DOI: https://doi.org/10.26565/2304-6201-2025-66-02

УДК (UDC) 004.056.53:004.89

Hleha Kateryna *master student; Institute of Special Communications and Information Protection of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Verkhnyoklyuchova, 4, Kyiv, Ukraine, 03056*
e-mail: katerynaglega54@gmail.com;
https://orcid.org/0009-0004-9337-5836

Hol Vladyslav *professor; head of department; Institute of Special Communications and Information Protection of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Verkhnyoklyuchova, 4, Kyiv, Ukraine, 03056*
e-mail: vladgol1971@gmail.com;
https://orcid.org/0000-0002-9995-9590

XAI Optimization for Low-Latency Neural-Based Intrusion Detection Systems in Network Environments

Relevance. In contemporary network environments, deep learning-based intrusion detection systems (IDS) provide significant improvements in detecting complex and evolving cyber threats. However, their practical deployment in real-time applications is severely limited by computational complexity, latency, and a lack of interpretability, commonly referred to as the "black-box" problem. Integrating eXplainable Artificial Intelligence (XAI) methods into IDS is crucial for enhancing the transparency, trustworthiness, and operational effectiveness of security systems. **Goal**. The aim of this research is to explore and optimize XAI methods to achieve low-latency, explainable neural-based intrusion detection systems suitable for real-time network traffic analysis, thus balancing interpretability with computational efficiency and detection accuracy. **Research methods.** The study conducted a systematic review and comparative analysis of existing deep learning (DL) models (CNN, LSTM, GRU, Autoencoders, CNN-LSTM hybrids) and prominent XAI techniques (SHAP, LIME, Integrated Gradients, DeepLIFT, Grad-CAM, Anchors). Optimization strategies were proposed, including hardware acceleration, lightweight gradient-based attribution methods, hybrid architectures, and selective explanation strategies. Empirical validation was performed on standard datasets (CICIDS2017, NSL-KDD, UNSW-NB15). **The results.** The analysis revealed that gradient-based attribution methods (DeepLIFT, Integrated Gradients) are optimal for real-time IDS due to minimal latency and high fidelity. Hybrid explainable-by-design frameworks, specifically CNN-LSTM models enhanced with attention mechanisms (ELAI framework), demonstrated significant performance gains with detection accuracy exceeding 98% and inference times below 10 ms. Optimized methods notably improved zero-day attack detection rates up to 91.6%. **Conclusions**. The research successfully demonstrated practical methods for integrating explainability into real-time neural-based IDS, significantly enhancing both detection performance and decision transparency. Future research should focus on standardizing evaluation metrics, refining attention-based models, and extending these optimization approaches to other cybersecurity applications.

Keywords: *cybersecurity, intrusion detection system, deep learning, explainable artificial intelligence, real-time detection, anomaly detection, neural networks, XAI optimization.*

How to quote: Hleha K., Hol V., "XAI Optimization for Low-Latency Neural-Based Intrusion Detection Systems in Network Environments", *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 66, pp. 19-36, 2025. https://doi.org/10.26565/2304-6201-2025-66-02

Як цитувати: Hleha K., Hol V., XAI Optimization for Low-Latency Neural-Based Intrusion Detection Systems in Network Environments. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 66. С.19-36. https://doi.org/10.26565/2304-6201-2025-66-02

1. Introduction

In the digital age, cybersecurity – the practice of protecting systems, networks, and confidential data that is coursing through them from unauthorized access, damage, and hijacking – has become an essential part of any organization's security policy. As time progresses, the reliance of individuals, organizations, and governments on digital infrastructure increases. Consequently, it is crucial to ensure the confidentiality, integrity, and availability (CIA triad) of data.

To safeguard sensitive information from a rapidly expanding array of cyberattacks, including the most recent “zero-day” and “slow” attacks, it is imperative that each organization establish a multilayered cybersecurity system, which includes intrusion detection systems (IDS). An intrusion detection system (IDS) scans for malicious activity or unauthorized access through analysis of traffic dynamics, applications and session behavior, and signature-based features. By emphasizing anomalies and recognized attack patterns, IDS alerts help organizations promptly address possible security risks [1].

As the volume of network traffic increases and cyberattacks become more sophisticated, conventional IDS systems, including signature-based and anomaly-based mechanisms, become less effective in detecting intrusions. Currently, approximately 50% of startups disclose experiencing information theft, underscoring the urgency of robust IDS solutions in institutional security frameworks [2].

The integration of machine learning (ML) and deep learning (DL) technologies with IDS systems allows us to significantly enhance the efficiency and classification accuracy of security systems. By utilizing a wide range of ML and DL models with varying characteristics, cybersecurity personnel can create IDS systems that are unique in their ability to perform specific tasks and guarantee the highest level of data protection for each distinct network.

In order to determine whether network traffic is normal or demonstrates indicators of potential malicious activity, machine learning techniques employ algorithms such as decision trees, K-nearest neighbors, and support vector machines (SVMs). These methods significantly enhanced detection rates in comparison to conventional solutions upon implementation; however, ML-based IDS systems were still incapable of managing high-dimensional and imbalanced data. Their reliance on centralized data storage and transmission has resulted in substantial privacy and security vulnerabilities. Two additional challenges to the development of effective ML solutions are the vast quantity of network information and the prevalence of imbalanced data sets. Minor but critical attack types are underrepresented and simply insufficient for proper training. While there are numerous preprocessing techniques, feature selection methods, and ensemble strategies available to improve performance, they are insufficient to guarantee highly accurate detection when it comes to capturing the complex patterns and relationships present in network traffic data. Therefore, researchers initiated the development of more complicated deep learning algorithms, which have demonstrated an exceptional ability to learn hierarchical representations from high-dimensional data [3, 4].

