Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 06 June-2026, Page No.- 10377-10393

DOI: 10.47191/etj/v11i06.13, I.F. – 8.595

Trade-off Between Performance, Efficiency, and Explainability in AI-Based Intrusion Detection Systems: A Systematic Review of Models, XAI, and Research Challenges

Ahmed Saleh Khaled AL-hurdi
PhD Student IT Section Aden, Yemen

ABSTRACT: Artificial Intelligence (AI)-based Intrusion Detection Systems (IDS) have significantly enhanced the detection of cyber-attacks through the adoption of machine learning and deep learning techniques. However, achieving high detection results often comes at the expense of interpretability, computational efficiency, and generalization. This paper introduces a systematic review of AI-based IDS, focusing on the integration of deep learning algorithms with Explainable Artificial Intelligence (XAI) methods. The paper analyzes a varied range of techniques, including traditional machine learning, deep learning, and hybrid models, alongside commonly used XAI methods such as SHAP and LIME. Furthermore, the review studies widely used benchmark and real-world datasets, emphasize their limitations in representing modern cyber-attack scenarios. Evaluation metrics used in the literature are also critically analyzed, revealing a lack of comprehensive assessment frameworks that incorporate performance, efficiency, and explainability. The findings determine a clear trade-off between accuracy, interpretability, and computational complexity, with most high-performing models lacking intrinsic transparency and relying heavily on post-hoc explanation methods. Finally, this paper recognizes key research gaps, including a limited combination of XAI within model architectures, over-reliance on outdated datasets, and the absence of unified Intrusion Detection Systems frameworks. These insights deliver a foundation for developing next-generation Intrusion Detection Systems solutions that balance performance, explainability, and real-world applicability.

KEYWORD: Intrusion detection system, SHAP, LIME, XAI, Deep Learning, Machine Learning, Post-hoc interpretable.

I. INTRODUCTION

The rapid growth of digital technologies, cloud computing, and Internet of Things (IoT) environments has significantly increased the volume, complexity, and sophistication of cyber threats. Modern cyber-attacks, including zero-day exploits, advanced persistent threats (APT), and polymorphic malware, pose serious risks to critical infrastructures and sensitive data across multiple domains such as healthcare, finance, and government systems. Consequently, ensuring robust and real-time cybersecurity mechanisms has become a fundamental requirement in contemporary networked systems [1].

Intrusion Detection Systems (IDS) have emerged as essential components of cybersecurity frameworks, designed to monitor network traffic and system activities to identify malicious behavior and potential security breaches. Traditional IDS approaches, particularly signature-based and rule-based systems, have demonstrated effectiveness in detecting known attacks; however, they suffer from significant limitations, including high false positive rates, inability to detect unknown threats, and lack of adaptability to evolving attack patterns. These limitations have motivated the integration of Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), into IDS to enhance detection capabilities and improve system responsiveness [2].

AI-based IDS leverages advanced learning algorithms to analyze large-scale and high-dimensional network data, enabling the detection of both known and unknown attacks with improved accuracy and reduced false alarms. Machine learning techniques allow IDS to learn patterns from historical data and adapt to new threats over time, while deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) provide powerful feature extraction and temporal modeling capabilities for complex network behaviors. As a result, AI-driven IDS has become a dominant research direction in modern cybersecurity [3].

Despite these advancements, several critical challenges remain unresolved. One of the most significant issues is the lack of interpretability in AI-based IDS models, particularly deep learning approaches, which are often considered “black-box” systems. While these models achieve high detection accuracy, their decision-making processes are not transparent, limiting trust and hindering their adoption in real-world security operations. To address this issue, Explainable Artificial Intelligence (XAI) techniques have been introduced to provide insights into model predictions and enhance transparency [1].

Furthermore, existing studies reveal additional limitations related to dataset quality, model generalization, and evaluation methodologies. Many IDS models are trained and evaluated on benchmark datasets that do not fully represent real-world attack scenarios, which affects their robustness and deployment readiness. Additionally, the lack of unified evaluation frameworks that consider accuracy, efficiency, and explainability further complicates the assessment of IDS performance [4].

In this context, this paper presents a comprehensive systematic review of AI-based intrusion detection systems, focusing on the integration of machine learning, deep learning, and explainable AI techniques. The study aims to

analyze existing models, identify key challenges, and highlight critical research gaps related to performance, interpretability, datasets, and evaluation strategies. By providing a structured and in-depth analysis, this work contributes to guiding future research toward the development of more accurate, interpretable, and practical IDS solutions.

II. SYSTEMATIC REVIEW METHODOLOGY

A. Study Design and Objective

This study adopts a systematic literature review (SLR) methodology to investigate AI-based Intrusion Detection Systems (IDS), with a particular focus on deep learning models, Explainable Artificial Intelligence (XAI) techniques, datasets, evaluation metrics, and associated research challenges. The primary objective of this review is not merely to summarize existing literature, but to systematically analyze and synthesize evidence from selected studies (Refs [1–34]) to identify key patterns, strengths, limitations, and unresolved research gaps in the domain. Unlike traditional narrative reviews, this study follows a structured, question-driven, and taxonomy-based analytical framework, enabling a critical examination of IDS research across multiple dimensions, including model architecture, interpretability, data quality, and evaluation methodologies.

B. Research Questions (RQs)

This review is guided by the following research questions:

RQ1: What types of machine learning, deep learning, and hybrid models have been proposed for intrusion detection systems, and how effective are they in terms of detection performance?

RQ2: Explainability in IDS To what extent have explainable AI techniques been integrated into IDS models to improve interpretability alongside detection accuracy?

RQ3: XAI Techniques Usage Which explainable AI techniques are most commonly used in IDS literature, and what are their strengths and limitations in interpreting model decisions?

RQ4: Dataset Representativeness How representative are existing datasets (e.g., NSL-KDD, CICIDS2017) of real-world cyber-attacks, and what limitations affect their use in IDS research?

RQ5: Evaluation Practices How are IDS models evaluated in the literature, and do current evaluation metrics sufficiently capture performance, efficiency, and explainability?

RQ6: Generalization and Robustness To what extent do existing IDS models generalize across different datasets and real-world scenarios?

RQ7: Computational Efficiency What are the computational and scalability limitations of current IDS models, especially in real-time and IoT environments?

RQ8: System-Level Integration. Is there a unified IDS framework that integrates detection accuracy, explainability, efficiency, and robustness, or are existing approaches fragmented?

Data Source and Research Strategy

The Fig.1. below presents a PRISMA flow diagram illustrating the systematic process of study selection [46]. The identification phase involved searching multiple academic databases, including IEEE Xplore, ACM Digital Library, Springer Nature, ScienceDirect, MDPI, arXiv, and Google Scholar, resulting in a total of 1100 records. Following this, 550 duplicate records were removed, leaving 550 studies for the screening phase. During screening, titles and abstracts were evaluated, leading to the exclusion of 350 records due to irrelevance to the research scope. 200 full-text articles were assessed against predefined inclusion and exclusion criteria. Of these, 160 articles were excluded for not meeting the required conditions. Finally, 40 studies were included in the systematic review, representing. The most relevant and high-quality research is aligned with the objectives of this study. One of the common search strings used with slight adjustments on different sites is as follows: "Intrusion Detection Systems" AND "deep learning" OR "neural network" OR "CNN" OR "LSTM" OR "Transformer" AND ("explainable AI" OR "XAI" OR "interpretability" OR "explanation" OR "SHAP" OR "LIME" OR "attention").

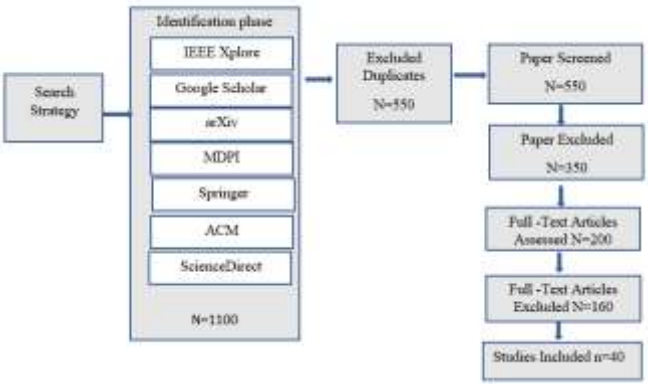

Fig. 1. Systematic Literature Review Process

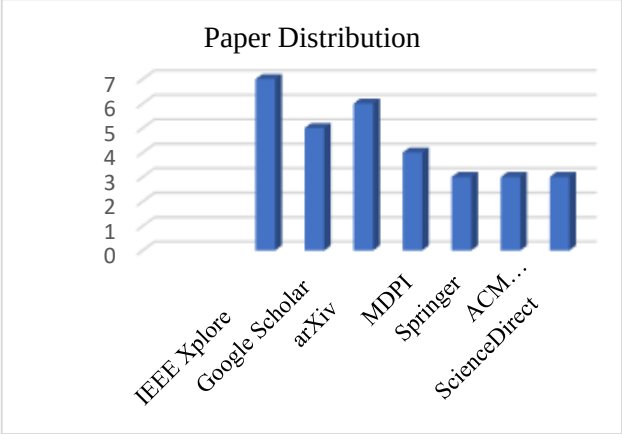

Fig. 2. Distribution of Papers by Publisher/Source

Background of AI- Models, Post-hoc Interpretability, and Datasets

AI- Models

Recent research has highlighted both the strengths and limitations of machine learning and deep learning in modern intelligent systems. Deep learning has significantly advanced the field of artificial intelligence by enabling highly accurate predictions across complex tasks such as image recognition and natural language processing. However, despite these advantages, deep learning models are often considered black-box systems, where "the reason for predictions is unknown," raising concerns about their reliability in critical applications [39]. Similarly, recent studies emphasize that the inherent opacity of AI systems can undermine user trust, accountability, and effective decision-making, particularly in high-stakes domains such as healthcare [40]. In addition, the rapid adoption of machine learning techniques has made them central to knowledge discovery across various data domains; nevertheless, the "black-box nature of contemporary ML" remains a primary challenge, limiting interpretability and transparency [41]. Furthermore, recent literature points to a

fundamental trade-off between accuracy and interpretability in deep learning models, where achieving high predictive performance often comes at the cost of model transparency [42]. These limitations highlight the urgent need for developing hybrid and explainable approaches that balance predictive power with interpretability, ensuring both high performance and trustworthy decision-making.

