improve intrusion detection performance.

To enhance transparency, the framework embeds a dedicated explainability layer, capable of both local and global interpretation of model decisions. LIME generates instance-specific explanations at the local level, attempting to explain why a particular network activity is classified as malicious or benign. Finally, SHAP and LRP are useful when an overall view of feature importance and model behaviour across the entire dataset is needed at a higher level. This duality of explainability ensures interpretability at both the individual alert level and the model level. Dual-Level Explainability represented as:

$$E_{total} = \lambda E_{local} + (1 - \lambda)\, E_{global}$$

where:

- E_{local} corresponds to LIME-based local explanations,
- E_{global} corresponds to SHAP-based global explanations,
- λ controls the relative importance of local and global interpretability.

This conceptual formulation highlights the integration of instance-level and model-level interpretability within the proposed framework.

Finally, it includes a framework for visualizing explainability results, presenting intelligence in an informal, simple, and easily understandable way. In fact, security experts can interpret model decisions using visual plots from SHAP and LIME in real time as a way to validate alerts. In short, the proposed framework increases detection efficiency by improving interpretability, making IDS results more transparent, trustworthy, and deployable in real-world security operations.

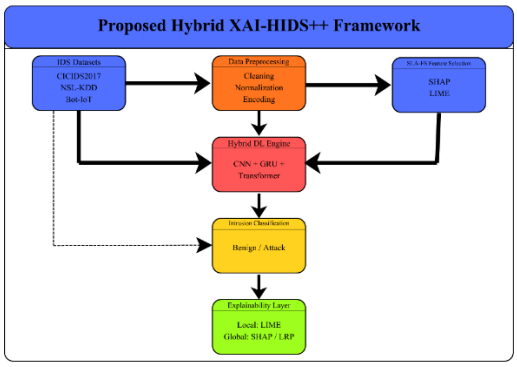

Figure 3: Proposed Hybrid XAI-HIDS++ Framework

The suggested framework is proposed in figure 3 combines hybrid deep learning and XAI to detect intrusion better. All network traffic data undergo cleaning, normalization, and encoding, and then the features are selected with explainability methods including SHAP and LIME. This processed data is then run through a hybrid deep learning engine that merges CNN, GRU, and Transformer network models to determine network traffic as malicious or benign. Lastly an explainability layer gives both local and global explanations of the model choices enhancing the transparency and confidence in the IDS.

V. CONCLUSION & FUTURE SCOPE

XAI is transforming IDS from black-box decision-making systems to transparent, accountable, and trustworthy decision-making systems. The reviewed works emphasize that the SHAP, LIME, and SLA-FS frameworks play a vital role in improving interpretability without compromising detection accuracy. The proposed hybrid XDL framework combines a CNN-GRU-Transformer model with dual-level explainability to provide an interpretable, efficient, and trustworthy IDS solution. Feature research should focus on using standardized evaluation metrics and adding real-time explainability capabilities for large-scale network data. This paradigm shift in explainable IDS is essential to ensuring ethical, interpretable, and resilient AI-based cybersecurity systems. XAI continues to transform IDS from "black box" models to fully transparent, accountable, and reliable decision support models. From a cursory analysis of the reviewed papers, it is reasonable to conclude that the use of explainability models such as SHAP and LIME, and cumulative models such as SLA-FS, is crucial as a robust way to increase model interpretability without compromising detection accuracy. This is typically achieved by

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

demonstrating the reasons behind specific intrusion detection decisions.

Hybrid deep learning models, combining CNN, GRU, and Transformer models, enhance the capabilities of IDS systems by analysing specific temporal and contextual patterns in network flow. Data models with dual-level explainability, i.e., both local and global explainability, provide valid support for IDS models to meet the objectives of human-centric cybersecurity, in accordance with ethical AI frameworks and emerging regulations to support transparency and explainability in AI and its associated models.