Post-hoc Interpretability

Recent research has raised significant concerns regarding post-hoc interpretability methods in machine learning. Argues that interpretability is "both important and slippery," and that many proposed techniques lack a clear and unified definition, making their scientific validity questionable [35]. Similarly, emphasizes that post-hoc explanations may be unreliable in high-stakes domains, recommending the use of inherently interpretable models instead of black-box explanations [36]. Furthermore, demonstrates that post-hoc methods can generate explanations based on artifacts rather than true data-driven reasoning, which raises concerns about their faithfulness to the model's actual behaviour [37]. In addition, highlights that human interpretability itself remains poorly defined, which complicates the evaluation of explanation quality [38]. These limitations suggest that post-hoc interpretability may introduce misleading or non-faithful explanations rather than a genuine understanding of model decision-making processes.

Datasets

NSL-KDD has been increasingly criticized in recent studies due to several inherent limitations. KDD'99 suffers from significant issues such as redundant records and unrealistic data distribution, which can bias machine learning models and lead to overly optimistic performance results [43]. Furthermore, both KDD'99 and NSL-KDD are derived from the DARPA 1998 dataset, which has been criticized for not accurately representing real-world network traffic, resulting in misleading evaluation outcomes [44]. Although NSL-KDD was introduced to address some of these issues, recent studies indicate that it still suffers from fundamental limitations, including class imbalance, static traffic features, and a lack of representation of modern attack types such as IoT and zero-day attacks [45]. Consequently, models trained on these datasets often fail to generalize effectively to real-world environments, limiting their practical applicability. These findings highlight the need to adopt more realistic and up-to-date datasets to ensure reliable evaluation of modern intrusion detection systems.

III. TAXONOMY OF MODELS, XAI - BASED INTRUSION DETECTION SYSTEMS

Fig.3. below demonstrates the taxonomy diagram provides a structured classification of AI-based Intrusion Detection Systems (IDS) across five key dimensions: AI models, explainability, datasets, evaluation, and research challenges. It categorizes IDS models into Deep Learning, Machine Learning, and XAI-supported approaches, while distinguishing explainability methods into post-hoc and intrinsic techniques. The taxonomy also highlights the role of datasets and evaluation metrics in assessing system performance. Finally, it summarizes the main research challenges, emphasizing the existing gap between performance, interpretability, data quality, and evaluation practices.

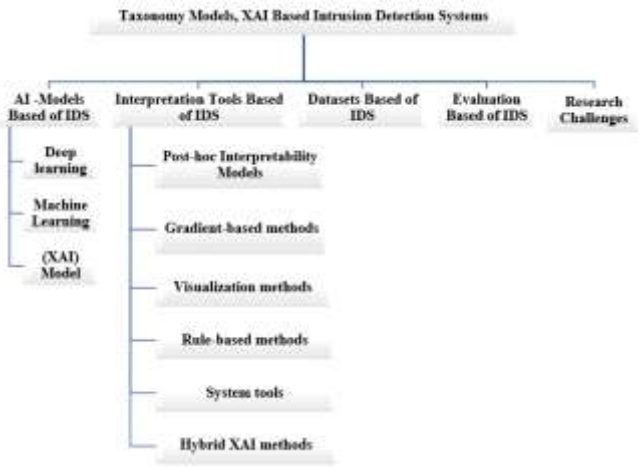

Fig.3. Taxonomy of Models, XAI - Based Intrusion Detection Systems

A. AI -Models Based on IDS

This dimension focuses on the different artificial intelligence models used in building Intrusion Detection Systems (IDS), as these models form the core of detecting attacks and analyzing network traffic. It includes Deep Learning models, traditional Machine Learning models, as well as models supported by Explainable AI (XAI) techniques. This classification aims to highlight the differences among these models in terms of performance, complexity, and interpretability, providing a foundation for a detailed analysis of their strengths and limitations.

Deep Learning Models

This category includes advanced deep learning models used in Intrusion Detection Systems to automatically learn complex patterns from high-dimensional network data. Models such as DNN, CNN, RNN, and LSTM are widely applied for their strong feature learning and temporal analysis capabilities, while AE and GAN are used for anomaly detection and handling imbalanced data. Transformer models further enhance the ability to capture global dependencies, and hybrid models combine multiple approaches to improve overall performance. Despite their high accuracy, these models often suffer from high computational cost and limited interpretability, as shown in Fig.4. below.

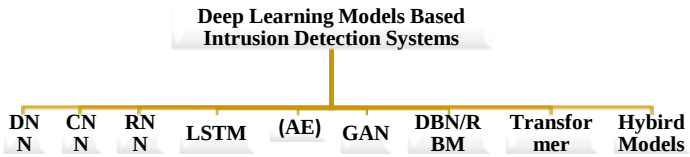

Fig.4. Deep Learning Models Based on IDS

Deep Neural Network (DNN)

The Deep Neural Network (DNN) is frequently employed in intrusion detection systems because of its capacity to learn hierarchical feature representations from high-dimensional data, as shown in several studies, including [16], [19], [22], [23], and [34]. Numerous studies have suggested that DNN models may successfully identify intricate and non-linear patterns in network traffic, with high detection accuracy on various datasets [16], [19], [23], and [34]. Additionally, several studies have improved DNN interpretability by including explainability strategies like SHAP [22]. Despite these benefits, DNN has several serious drawbacks, including a high computational cost, a reliance on huge labelled

datasets, and a black-box design that limits transparency. Furthermore, overfitting might result from poor tuning, which would reduce its capacity for generalisation.

Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is used in a number of papers [16], [17], [19], [23], [26], [27], [32]. CNN models are known for their strong automatic feature extraction capability and robustness when handling structured data [16], [26], [27]. However, CNN is less effective in modelling temporal dependencies than sequence-based models like LSTM; additionally, it suffers from high computational requirements and limited interpretability; in certain situations, it may also perform worse in recall and F1-score metrics [17], [23].

Recurrent Neural Network (RNN)

In [16], [19], [23], [26], and [27], an RNN is suggested. The Recurrent Neural Network (RNN) retains internal memory of prior inputs to interpret sequential data. Analysing time-series network data and identifying sequential attack patterns are two common uses for it [16], [19], and [23]. Although RNNs are good at capturing temporal connections, their capacity to learn long-term dependencies is constrained by the vanishing gradient problem. Additionally, it has restricted interpretability, a high computational cost, and a very lengthy training time [23], [26].

Long Short-Term Memory (LSTM)

In [16], [17], [20], [23], [26], and [27], LSTM is highlighted. An improved version of RNN called the Long Short-Term Memory (LSTM) uses gated mechanisms to solve the vanishing gradient issue. It has demonstrated excellent performance in identifying sequential assaults and modelling temporal relationships in network data [17], [20], and [26]. Because LSTM can retain long-term knowledge, it obtains great accuracy. On the other hand, it results in longer training times, greater computing costs, and more sophisticated architecture. Furthermore, it is still challenging to understand, and its performance is susceptible to hyperparameter adjustment [20], [23].

Autoencoder (AE)

AE is highlighted in [19], [20], [23], [26], and [27]. By recreating input data and spotting deviations, the Autoencoder (AE), an unsupervised deep learning model, may detect anomalies [19], [23]. It has demonstrated great accuracy (up to 97.7%) in certain experiments and is especially useful in identifying unknown assaults [26]. Its primary benefit is that it can function without labelled data. However, the quality of the training data has a significant impact on AE performance, and it could have trouble identifying infrequent assaults. It is also not interpretable and might result in reconstruction bias [20], [23].

Generative Adversarial Network (GAN)

GAN verifies in [19], [23]. By creating artificial data samples, the Generative Adversarial Network (GAN) is employed to correct class imbalance [19]. In situations where assault events are infrequent, it greatly enhances detection performance. However, instability and convergence problems make training GAN models challenging. Additionally, they are intrinsically challenging to comprehend and need large computing resources [23].

Transformer

Transformers are mentioned in [20], [23], and [26]. Without using sequential processing, the Transformer model uses attention mechanisms to identify global relationships in data [20]. It has proven to perform exceptionally well in intrusion detection, frequently above 95% accuracy [23]. It is very successful because it can simulate complicated interactions. Nevertheless, it has limited explainability, complicated training processes, and significant computing cost [20], [26].

DBN / RBM

Deep Belief Networks (DBN) and Restricted Boltzmann Machines (RBM) are unsupervised models used for feature extraction, according to DBN/RBM in [16], [27]. They were extensively employed in previous studies, but as more effective deep learning architectures have emerged, their use has decreased. Longer convergence times and greater training complexity are their primary drawbacks [16].

LSTM + SPIP

In [5], the hybrid model is demonstrated. This hybrid model combines SPIP for improved data encoding with LSTM for temporal feature learning [5]. Although it achieves great accuracy in sequential data processing, its real-time application is restricted, and its computing cost is high.

RNN + Attention + Optimization

[6] reflects the hybrid model. To improve feature importance and efficiency, this model combines RNN with optimisation techniques and attention mechanisms [6]. It increases detection accuracy but adds significant processing complexity and cost.

KD-XVAE

KD-XVAE: This hybrid discloses [14]. To achieve lightweight and real-time performance, this model combines knowledge distillation with a Variational Autoencoder [14]. Nevertheless, it has a poor capacity for generalisation.

SA-DNN + LFG

This hybrid, illustrated in [25], The SA-DNN + LFG model represents a hybrid deep learning approach that integrates a Deep Neural Network enhanced with self-attention mechanisms and a dynamic feature selection strategy referred to as LFG (Learning Feature Grouping) [25]. The incorporation of self-attention allows the model to focus on the most relevant features within the input data, thereby improving feature representation and classification performance. Meanwhile, the LFG component contributes to dynamically selecting and grouping important features, reducing noise and enhancing learning efficiency. In terms of strengths, the model demonstrates high detection accuracy and improved feature learning capability due to the combination of attention mechanisms and adaptive feature selection. However, it also presents several limitations, including the need for careful manual hyperparameter tuning, limited experimental validation across diverse datasets, and potential scalability challenges in large-scale real-world environments [25].

CNN + LSTM + AE

This hybrid approach incorporates temporal, spatial, and anomaly detection capabilities, as demonstrated in [30]. Although it is computationally costly and unsuitable for real-time applications without optimisation, it yields very high detection accuracy.

N. EnsembleGuard

In [15], EnsembleGuard is shown. EnsembleGuard is a powerful detection system that integrates deep learning, distillation methods, and tree-based models [15]. It adds system complexity and lacks integration clarity despite its great accuracy. Table 1 below illustrates clearly the key strengths and key limitations of deep learning models.

Table 1 Deep Learning Models

Model	Key Strengths	Key Limitations
Deep Neural Network (DNN)	Learns complex patterns, has high accuracy, and handles high-dimensional data	High computational cost, large data requirement, black-box
Convolutional Neural Network (CNN)	Strong spatial feature extraction, automatic learning, and high accuracy	Weak temporal modelling, high cost, low interpretability
Recurrent Neural Network (RNN)	Handles sequential data, captures temporal patterns	Vanishing gradient, slow training, high cost
Long Short-Term Memory (LSTM)	Captures long-term dependencies, strong temporal learning	Complex, high computational cost, and tuning difficulty
Autoencoder (AE)	Unsupervised anomaly detection, high accuracy	Weak for rare attacks, reconstruction bias
Generative Adversarial Network (GAN)	Handles imbalance, improves rare attack detection	Training instability, high cost
Transformer	Captures global dependencies, high performance	Very high cost, complex training
DBN / RBM	Feature extraction, unsupervised learning	Outdated, slow training
LSTM + SPIP (Hybrid DL)	Enhanced temporal learning, improved accuracy	High complexity
RNN + Attention + Optimization (Hybrid DL)	Better feature focus, improved performance	High computational cost
KD-XVAE (VAE + Distillation)	Lightweight, real-time capable	Limited generalization
SA-DNN + LFG	Dynamic feature selection, improved accuracy	Requires tuning, limited validation
CNN + LSTM + AE	Combines spatial, temporal, and anomaly detection capabilities; high detection accuracy; strong feature extraction	Very high computational cost, complex architecture, not suitable for real-time without optimization
EnsembleGuard	Robust ensemble combining tree-based models, deep learning, and distillation; high accuracy and resilience	High complexity, unclear integration strategy, and difficult to interpret

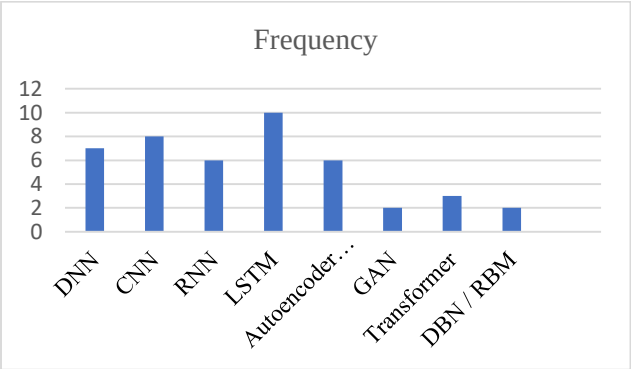

Fig. 5. Deep Learning Models Frequency

Machine Learning Models

Fig. 6. Describe traditional machine learning models that are employed in intrusion detection systems to categorise and identify malicious network activity that falls under this category. Due to their effectiveness, ease of use, and comparatively better interpretability than deep learning models, models like RF, SVM, KNN, LR, and NB are frequently employed. Furthermore, sophisticated methods like AdaBoost and LightGBM raise performance through boosting strategies, while ensemble and hybrid models integrate many algorithms to improve detection robustness and accuracy. Despite these benefits, these models could have trouble scaling in large-scale network systems and handling complicated data patterns.

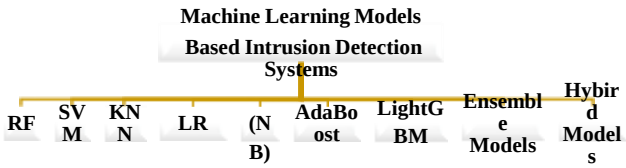

Fig. 6. Machine Learning Models-Based Intrusion Detection Systems

Random Forest (RF)

RF is validated in [18], [22], [29], [31], and [32]. Several decision trees are combined in the Random Forest (RF), a reliable ensemble learning approach that enhances classification performance. Because it can handle complicated data and decrease variation, it has continuously shown good performance in intrusion detection tasks [18], [29], [31]. Furthermore, compared to deep learning models, RF provides a certain degree of interpretability. However, its computational cost rises with dataset size [22], [32], and it may experience overfitting when the number of trees is too high.

Support Vector Machine (SVM)

SVM is supported in [18], [21], [22], [29], and [31]. An effective classification technique that creates ideal decision boundaries is the Support Vector Machine (SVM). It can use kernel functions to represent non-linear interactions and works well on small and medium-sized datasets [18]. However, choosing the right kernel is difficult, and SVM training becomes computationally costly for big datasets. Moreover, it is not scalable in high-dimensional settings [31].

K-Nearest Neighbours (KNN)

KNN is highlighted in [18], [22], and [31]. A straightforward distance-based classification technique is called K-Nearest

Neighbours (KNN). It doesn't require training and is simple to apply. Nevertheless, it has poor scalability with big datasets, high prediction time, and susceptibility to noise [22], [31].

Logistic Regression (LR)

LR provides evidence in [7], [18], and [29]. A popular linear model for binary classification is logistic regression (LR). It is appropriate for baseline intrusion detection systems because it is interpretable and computationally efficient [7]. Its main drawback, though, is that it is less useful in complex network situations due to its incapacity to capture non-linear interactions [18].

Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' theorem, as NB confirms in [18], [29]. It is quick and effective, particularly when dealing with big datasets. However, its feature independence assumption is frequently incorrect, which might have a detrimental effect on performance in practical situations [18], [29].

AdaBoost

AdaBoost is an ensemble learning strategy that iteratively improves weak classifiers; it can increase detection performance, but it is very sensitive to noise and outliers, and it has scalability issues when dealing with big datasets [18], [31].

LightGBM

[22], [31] Make use of LightGBM. The gradient boosting framework LightGBM is renowned for its effectiveness and quickness. Although it needs careful hyperparameter optimisation, it works effectively on large-scale datasets. Overfitting can result from improper setup [22], [31].

Ensemble Voting (DT, RF, SVM)

The ensemble is described in [21]. The Ensemble vote model generates a final prediction based on majority vote by combining many machine learning classifiers, such as Decision Tree, Random Forest, and Support Vector Machine [21]. This method improves overall classification accuracy by reducing individual model faults by using the strengths of many models. Strengths: By combining judgements from several classifiers, Ensemble Voting improves resilience and lowers false positives. However, the mixing of different models may result in decreased interpretability, greater processing costs, and increased system complexity [21].

Logistic Regression + Active Learning + Optimization

[7] Make advantage of this hybrid. To increase training effectiveness and model performance, this hybrid model combines active learning methodologies, optimisation approaches, and logistic regression [7]. While optimisation techniques improve parameter adjustment, active learning carefully selects the most instructive samples. The model is appropriate for real-world applications with less data since it is effective and eliminates the requirement for large labelled datasets. However, logistic regression's primary drawback is its linear character, which limits its capacity to represent intricate non-linear correlations in network traffic data [7].

RF + SVM + DNN

[8] illustrates the hybrid concept. In order to take advantage of their complementary capabilities, this hybrid approach blends deep learning and machine learning models [8]. Although it offers balanced performance, it raises system complexity and latency. Table 2 below highlights key strengths and key limitations of machine Learning models

Table 2 Machine Learning Models

Model	Key Strengths	Key Limitations
Random Forest (RF)	Robust, reduces variance, strong performance	Overfitting risk, computational cost
Support Vector Machine (SVM)	High accuracy, effective non-linear classification	Slow training on large data
K-Nearest Neighbours (KNN)	Simple, easy implementation	Slow prediction, noise sensitive
Logistic Regression (LR)	Efficient, interpretable	Cannot model non-linear relationships
Naive Bayes (NB)	Fast, scalable	Unrealistic independence assumption
AdaBoost	Improves weak learners	Sensitive to noise, scalability issues
LightGBM	Fast, efficient	Requires tuning
Ensemble Voting (DT, RF, SVM)	Reduces errors, improves accuracy	Computational cost, complexity
LR + Active Learning + Optimization	Efficient, improved data usage	Limited to linear relationships
RF + SVM + DNN	Combines the strengths of machine learning and deep learning, balanced performance, and improved classification accuracy	High latency, complex integration, and increased computational cost.

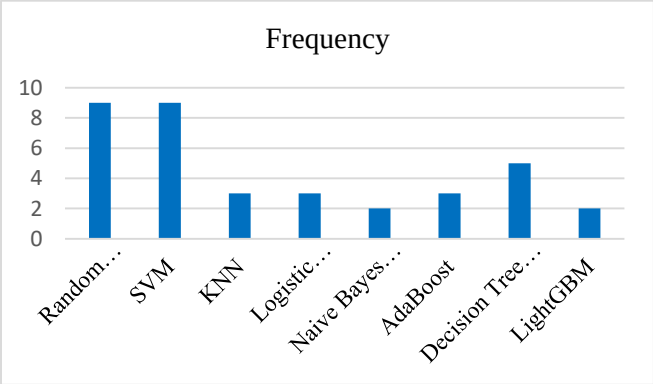

Fig. 7. Machine Learning Models Frequency

Explainable AI (XAI)

SHAP

SHAP is used in [22], [28], and [1]. The SHAP approach offers reliable and logically supported explanations of feature significance. It is computationally costly, albeit [24].

LIME

LIME is applied in [28], [1]. The LIME method explains model predictions locally. However, it suffers from instability and limited generalization [28].

LRP / DeepLift

These methods, as shown in [24], offer intrinsic explanations for neural networks, but they are memory-intensive and

restricted to particular topologies. Even though individual models like DNN, CNN, LSTM, and Random Forest have proven to perform well and consistently in numerous studies, their interpretability, computational cost, and generalisation limitations underscore the need for hybrid and explainable approaches to create a balanced and effective intrusion detection system. Table 3 below indicates the key strengths and key limitations of Explainable AI (XAI) Models.

Table 3 Explainable AI (XAI) Models

Model	Key Strengths	Key Limitations
SHAP	Consistent, theoretically sound explanations	High computational cost
LIME	Local interpretability	Instability
LRP/ DeepLift	Internal explanations for neural networks	Limited applicability

Overall Most Used Models Ranking

Fig. 8. Demonstrates that the most popular deep learning model, according to the analysis of the reviewed literature, is LSTM because of how well it captures temporal relationships in network traffic. Similar to this, machine learning-based techniques are dominated by Random Forest and Support Vector Machine because of their resilience and potent generalisation abilities. Although hybrid models are less common, their use is growing in order to combine the complementary characteristics of many algorithms and improve detection accuracy.

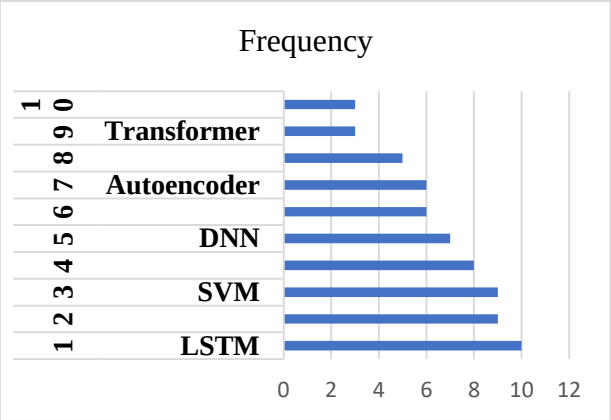

Fig. 8. Overall Most Used Models Ranking

B. Interpretation Tools

The many sorts of explainability approaches used in intrusion detection systems are presented in this section. After model training, post-hoc interpretability models are used to explain predictions without changing the original model. Gradient-based approaches, particularly in deep learning models, use model gradients to pinpoint crucial elements impacting choices. Graphical representations are used in visualisation techniques to assist in comprehending patterns and modelling behavior. Rule-based approaches increase transparency by offering logical rules that are understandable to humans. In order to facilitate interpretability, system tools concentrate on tracking and evaluating system-level data, such as logs and events. Furthermore, hybrid XAI approaches integrate many explanation strategies to enhance interpretability and accuracy. When combined, these strategies seek to improve IDS models' dependability, transparency, and credibility.