In the future, explainable IDS research should focus more on providing standardized and universally accepted evaluation metrics to assess the quality, faithfulness, and stability of explanations. Other specific directions include optimizing computational efficiency for real-time explainability in high-speed and large network environments. Furthermore, adaptive explanation mechanisms that can transform human-AI collaboration into patterns of questioning, refinement, and context for model decisions also hold potential. Research on the adversarial robustness of explanation methods and the integration of explainable IDS with automated response systems also holds great promise. This paradigm shift to an explainable approach to IDS is essential for developing ethical and trustworthy AI-driven cybersecurity systems.

VI. REFERENCES

[1] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access, 10, 112392-112415.

[2] Matta, S. S., & Bolli, M. (2023). Trustworthy ai: Explainability & fairness in large-scale decision systems. Review of Applied Science and Technology, 2(04), 54-93.

[3] Lansky, J., Ali, S., Mohammadi, M., Majeed, M. K., Karim, S. H. T., Rashidi, S., ... & Rahmani, A. M. (2021). Deep learning-based intrusion detection systems: a systematic review. IEEE access, 9, 101574-101599.

[4] Chauhan, D., Bahad, P., & Jain, J. K. (2024). Sustainable AI: environmental implications, challenges, and opportunities. Explainable AI (XAI) for sustainable development, 1-15.

[5] Liu, M., Xue, Z., Xu, X., Zhong, C., & Chen, J. (2018). Host-based intrusion detection system with system calls: Review and future trends. ACM computing surveys (CSUR), 51(5), 1-36.

[6] Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *Ieee Access*, *10*, 93575-93600.

[7] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, *10*, 112392-112415.

[8] Masud, M. T., Keshk, M., Moustafa, N., Linkov, I., & Emge, D. K. (2024). Explainable artificial intelligence for resilient security applications in the Internet of Things. *IEEE Open Journal of the Communications Society*, *6*, 2877-2906.

[9] Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Systems with Applications*, *238*, 121751.

[10] Arreche, O., & Abdallah, M. (2025). A comparative analysis of DNN-based white-box explainable AI methods in network security. *EURASIP Journal on Information Security*, *2025*(1), 16.

[11] Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. *Information*, *16*(12), 1036.

[12] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access*, *10*, 93104-93139.

[13] Adhikari, D., & Thapaliya, S. (2024). Explainable AI for cyber security: interpretable models for malware analysis and network intrusion detection. *NPRC Journal of Multidisciplinary Research*, *1*(9), 170-179.

[14] Arreche, O., Guntur, T. R., Roberts, J. W., & Abdallah, M. (2024). E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection. *IEEE Access*, *12*, 23954-23988.

[15] Chauhan, D., & Jain, J. K. (2025). Hybrid Deep Learning and Blockchain-Enabled Intrusion Detection System for IoT Networks using Enhanced Dataset Fusion. *Journal of Engineering Science & Technology Review*, *18*(4).

[16] Shafin, S. S. (2024). An explainable feature selection framework for web phishing detection with machine learning. *Data Science and Management*.

[17] Ghadami, R., & Rahebi, J. (2025). An explainable hybrid deep learning-optimization framework for robust phishing attack detection using GAN and transformer-based feature learning. *Ain Shams Engineering Journal*, *16*(12), 103745.

[18] Zebin, T., Rezvy, S., & Luo, Y. (2022). An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks. *IEEE Transactions on Information Forensics and Security*, *17*, 2339-2349.

[19] Kaur, N., & Gupta, L. (2024). An approach to enhance iot security in 6g networks through explainable ai. *arXiv preprint arXiv:2410.05310*.

[20] Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P. F., Han, Y., Jmila, H., ... & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, *77*(11), 789-812.

[21] Radhika, M. S., Vimala, P., Balakrishnan, C., & Sudha, K. (2025, June). Enhancing Network Security: the Role of Deep Learning and Explainable AI in Intrusion Detection Systems. In *2025 11th International Conference on Communication and Signal Processing (ICCSP)* (pp. 452-457). IEEE.

[22] Songma, S., Sathuphan, T., & Pamutha, T. (2023). Optimizing intrusion detection systems in three phases on the CSE-CIC-IDS-2018 dataset. *Computers*, *12*(12), 245.