Post-hoc Explainability Methods

This idea appears as the primary node in the picture, which depicts a hierarchical structure of Intrusion Detection Systems based on Post-hoc Explainability Methods. It branches into three crucial techniques: SHAP, LIME, and Permutation Feature Importance, as shown in Fig. 9. Since these techniques come within the post-hoc explainability framework and each uses a distinct strategy to understand model decisions and improve transparency, this structure indicates a taxonomical link.

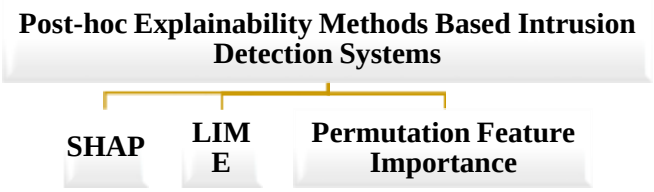

Fig. 9. Post-hoc Explainability Methods-Based Intrusion Detection Systems

SHAP

SHAP is employed in [5], [6], [7], [8], [9], [12], [13], [14], [17], [18], [21], [22], [23], [25], [26], [28], [29], [30], [31], [32], [33], [1], [34]. One of the most popular post-hoc explainability methods is the SHAP approach, which offers a theoretically sound framework for feature attribution based on cooperative game theory. It allows for both global and local interpretability across various machine learning and deep learning models by quantifying each feature's contribution to the final prediction [5] to [34]. SHAP is especially well-suited for security-sensitive applications like intrusion detection systems since it is thought to be extremely dependable and consistent in describing black-box models. Explicitly identifying influential characteristics, it enhances transparency and facilitates model evaluation. However, when applied to big datasets or intricate deep learning systems, its primary drawbacks are its high computational cost, memory usage, and scalability problems.

LIME

[7], [8], [9], [12], [13], [14], [21], [22], [23], [25], [26], [28], [30], [31], [32], [33], [1], [34] all include LIME. The LIME approach is a local post-hoc explanation methodology that uses an interpretable surrogate model to approximate a black-box model's decision boundary. By altering input data and tracking changes in output, it explains individual predictions [7] through [34]. LIME is extensively utilised because of its simplicity, versatility, and model-agnostic nature, making it adaptable to both machine learning and deep learning models. However, in comparison to global approaches like SHAP, it has very low fidelity, susceptibility to input disturbances, and instability in explanations. It also adds computing cost and might result in inconsistent explanations across runs.

Permutation Feature Importance

[5] adopts the Permutation Feature Importance. Permutation Feature Importance measures how each feature affects the model's performance when feature values are shuffled at random. This approach may be used with a variety of algorithms because it is straightforward and model-independent [5]. It helps determine the most important variables in prediction tasks by offering a simple global interpretation of feature importance. However, its depth of

explanation is limited since it can not capture complicated feature connections and becomes computationally costly for big datasets. Table 4 below describes key strengths and key limitations of Post-hoc Feature Attribution Methods.

Table 4 Post-hoc Feature Attribution Methods

Interpretation Tool	Key Strengths	Key Limitations
Integrated Gradients (IG)	Stable attribution, mathematically grounded	Only for differentiable models, the computational cost
Grad-CAM	Visual explanations, CNN interpretability	Limited to CNNs, coarse heatmaps
Saliency Maps	Simple visual explanation, gradient-based	Noisy, unstable, sensitive to perturbations

Gradient-based Explainability Methods

This idea is the central node in the figure, which depicts a hierarchical structure of Gradient-based Explainability Methods. It branches out into three important methods: Integrated Gradients, Grad-CAM, and Saliency Maps, as shown in Fig. 10. Since these techniques are all part of the gradient-based explainability framework and use gradient information to analyse model decisions and highlight the most important characteristics influencing predictions, this structure represents a taxonomical link.

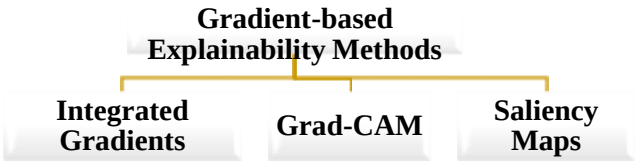

Fig. 10. Gradient-based Explainability Methods

Integrated Gradients (IG)

IG is described in [12], [24], and [26]. By building up gradients along a straight path from a baseline input to the actual input, the Integrated Gradients method is a gradient-based attribution technique intended to explain deep neural network predictions. It offers a method for measuring feature relevance in differentiable models that is based on mathematics [12], [24], and [26]. Deep learning-based intrusion detection systems might benefit from Integrated Gradients, as it is regarded as robust and theoretically sound. Compared to simple gradient algorithms, it guarantees consistency in feature assignment and lowers noise. However, when applied to large-scale neural networks, it can be computationally costly, and its usefulness is restricted to differentiable models.

Grad-CAM

There are references to Grad-CAM in [9], [12], [13], and [28]. By creating class-discriminative localisation maps, the Grad-CAM approach is frequently used to explain convolutional neural networks. It draws attention to the areas of input data that have the greatest influence on the model's choice [9], [12], and [28]. Grad-CAM offers clear visual explanations that are especially helpful for CNN-based and image-based intrusion detection models. By displaying spatially significant areas, it improves interpretability. Nevertheless, it is limited to convolutional designs and may result in noisy or coarse visualisations, which restrict accurate feature-level interpretation.

Saliency Maps

In [28], [1], saliency maps are presented. Gradient-based visualisation methods called saliency maps show how sensitive model outputs are to input inputs. By calculating gradients of the output with respect to input factors, they give pixel-level or feature-level relevance [28], [1]. These techniques are straightforward and offer clear visual explanations. They are less reliable in intricate intrusion detection systems, though, because they are frequently loud, unstable, and sensitive to minute changes in input.

Table 5 below explains the key strengths and key limitations of Gradient-based Explainability Methods.

Table 5 Gradient-based Explainability Methods

Interpretation Tool	Key Strengths	Key Limitations
Integrated Gradients (IG)	Stable attribution, mathematically grounded	Only for differentiable models, the computational cost
Grad-CAM	Visual explanations, CNN interpretability	Limited to CNNs, coarse heat maps
Saliency Maps	Simple visual explanation, gradient-based	Noisy, unstable, sensitive to perturbations

Visualization-based XAI Methods

SHAP Plots & Feature Importance Visualization (FIV), Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE), and Accumulated Local Effects (ALE) are important techniques that branch out from this concept, which serves as the main node in the diagram's hierarchical structure of Visualization-based XAI Methods, as shown in Fig.11. Since these techniques come under visualization-based explainability and each focuses on visually depicting model behavior and feature impacts to enhance interpretability and understanding, this structure represents a taxonomical link.

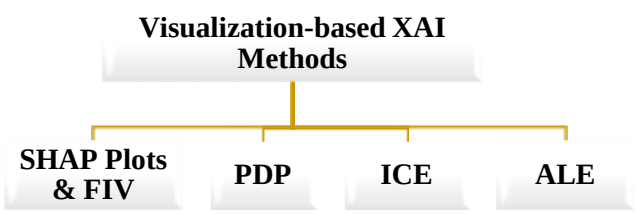

Fig.11. Visualization-based XAI Methods

SHAP Plots & Feature Importance Visualizations

[18] discusses SHAP Plots and Feature Importance Visualizations. Visualization-based SHAP plots, such as waterfall and summary plots, offer graphical depictions of feature contributions. These techniques facilitate more intuitive interpretation of both local and global model

behavior [18]. They are often employed in security analytics and enhance comprehension of model conclusions. Nevertheless, they continue to rely on the underlying SHAP calculation cost and do not directly address misclassification problems.

Partial Dependence Plots (PDP)

PDP is explained in [5], [12], and [18]. By averaging over all other data, the Partial Dependence Plot displays a feature's marginal impact on model predictions. It offers a comprehensive understanding of feature influence [5], [12], [18]. PDP is helpful for general model analysis since it is straightforward to understand. Nevertheless, it ignores feature interactions and presumes feature independence, which might result in incorrect interpretations in coupled datasets.

ICE (Individual Conditional Expectation)

[5], [12] citing ICE. By displaying feature impacts at the instance level, ICE charts go beyond PDP. This enables a thorough examination of each sample's model behavior [5], [12]. Although ICE offers greater insights into model heterogeneity, it has scaling issues and becomes challenging to comprehend in big datasets.

ALE (Accumulated Local Effects)

By taking feature dependencies into account and offering objective local explanations, the Accumulated Local Effects approach outperforms PDP [8].

Compared to PDP, ALE offers more correct interpretations and is more reliable in correlated feature spaces. However, it is less user-friendly for non-expert users and computationally more costly. Table 6 below explains the key strengths and key limitations of visualisation-based XAI methods.

Table 6 Visualization-based XAI Methods, Rule-based Explainability Methods

Interpretation Tool	Key Strengths	Key Limitations
PDP (Partial Dependence Plot)	Global feature effect visualization, simple	Assumes feature independence, ignores interactions
ICE (Individual Conditional Expectation)	Instance-level explanations	Hard to scale, complex interpretation
ALE (Accumulated Local Effects)	Handles correlated features better than PDP	Computationally expensive
SHAP Plots	Clear feature contribution visualization	High computation cost dependency
Feature Importance Plots	Easy interpretation	No causal explanation

This idea functions as the central node in the diagram's hierarchical structure of Rule-based Explainability Methods, which branches off into two crucial methods: Rule Extraction and RuleFit, as emphasized in Fig.12. Since these techniques come under rule-based explainability and concentrate on producing comprehensible rules that explain model decisions and improve transparency, this structure represents a taxonomical link.

.

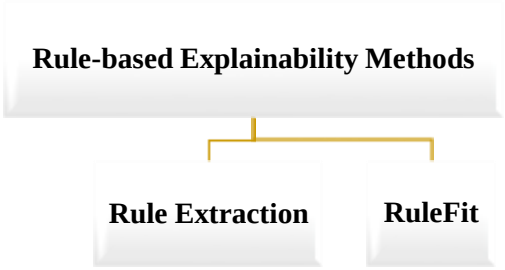

Fig. 12. Rule-based Explainability Methods

Rule Extraction

The documentation for Rule Extraction is found in [10]. Rule extraction techniques translate intricate deep learning or machine learning models into decision rules that are comprehensible to humans. These guidelines provide an interpretable approximation of model behavior [10]. This method greatly improves transparency and is appropriate for compliance and auditing. But for deep models, rule development may get quite complicated, and scalability is still a significant drawback.

RuleFit

[34] reports on RuleFit. RuleFit creates interpretable prediction rules by combining rule-based learning with linear models. By combining rule sets with feature weights, it strikes a compromise between interpretability and accuracy [34]. Although it increases explainability, its high computing cost, intricate rule structures, and decreased scalability in high-dimensional datasets are its drawbacks. Table 7 below describes the key strengths and key limitations of rule-based explainability methods.

Table 7 Rule-based Explainability Methods

Interpretation Tool	Key Strengths	Key Limitations
Rule Extraction	Human-readable rules, interpretable decisions	Complex rule generation, scalability issues
RuleFit	Combines rules + linear models, interpretable	High complexity, reduced scalability

System-level Monitoring & Security XAI Tools

ETW / auditd / ELK / sysdig / Drain

[16] mentions this instrument, Drain. For intrusion detection systems, system-level monitoring technologies like ELK Stack and audit frameworks offer real-time recording, anomaly tracking, and behavioral analysis [16]. These tools work quite well for scalable log analysis and real-time monitoring. Nevertheless, they are not appropriate as stand-alone XAI techniques due to their high computing resource consumption and lack of feature-level explainability.

Hybrid & Meta Explainability Methods

Fig.13. below explains a hierarchical structure of Hybrid & Meta Explainability Methods, with Model Distillation in conjunction with Visualization, Attention Mechanisms as Explainability techniques, General XAI Frameworks, and Counterfactual Explanations as the four main branches. Because these techniques combine many explainability techniques or higher-level frameworks to offer more thorough, adaptable, and reliable interpretations of model

behavior and decision-making processes, this structure represents a taxonomical link.

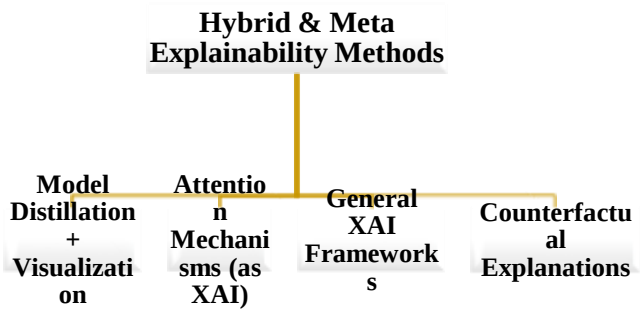

Fig. 13. Hybrid & Meta Explainability Methods

Model Distillation + Visualization

In [15], model distillation and visualisation are used. In order to mimic complicated black-box systems, model distillation-based explainability algorithms provide smaller surrogate models together with visualisation tools for interpretation [15]. This method simplifies model behavior, which enhances interpretability. Nevertheless, it lacks a standardised structure and might not adequately convey the intricacy of the original model, resulting in explanations that fall short.

Attention Mechanisms (as XAI)

[23], [27] discuss attention mechanisms. By allocating learned weights during model inference, attention-based explainability draws attention to significant input characteristics. Deep learning-based intrusion detection systems make extensive use of it [23], [28]. By dynamically displaying feature significance, attention enhances interpretability. However, it may not accurately reflect genuine causal importance and does not always offer faithful answers.

General XAI Frameworks

In [19], [27], general XAI frameworks are discussed. By combining methods and design concepts, general XAI frameworks seek to increase machine learning systems' transparency and trustworthiness [19], [27]. Despite improving interpretability, they are not commonly used in real-world intrusion detection systems and lack standardisation.

Counterfactual Explanations

CE is recognised in [1]. Alternative inputs that demonstrate how small adjustments can influence model predictions are produced via counterfactual explanation techniques. For the purpose of interpreting decisions, they offer intuitive "what-if" reasoning [1]. They are very helpful in enhancing user trust and comprehending choice limitations. Nevertheless, they can result in scenarios that are unrealistic or impractical and are computationally costly. Table 8 highlights key strengths and key limitations of Hybrid & Meta Explainability Methods.

Table 8 Hybrid & Meta Explainability Methods

Interpretation Tool	Key Strengths	Key Limitations
Model Distillation with Visualization	Simplifies black-box models, improves interpretability	No standard framework, incomplete explanations
Attention-based Explainability	Dynamic feature importance, partial interpretability	Not always faithful explanations
General XAI Frameworks	Improves transparency and trust	Lack of standardization
Counterfactual Explanations	"What-if" reasoning, intuitive explanations	High computational cost, unrealistic scenarios

Because of its solid theoretical underpinnings and capacity to offer both global and local explanations, SHAP is the most used approach in intrusion detection systems, according to an analysis of explainability methodologies as shown in Fig.14. Because LIME is model-agnostic, it is also widely utilised. Nevertheless, the employment of gradient-based and visualisation techniques like Grad-CAM, PDP, and ICE is less common, suggesting that post-hoc feature attribution techniques predominate in the XAI environment in IDS research.

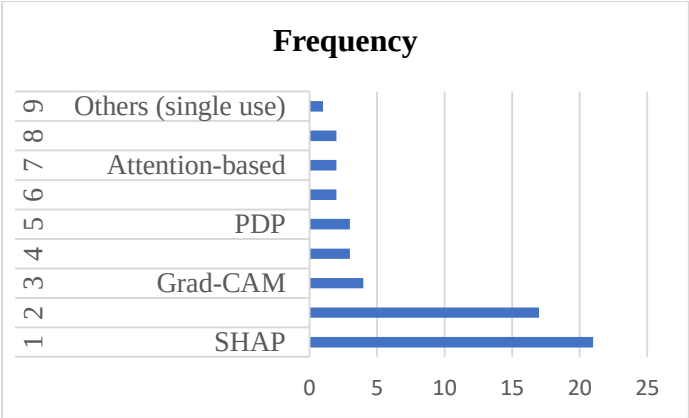

Fig.15. Most Used XAI Methods Ranking

C. Dataset Based on IDS

KDD99 / KDDCUP are used in [11], [16], [19], [23], [27], [32], and [34]. One of the first and most popular benchmarks in intrusion detection research is the KDD Cup 99 dataset. It has been widely used to assess machine learning models and offers a large-scale network traffic dataset with several attack types [11], [16], [19]. KDD99 has a number of drawbacks despite its historical significance, such as out-of-date traffic patterns, duplicate information, and a lack of representation of contemporary cyberthreats. Its usefulness in actual intrusion detection settings is greatly diminished by these problems [23], [32].

NSL-KDD is used in [5], [6], [8], [9], [11], [15], [16], [17], [19], [22], [23], [24], [25], [26], [27], [32], [34]. An enhanced version of KDD99, the NSL-KDD dataset, eliminates duplicate entries and offers a more equitable assessment standard. Because of its accessibility and ease of use, it is frequently employed for comparative research [6], [16], and [17]. It is still regarded as antiquated, though, and does not

account for contemporary network behaviors like IoT traffic or sophisticated persistent attacks. Furthermore, generalisation to real-world settings is constrained by its restricted variety [23], [26].

UNSW-NB15 is emphasised in [5], [9], [10], [15], [16], [19], [23], [25], [26], [27], and [34]. By including genuine contemporary assault scenarios, the UNSW-NB15 dataset is a contemporary benchmark created to address the shortcomings of previous datasets. Compared to NSL-KDD, it offers more varied and current network traffic patterns [10], [16]. It does not accurately reflect real-world IoT contexts, though, and it still has a class imbalance [23], [25].

CICIDS Family (CICIDS2017 / CICIDS2018 / CICIoT / CIC-IoV)

[8], [9], [11], [16], [18], [19], [21], [22], [23], [24], [27], [28], and [31] all use the CIC family. Among the most popular contemporary intrusion detection benchmarks are the CICIDS2017 and associated CIC datasets. They are appropriate for both machine learning and deep learning assessment, and they provide a variety of attack types and realistic traffic patterns [21], [28]. Compared to KDD-based datasets, these datasets are thought to be more typical of real-world settings. Nevertheless, they continue to face problems such as labelling errors, data imbalance, and artificial traffic artefacts [22], [24].

IoT & IoT-specific datasets (TON_IoT, Bot-IoT, N-BaIoT)

[5], [16], [19], [23], [25], [26], and [30] make use of IoT and IoT-specific datasets (TON_IoT, Bot-IoT, and N-BaIoT). IoT and cyber-physical systems are the focus of the TON_IoT and related IoT datasets, such as Bot-IoT and N-BaIoT. They offer multi-source data, including network logs, realistic IoT traffic, and a variety of attack scenarios [25], [30]. However, a lot of these datasets still have problems with class imbalance, restricted generalisation, and synthetic generation bias, which makes them less useful in practical applications [23].

Industrial & SCADA datasets (SWaT, WADI, IEC 60870-5-104)

Industrial and SCADA datasets are used in [20], [29] (SWaT, WADI, IEC 60870-5-104). Industrial control environments and real-world cyber-physical systems are represented by industrial datasets like SWaT and WADI. Their actual operating data makes them very helpful for anomaly detection in critical infrastructure [20]. However, they frequently have poor generalisation across many settings, restricted availability, and domain-specific limitations [29].

Malware & Security datasets (DREBIN, EMBER, VirusShare)

The DREBIN, EMBER, and VirusShare malware and security datasets are confirmed in [13]. Binary analysis and malware detection are done using malware datasets like DREBIN and EMBER. Large-scale security data and actual malware samples are provided by these databases [13]. However, their long-term efficacy is diminished by a significant class disparity and changing virus trends.

System-call datasets (ADFA-LD, CTU-13)

In [27], [33], system-call datasets (ADFA-LD, CTU-13) are verified. By examining system behavior, system-call-based datasets, like ADFA-LD, concentrate on host-based intrusion detection. They are helpful for IoT and endpoint security studies, and they offer realistic behavioral patterns [33]. Nevertheless, they frequently have an imbalance and poor generalisation, and they need extensive preprocessing.

General multi-dataset studies

In [12], [1], general multi-dataset investigations are validated. Several datasets are used in certain research to enhance comparability and generalisation in various contexts. Although this method improves coverage and robustness, it has uneven assessment metrics across datasets and lacks standardisation [12], [1]. Finally, Table 9 demonstrates the comparative analysis of datasets in terms of key strengths and key limitations.

Table 9 Comparative Analysis of Datasets

Dataset Category	Key Strengths	Key Limitations
KDD99 / KDDCUP	Large-scale, widely used benchmark	Outdated, redundant records, unrealistic traffic
NSL-KDD	Improved version of KDD99, balanced evaluation	Still outdated, limited real-world relevance
UNSW-NB15	Modern traffic, realistic attacks	Class imbalance, limited IoT realism
CICIDS Family	Realistic traffic, rich features, modern attacks	Data imbalance, labelling noise
IoT Datasets (TON_IoT, Bot-IoT, N-BaIoT)	IoT-focused, multi-source data	Synthetic bias, imbalance, and limited generalization
Industrial Datasets (SWaT, WADI, IEC 104)	Real cyber-physical systems	Limited availability, domain-specific
Malware Datasets (DREBIN, EMBER)	Real malware samples, large-scale	High imbalance, evolving threats
System-call Datasets (ADFA, CTU-13)	Host-based realistic behavior	Preprocessing complexity, imbalance
Multi-dataset Studies	Broad coverage, improved generalization	Lack of standardization

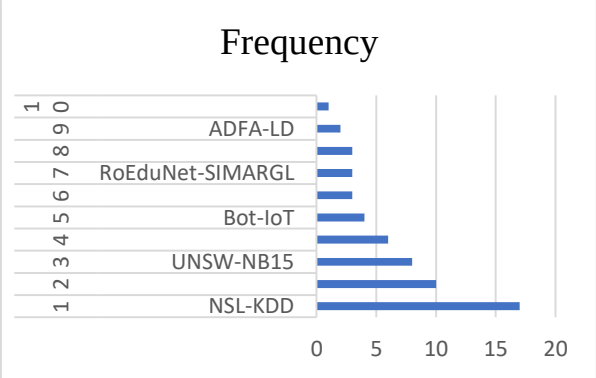

Fig. 16. Top Most Used Datasets Ranking

The dataset analysis reveals that because of its historical relevance and ease of use in intrusion detection research,

NSL-KDD continues to be the most often utilised benchmark dataset, as shown in Fig.15. However, because of their accurate traffic representation, more recent datasets like CICIDS2017 and UNSW-NB15 are being used more and more. The trend toward IoT security research is also reflected in the growing interest in IoT-based datasets like Bot-IoT and TON_IoT. However, the research continues to rely mostly on out-of-date benchmark datasets, suggesting a gap in the adoption of real-world datasets.

D. Evaluation Based on IDS

Accuracy, Precision, Recall, F1-score Metrics

Accuracy, Precision, Recall, F1-score Metrics are supported in [5], [7], [8], [9], [10], [11], [14], [16], [17], [19], [21], [23], [26], [27], [30], [31], [33], [34] The most widely used evaluation metrics in intrusion detection systems are accuracy, precision, recall, and F1-score, which collectively measure classification effectiveness under different perspectives. These metrics are widely adopted across both machine learning and deep learning models based on IDS papers due to their simplicity and interpretability [5], [16], [19], [23]. Accuracy delivers an overall performance measure, while precision and recall evaluate false positive and false negative rates, respectively. The F1-score balances both metrics, making it particularly appropriate for imbalanced datasets commonly found in cybersecurity applications [21], [27]. Despite their popularity, these metrics alone are insufficient for real-world evaluation as they do not account for computational cost, latency, or energy consumption.

ROC-AUC and PR-AUC Metrics

ROC-AUC and PR-AUC Metrics are substantiated in [7], [8], [9], [16], [20], [30], [31]. The ROC-AUC and Precision-Recall AUC metrics are broadly used to evaluate model discrimination capability across different threshold settings. These metrics afford a more comprehensive evaluation compared to accuracy alone, especially in imbalanced datasets [7], [20]. ROC-AUC measures the trade-off between true positive rate and false positive rate, while PR-AUC is more sensitive to class imbalance. However, much research relies heavily on these metrics without integrating real-time constraints or system efficiency, limiting their practical applicability [9], [26].

False Positive Rate (FPR), False Negative Rate (FNR), and Confusion Matrix Metrics

False Positive Rate (FPR), False Negative Rate (FNR), and Confusion Matrix Metrics are emphasized in[10], [14], [19], [21], [27], [33]. FPR, FNR, and confusion matrix-based metrics are vital for evaluating classification errors in intrusion detection systems. These metrics offer detailed insights into model misclassification behavior [10], [21]. Low FPR is particularly important in cybersecurity applications to decrease false alarms, while FNR indicates the system's ability to detect real attacks. However, many studies report these metrics without considering system-level performance such as latency or real-time deployment constraints [14], [33].

Advanced Performance and System-level Metrics

Advanced Performance and System-level Metrics are reinforced in [20], [22], [25], [26], [28], [31]. Advanced evaluation metrics such as latency, memory consumption, energy efficiency, Mean Time to Detect (MTTD), and VUS (Volume Under Surface) have been presented to evaluate the real-world applicability of IDS models. These metrics deliver a more holistic evaluation beyond classification accuracy [20], [26]. Despite their importance, most papers still lack standardized evaluation frameworks incorporating both performance and efficiency metrics. This leads to inconsistencies in comparing different models across datasets and architectures [25], [28].

Explainability and Trust-based Metrics

Explainability and Trust-based Metrics are demonstrated in [22],[28],[32],[1]. Explainability-related metrics such as fidelity, sparsity, stability, robustness, completeness, and user trust have been introduced to evaluate interpretability in AI-based IDS systems. These metrics aim to quantify how well a model explanation reflects its actual behavior [22], [28]. Although these metrics improve interpretability assessment, they are often subjective, lack standardization, and vary significantly across studies. Additionally, many works participate in human trust evaluation, which announces subjectivity and inconsistency in results [32], [1].

Hybrid Evaluation Metrics (Combined Performance + Explainability)

Hybrid Evaluation Metrics are illustrated in [18], [22], [24], [26], [28]. Some recent papers propose hybrid evaluation frameworks that combine classification metrics with explainability and efficiency measures. These contain combinations of accuracy, F1-score, robustness, sparsity, and latency to deliver a more comprehensive evaluation [18], [26]. However, these frameworks still suffer from the absence of unified benchmarking standards, making cross-study comparison difficult [24], [28]. Finally, Table 10 below demonstrates a comparative analysis of evaluation metrics.

Table 10 Comparative Analysis of Evaluation Metrics

Metric Category	Key Strengths	Key Limitations
Accuracy / Precision / Recall / F1-score	Simple, widely used, interpretable	Insufficient for real-world evaluation
ROC-AUC / PR-AUC	suitable for imbalanced data, threshold-independent	No real-time evaluation consideration
FPR / FNR / Confusion Matrix	Measures classification errors precisely	No system-level performance evaluation
Efficiency Metrics (Latency, Memory, Energy, MTTD)	Real-world applicability, system-level evaluation	Lack of standardization
Explain ability Metrics (Fidelity, Trust, Stability)	Improves interpretability evaluation	Subjective, non-standardized
Hybrid Evaluation Metrics	Combines performance + interpretability	No unified benchmark

The analysis of top-performing IDS papers reveals that hybrid and ensemble-based models, particularly those combining deep learning with optimization or distillation techniques, achieve the highest result levels, often above 99%. SHAP emerges as the most commonly used explainability tool among high-performance models, followed by LIME. Additionally, modern datasets such as CICIDS2017 and IoT-based datasets are frequently used in high-accuracy studies, although traditional datasets like NSL-KDD still appear in top-performing results, as shown in Table 11.

Table 11: Top 10 Highest Accuracy Studies in IDS

Rank	Ref	Model	XAI Technique	Dataset	Accuracy
1	[10]	KD-XVAE (VAE With Distillation)	SHAP	HCRL, CIC-IoV2024	≈ 100%
2	[9]	Transparent with Black-box Models	SHAP	DREBIN, EMBER, Virus Share	> 99%
3	[13]	CNN and LSTM	SHAP	NSL-KDD	~ 99%
4	[28]	CNN and RF	SHAP / LIME	KDD99, NSL-KDD	~ 98% – 99.2%
5	[11]	Ensemble Guard (Tree with DL and Distillation)	Model Distillation	NSL-KDD, UNSW-NB15, CICIDS2017	0.98 – 0.99
6	[21]	SA-DNN with LFG	SHAP / LIME	IoT datasets	97.9% – 99.6%
6	[22]	DL Models (CNN, LSTM, AE, Transformer)	SHAP / LIME	Multiple datasets	97.7% – 99.9%
7	[26]	CNN with LSTM and Auto encoder (Ensemble)	SHAP / LIME	ToN-IoT	98.50%
8	[27]	RF, DNN, MLP, LightGBM	SHAP	CICIDS2017, NSL-KDD	~ 95% – 99%
9	[17]	Ensemble Voting (DT, RF, SVM)	LIME / SHAP	CICIDS2017	~ 95.5% – 96.2%

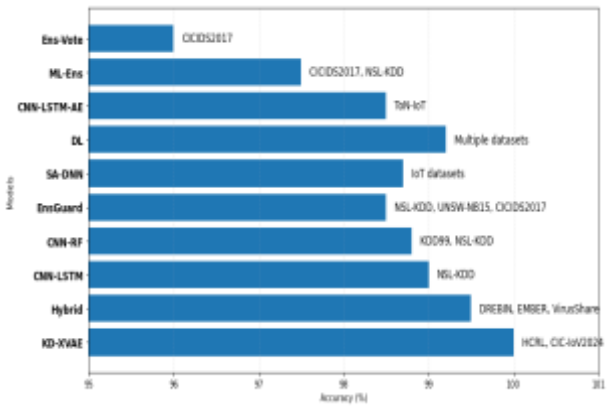

Fig. 16. Model Accuracy and Dataset Used

Fig. 16. Above: Illustrated comparison of IDS models based on detection accuracy. Each bar represents a model, while the annotations indicate the dataset used. The results show that hybrid and deep learning models such as KD-XVAE and CNN-LSTM achieve the highest accuracy, particularly when evaluated on benchmark datasets such as NSL-KDD and CICIDS2017.

E. Research Challenges

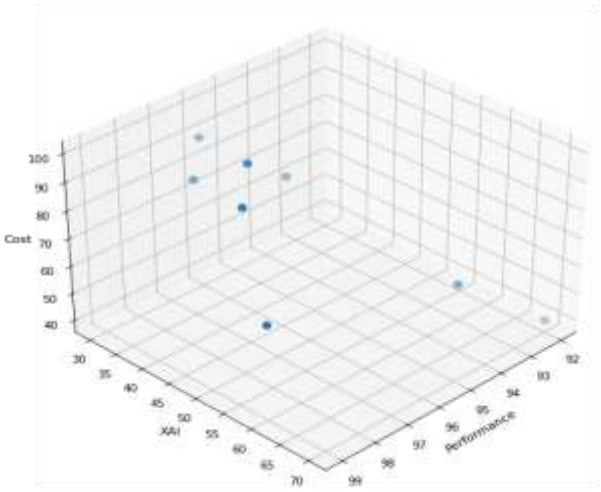

Fig.17. Trade-off Between Performance and Efficiency

Fig.17. The above illustrates a three-dimensional representation of the trade-off among performance, interpretability, and computational cost in AI-based intrusion detection systems. Each point represents a specific model positioned according to its detection accuracy (Performance), level of explainability (Interpretability), and computational requirements (Cost). The visualization shows that high-performance models, particularly deep learning and hybrid approaches, tend to achieve higher accuracy but at the expense of increased computational cost and reduced interpretability. In contrast, traditional machine learning models provide better interpretability and lower computational cost, but with relatively lower detection performance. This trade-off highlights a fundamental challenge in IDS design, where improving one dimension often leads to the degradation of others, emphasizing the need for balanced and unified approaches.

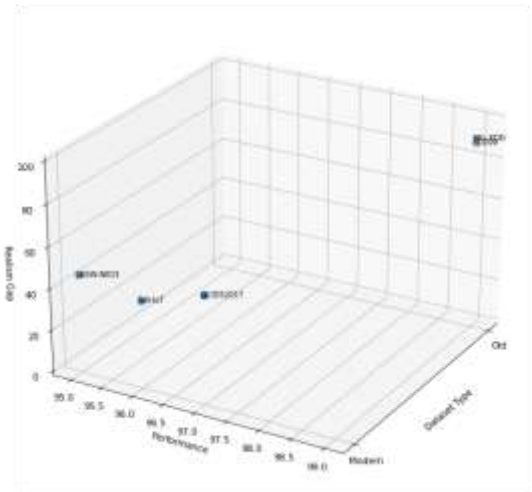

Fig.18. Relation Between Performance, Dataset, and Realism Gap

Fig.18. The above illustrates the relationship between dataset type, performance, and realism gap. It is evident that older benchmark datasets (e.g., NSL-KDD, KDD99) yield extremely high accuracy but are associated with a high realism gap. In contrast, modern datasets provide more realistic evaluation conditions, although with slightly lower performance.

Research Questions Analysis with Evidence from Literature

The findings indicate that IDS research is highly fragmented across models, datasets, explainability, and evaluation metrics, with no unified framework addressing all dimensions simultaneously, as shown in Table 12.

Table 12 Research Questions Analysis with Evidence from Literature

RQ	Domain	Key Answer (Literature Findings)
RQ1	IDS Models	ML, DL, and hybrid models dominate; CNN, LSTM, and DNN achieve the highest accuracy, but hybrid models are the most complex.
RQ2	Accuracy with Explain ability	High-accuracy models (DL/hybrid) lack inherent interpretability; explain ability is mostly post-hoc
RQ3	XAI Role	SHAP and LIME are the most widely used techniques, providing local and global explanations and improving transparency. However, they are computationally expensive, post-hoc in nature, and lack stability and real-time suitability.
RQ4	Datasets	NSL-KDD and CICIDS2017 dominate but are outdated or partially synthetic; limited real-world representation
RQ5	Evaluation Metrics	Mostly accuracy, F1, precision; lacks latency, efficiency, and standardized explain ability metrics
RQ6	Generalization	Models often trained on single datasets → weak cross-dataset generalization.
RQ7	Efficiency	DL and hybrid models are computationally expensive and unsuitable for real-time/IoT deployment.
RQ8	Unified Framework	No unified IDS framework exists integrating accuracy, explain ability, efficiency, and robustness.

The gained results addressed several critical limitations in the existing literature. First, a significant inconsistency is observed between model performance and interpretability outcomes, representing a persistent trade-off between predictive accuracy and explanation quality. Second, many papers fail to engage real-world datasets, instead relying on outdated or synthetic data, which limits the external validity and practical applicability of the findings. Third, there is an overdependence on legacy datasets, which may not adequately represent contemporary and evolving cybersecurity threats. Finally, the current techniques largely lack a well-defined unified framework in which explainability is inherently embedded within the learning process; rather, most methods depend on post-hoc explainability methods, where interpretability is applied after model training instead of being integrated as an intrinsic component of the model.

IV. OPEN PROBLEMS AND RESEARCH CHALLENGES

This section illustrates some steps related to the open problem and research challenges:

A. Limitations of Machine Learning and Deep Learning Models in IDS

Despite the strong performance achieved by machine learning and deep learning-based intrusion detection models, several inherent limitations persist across the literature. Traditional machine learning models such as Random Forest and SVM provide relatively respectable interpretability, but they struggle with high-dimensional and highly dynamic network traffic patterns [18], [29], [31]. On the other hand, deep learning models such as Convolutional Neural Network, Long Short-Term Memory, and Deep Neural Network achieve higher accuracy but suffer from significant drawbacks, including high computational cost, long training time, and lack of interpretability [16], [19], [23], [26], [27], [34]. Furthermore, ensemble and hybrid models, although capable of improving detection performance, introduce additional complexity and latency, making them less suitable for real-time deployment scenarios [8], [15], [30]. In addition, most models are highly dependent on training datasets and exhibit performance degradation when applied to unseen or cross-domain data, indicating limited generalization capability [20], [24], [27]. Gap: Current IDS models face a fundamental trade-off between accuracy, interpretability, computational efficiency, and generalization, with no unified architecture successfully addressing all these dimensions simultaneously.

B. Imbalance Between Accuracy and Explainability

The analysis of intrusion detection systems (IDS) shows that high detection performance is mainly achieved through deep learning models such as Convolutional Neural Network and Long Short-Term Memory, with several studies reporting accuracies exceeding 99% [14], [17], [23], [26]. However, the explainability component is predominantly handled using post-hoc methods such as SHAP and LIME [5], [7], [9], [12], [22], [28]. This separation between predictive performance and interpretability indicates that most high-accuracy models lack intrinsic transparency. Gap: There is a clear absence of IDS models that jointly achieve high accuracy and inherent interpretability.

C. Limited Integration of XAI with High-Performance Models

Although hybrid and deep learning-based IDS models demonstrate superior detection performance [4], [9], [15], [30], explainability techniques are usually applied as external post-processing tools rather than being integrated into the learning architecture. SHAP and LIME are commonly applied after model training without influencing the internal decision-making process [22], [25], [26]. This results in explanations that may not fully reflect the true internal reasoning of the model. Gap: There is a lack of tightly integrated frameworks combining deep learning models with built-in explainability mechanisms.

D. Dominance of Limited XAI Techniques

The literature review reveals a strong dominance of SHAP and LIME as the primary explainability techniques across IDS studies [5], [7], [8], [12], [22], [19], [30], [32]. In contrast, alternative methods such as gradient-based techniques, visualization methods (PDP, ICE), and system-level XAI approaches are rarely used [5], [12], [24], [26]. This over-reliance on a limited set of XAI methods restricts the diversity of interpretability perspectives. Gap: There is limited exploration and comparative evaluation of diverse XAI techniques beyond SHAP and LIME.

E. Over-reliance on Benchmark and Outdated Datasets

The dataset analysis shows that NSL-KDD and KDD Cup 99 remain widely used despite their age and limited representation of modern cyber threats [5], [6], [9], [15], [19], [27], [34]. Although newer datasets such as CICIDS2017 provide improved realism, they still suffer from issues such

as class imbalance and synthetic traffic generation [8], [10], [21], [28]. Gap: Existing IDS research heavily depends on outdated or partially synthetic datasets, limiting real-world applicability.

F. Weak Generalization Across Datasets

Most studies evaluate their models on a single dataset or a limited number of datasets without performing cross-dataset validation [7], [8], [17], [21]. This leads to models that perform well in controlled environments but fail to generalize effectively to unseen or real-world attack scenarios. Gap: There is insufficient cross-dataset evaluation, resulting in weak model generalization and poor robustness.

G. Incomplete Evaluation Metrics

The evaluation of IDS models is primarily based on traditional classification metrics such as accuracy, precision, recall, and F1-score [5], [7], [16], [19], [26]. However, critical aspects such as latency, computational cost, and explainability quality are rarely considered. Furthermore, explainability evaluation lacks standardized metrics across studies [22], [24], [28]. Gap: Current evaluation frameworks are insufficient to assess real-world IDS performance, efficiency, and explainability simultaneously.

H. High Computational Complexity of Advanced Models

High-performing IDS models, particularly deep learning and hybrid architectures, often require significant computational resources and long training times [9], [19], [23], [30]. This limits their deployment in real-time environments such as IoT networks and high-speed communication systems.

Gap: The high computational complexity of state-of-the-art models restricts their real-time and edge deployment.

I. Lack of Unified IDS Framework

The overall analysis indicates that IDS research remains fragmented across multiple dimensions, including model development, explainability techniques, dataset usage, and evaluation strategies [12], [22], [26], [1]. Each component is typically optimized independently without a unified framework that integrates detection accuracy, interpretability, efficiency, and real-world applicability. Gap: There is a lack of a unified IDS framework that jointly integrates performance, explainability, efficiency, and deployment readiness.

The summarized Table 13 below highlights that IDS research is fragmented across multiple dimensions, including model design, explainability, datasets, and evaluation metrics. While deep learning models achieve high detection accuracy, they suffer from interpretability and computational limitations. Similarly, explainability techniques are widely used but remain mostly post-hoc and lack integration into model architecture. Furthermore, the absence of standardized datasets and evaluation frameworks limits the reproducibility and real-world applicability of existing IDS solutions.

Table 11: Top 10 Highest Accuracy Studies in IDS

Section	Area	Models / Techniques	Descriptive	Class
A	Model limitations	ML, DL, Hybrid	Trade-off between accuracy, efficiency, and generalization	GAP
B	Accuracy with Explainability	CNN, LSTM + SHAP, LIME	High accuracy but low inherent interpretability	GAP
C	XAI integration	SHAP, LIME	XAI applied post-hoc, not embedded in models	GAP
D	XAI methods	SHAP, LIME, IG	Limited diversity of explainability techniques	GAP
E	Datasets	NSL-KDD, KDD99, CICIDS2017	Outdated and synthetic datasets	GAP
F	Generalization	Single-dataset models	Weak cross-dataset performance	GAP
G	Evaluation metrics	Accuracy, F1	Missing latency, cost, and explainability metrics	GAP
H	Complexity	DL, Hybrid models	High computational cost, not real-time	GAP
I	Framework	All IDS models	No unified IDS framework	GAP

V. CONCLUSION

particular focus on the interplay between detection performance, explainability, and computational efficiency. The analysis illustrated that deep learning and hybrid techniques achieve superior detection results, often above 99%, especially when evaluated on benchmark datasets. However, these techniques suffer from significant limitations, including high computational cost, lack of interpretability, and limited generalization across datasets. The paper further revealed that explainability in Intrusion Detection Systems is predominantly addressed using post-hoc methods such as SHAP and LIME, which, although effective, are not inherently integrated into the learning process. This results in a disconnect between model decisions and their explanations. Additionally, the heavy reliance on outdated or partially synthetic datasets, such as NSL-KDD and KDD99, raises concerns regarding the real-world applicability of current Intrusion Detection Systems techniques. Moreover, the evaluation of Intrusion Detection Systems models remains incomplete, as most papers focus on traditional performance metrics while neglecting critical factors such as latency, resource consumption, and explainability quality. This lack of comprehensive evaluation

frameworks limits the ability to assess Intrusion Detection Systems in real-world deployment scenarios. Based on these findings, several critical research challenges were identified, including the need for unified IDS frameworks that jointly optimize accuracy, interpretability, efficiency, and robustness. Future research should focus on developing intrinsically explainable models, incorporating diverse XAI techniques, utilizing realistic and evolving datasets, and establishing standardized evaluation metrics. Addressing these challenges is essential for advancing Intrusion Detection Systems toward practical, trustworthy, and real-time cybersecurity solutions.

REFERENCES

1. V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," Frontiers in Artificial Intelligence, vol. 8, p. 1526221, 2025.
2. M. I. J. Shekh Tareq Ali, Mahabub Alam Khan, Sheikh Md Kamrul Islam Rasel, Md Abdullah Al Nahid, Touhid Bhuiyan, "Enhancing Intrusion Detection Systems with AI: Examine the Integration of AI into Traditional IDS to Improve Detection Rates and Reduce False Positives," International Journal of Intelligent, vol. VOL. 12 NO. 22S (2024) no. 2147-6799, 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/7555?utm_source=chatgpt.com.
3. S. M. Yellepeddi, C. S. Ravi, V. K. R. Vangoor, and S. Chitta, "AI-powered intrusion detection systems: Real-world performance analysis," J AI-Assist Sci Discov, vol. 4, no. 1, pp. 279-289, 2024.
4. Y. Sanjalawe, S. Fraihat, S. Al-E'mari, and S. N. Makhadmeh, "A review of artificial intelligence-based intrusion detection in industrial internet of things," Discover Internet of Things, 2026.
5. M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, and A. Y. Zomaya, "An explainable deep learning-enabled intrusion detection framework in IoT networks," Information Sciences, vol. 639, p. 119000, 2023.
6. Y. Djenouri, A. Belhadi, G. Srivastava, J. C.-W. Lin, and A. Yazidi, "Interpretable intrusion detection for the next generation of Internet of Things," Computer Communications, vol. 203, pp. 192-198, 2023.
7. M. Hasnain, N. Javaid, A. K. J. Saudagar, and N. Kumar, "An intelligent and explainable intrusion detection framework for Internet of Sensor Things using generalizable optimized active Machine Learning," Journal of Network and Computer Applications, p. 104358, 2025.
8. A. Kwubeghari and N. G. Ezeji, "Designing an Explainable Intrusion Detection System (X-Ids) Using Machine Learning: A Framework for Transparency and Trust," ABUAD Journal of Engineering Research and Development (AJERD), vol. 8, no. 2, pp. 319-328, 2025.
9. S. Neupane et al., "Explainable intrusion detection systems (X-IDS): a survey of current methods, challenges, and opportunities. arXiv," arXiv preprint arXiv:2207.06236, 2022.
10. J. Ables et al., "Eclectic rule extraction for explainability of deep neural network-based intrusion detection systems," arXiv preprint arXiv:2401.10207, 2024.
11. D. J. K. Nkashama et al., "Deep learning for network anomaly detection under data contamination: evaluating robustness and mitigating performance degradation," in European Symposium on Research in Computer Security, 2024: Springer, pp. 68-87.
12. N. Khan, K. Ahmad, A. A. Tamimi, M. M. Alani, A. Bermak, and I. Khalil, "Explainable AI-based intrusion detection system for industry 5.0: an overview of the literature, associated challenges, the existing solutions, and potential research directions," arXiv preprint arXiv:2408.03335, 2024.
13. H. Manthena, S. Shajarian, J. Kimmell, M. Abdelsalam, S. Khorsandroo, and M. Gupta, "Explainable artificial intelligence (XAI) for malware analysis: A survey of techniques, applications, and open challenges," IEEE Access, 2025.
14. M. A. Yagiz, P. MohajerAnsari, M. D. Pesé, and P. Goktas, "Transforming in-vehicle network intrusion detection: VAE-based knowledge distillation meets explainable AI," in Proceedings of the Sixth Workshop on CPS&IoT Security and Privacy, 2024, pp. 93-103.
15. M. Adil, M. A. Jan, S. B. Hakim, H. H. Song, and Z. Jin, "xIDS-EnsembleGuard: An Explainable Ensemble Learning-based Intrusion Detection System," in 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2024: IEEE, pp. 93-100.
16. Z. Xu et al., "Deep learning-based intrusion detection systems: A survey," arXiv preprint arXiv:2504.07839, 2025.
17. M. M. J. Ayan et al., "Human-Centered Explainable AI for Security Enhancement: A Deep Intrusion Detection Framework," arXiv preprint arXiv:2602.13271, 2026.
18. W. Shao, "Interpretable Ensemble Learning for Network Traffic Anomaly Detection: A SHAP-based Explainable AI Framework for Embedded Systems Security," arXiv preprint arXiv:2603.28654, 2026.
19. Z. Li, W. Fang, C. Zhu, G. Song, and W. Zhang, "Toward deep learning-based intrusion detection system: A survey," in Proceedings of the 2024 6th International Conference on Big Data Engineering, 2024, pp. 25-32.
20. Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," ACM Computing Surveys, vol. 57, no. 1, pp. 1-42, 2024.
21. S. Patil et al., "Explainable artificial intelligence for intrusion detection system," Electronics, vol. 11, no. 19, p. 3079, 2022.
22. O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," IEEE Access, vol. 12, pp. 23954-23988, 2024.
23. E. C. P. Neto, S. Iqbal, S. Buffett, M. Sultana, and A. Taylor, "Deep learning for intrusion detection in emerging technologies: a comprehensive survey and

new perspectives," Artificial Intelligence Review, vol. 58, no. 11, p. 340, 2025.
24. O. Arreche and M. Abdallah, "A comparative analysis of DNN-based white-box explainable AI methods in network security," EURASIP Journal on Information Security, vol. 2025, no. 1, p. 16, 2025.
25. K. P. Sharma et al., "Interpretable intrusion detection for IoT environments using a self-attention-based explainable AI framework," Scientific Reports, vol. 15, no. 1, p. 39937, 2025.
26. T. B. Ogunseyi, G. Thiyagarajan, H. He, V. Bist, and Z. Du, "Performance Analysis of Explainable Deep Learning-Based Intrusion Detection Systems for IoT Networks: A Systematic Review," Sensors, vol. 26, no. 2, p. 363, 2026.
27. S. ISHTIAQ, F. U. REHMAN, S. SALEEM, and A. YAAR, "DEEP LEARNING FOR INTRUSION DETECTION SYSTEMS," Contemporary Journal of Social Science Review, vol. 3, no. 4, pp. 665-675, 2025.
28. N. Kalimuthu, "Explainable AI (XAI) for Enhanced Cyber Threat Intelligence: Building Interpretable Intrusion Detection Systems," IJSAT-International Journal on Science and Technology, vol. 16, no. 4, 2025.
29. M. Siganos et al., "Explainable AI-based intrusion detection in the internet of things," in Proceedings of the 18th International Conference on Availability, Reliability, and Security, 2023, pp. 1-10.
30. M. B. M. Shtayat, M. K. Hasan, R. Sulaiman, S. Islam, and A. U. R. Khan, "An explainable ensemble deep learning approach for intrusion detection in industrial internet of things," IEEE Access, vol. 11, pp. 115047-115061, 2023.
31. O. Arreche, T. Guntur, and M. Abdallah, "Xai-ids: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," Applied Sciences, vol. 14, no. 10, p. 4170, 2024.
32. A. I. Udofot, O. M. Oluseyi, and E. Bassey, "Explainable AI for cyber security. Improving transparency and trust in intrusion detection systems," International Journal of Advances in Engineering and Management, vol. 6, no. 12, pp. 229-240, 2024.
33. D. Gaspar, P. Silva, and C. Silva, "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron," IEEE Access, vol. 12, pp. 30164-30175, 2024.
34. Z. Abou El Houda, B. Brik, and L. Khoukhi, "Why should I trust your IDs?": An explainable deep learning framework for intrusion detection systems in Internet of Things networks," IEEE Open Journal of the Communications Society, vol. 3, pp. 1164-1176, 2022.
35. Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. Vol. 61, No. 10, pp. 36-43, Oct 1 2018, doi: 10.1145/3233231.
36. P. Zschech, S. Weinzierl, and M. Kraus, "Inherently Interpretable Machine Learning: A Contrasting Paradigm to Post-hoc Explainable AI: P. Zschech et al," Business & Information Systems Engineering, pp. 1-19, 2025.
37. T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," arXiv preprint arXiv:1907.09294, 2019.
38. I. Lage et al., "An evaluation of the human-interpretability of explanation," arXiv preprint arXiv:1902.00006, 2019.
39. T. Qamar and N. Z. Bawany, "Understanding the black-box: towards interpretable and reliable deep learning models," PeerJ Computer Science, vol. 9, p. e1629, 2023.
40. R. Hassan, N. Nguyen, S. R. Finserås, L. Adde, I. Strümke, and R. Støen, "Unlocking the black box: Enhancing human-AI collaboration in high-stakes healthcare scenarios through explainable AI," Technological Forecasting and Social Change, vol. 219, p. 124265, 2025.
41. L. Gao and L. Guan, "Interpretability of machine learning: Recent advances and prospects," IEEE MultiMedia, vol. 30, no. 4, pp. 105-118, 2023.
42. J. C. Bjerring, J. Mainz, and L. Munch, "Deep learning models and the limits of explainable artificial intelligence," Asian Journal of Philosophy, vol. 4, no. 1, p. 22, 2025.
43. H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," Applied Sciences, vol. 9, no. 20, p. 4396, 2019.
44. A. Devarakonda, N. Sharma, P. Saha, and S. Ramya, "Network intrusion detection: A comparative study of four classifiers using the NSL-KDD and KDD’99 datasets," in Journal of Physics: Conference Series, 2022, vol. 2161, no. 1: IOP Publishing, p. 012043.
45. H. M. R. U. Rehman et al., "A systematic literature study of machine learning techniques based intrusion detection: datasets, models, challenges, and future directions," Journal of Big Data, vol. 12, no. 1, p. 264, 2025.
46. S. A. AL-Shami and M. F. Abdullah, "Deep Learning for DDoS Detection: A Systematic Review of Explainability and Adversarial Robustness."