Artificial neural networks have recently garnered significant attention for their potential to improve IDS systems. For instance, convolutional neural networks (CNNs) performed exceptionally well in the identification of spatial features – correlations and relationships within data at a specific time. Conversely, recurrent neural networks (RNNs) had proven to be more adept at detecting temporal features – patterns and sequences across data over time. Several studies have illustrated the successful implementation of deep learning techniques in network intrusion detection, either independently or as part of ML/DL hybrid systems. More recently, researchers have examined the use of advanced deep learning architectures, including long short-term memory (LSTM) and gated recurrent unit (GRU) networks, for intrusion detection in network traffic data [5]. These models demonstrated a high level of ability to capture temporal dependencies in sequential data, which is particularly important when examining network traffic.

Despite their numerous benefits, the DL technologies have significant issues that must be resolved before they can be widely implemented in IDS systems. The deployment of complex deep learning models, such as LSTM, is restricted in real-time or high-throughput environments, such as backbone networks, due to their resource-intensive and slow nature. In addition, they may be highly vulnerable to adversarial inputs, be challenging to scale, or require the implementation of specialized techniques to identify rare but critical attack patterns. DL IDS systems’ biggest shortcomings, however, are their high latency, low throughput, and inability to be explained [6].

Processing large-scale traffic flows, particularly in real-time operations, is restricted by the computational complexity and high resource consumption of DL architectures. DL-based IDS systems are frequently treated as “black boxes” by both developers and users due to their inability to clarify their inference processes and final results. This lack of transparency hinders forensic analysis, complicates auditing and compliance processes, and reduces the overall trust in automated security decisions made by DL models.

eXplainable Artificial Intelligence (XAI) approaches can be integrated into the DL-based IDS frameworks to enhance transparency and interpretability and to facilitate a more comprehensive understanding of model decisions. Transparency helps build trust in AI-driven frameworks by explaining the logic behind some outcomes, which is essential for meeting legal and regulatory requirements [7].

However, despite XAI technologies offering a promising solution to the explainability problem, the most popular methods, such as SHAP and LIME [8], are computationally expensive and not well-suited for real-time deployment. Their reliance on repeated model evaluations or surrogate approximations significantly increases latency, making them impractical for high-throughput environments where fast decision-making is critical.

This research attempts to come up with a proper solution for integrating explainable AI techniques into deep learning-based IDS models in a way that preserves low latency and high throughput while maintaining sufficient interpretability for real-time security decision-making.

2. Objective of the study and research tasks

The primary objective of this study is to explore and evaluate optimization strategies for integrating explainable artificial intelligence (XAI) into deep learning-based intrusion detection systems (IDS) operating in real-time network environments. The goal is to balance detection performance, computational efficiency, and interpretability to enhance trust and operational usability.

To achieve this objective, the following research tasks are defined:

- to review existing explainable AI methods (e.g., LIME, SHAP) and analyze their applicability to IDS;
- to identify the main challenges of implementing XAI in low-latency, high-throughput intrusion detection systems;
- to compare traditional (offline) and real-time XAI approaches in terms of performance, scalability, and explainability;
- to investigate potential optimization strategies for deploying XAI in real-time DL-based IDS;
- to propose conceptual guidelines for integrating interpretable components into deep IDS models without compromising detection speed and accuracy.

3. Review of existing DL models and XAI methods suitable for real-time IDS systems

3.1. Deep learning models for real-time intrusion detection

DL technologies have the potential to significantly improve intrusion detection systems by removing the primary limitations of traditional methods. In contrast to signature-based IDS, which fail to identify emerging threats, deep neural networks automatically identify intricate patterns from raw network data, thereby capturing non-linear feature relationships without the need for manual feature engineering. This allows DL-based IDS systems to detect both known attack signatures and previously unseen or evolving attack patterns with greater accuracy and adaptability. Simply put, deep learning improves on previous methods' high false alarms and blind spots for novel attacks by enabling more precise, flexible, and comprehensive threat detection in IDS [4].

In order to effectively prevent the intrusion, it is crucial to immediately identify any potential anomalies and unusual behavioral changes in the network traffic. As a result, the faster the IDS system operates, the greater is the chance of stopping an attack before it fully corrupts the network. The most effective approach for the majority of contemporary networks is to implement real-time IDS systems that can immediately process traffic and identify changes in the present.

Deploying an IDS in real-time operational networks imposes strict requirements on both the system and the DL models used in it. Key demands include the following:

1. Low detection latency. The IDS system must analyze traffic and detect intrusions with minimal delay (near-instantaneously) to prevent or contain attacks as they occur. Real-time network applications require ultra-low latency processing; even minor delays in traffic analysis can degrade an IDS’s effectiveness [9].

2. High throughput and scalability. It is critical that the real-time IDS system be able to handle large volumes of continuous network data (high bandwidth traffic) without becoming a bottleneck. This implies that the detection DL model must be capable of scaling to high-speed networks and large data streams, processing events in milliseconds, and maintaining a pace with network line rates. As networks expand, the IDS system must ensure that it operates efficiently in heterogeneous or distributed environments.

3. Computational efficiency. To operate in real time, the algorithms must be resource-efficient. Deep models with extremely high complexity (e.g., very deep CNNs or LSTMs) can require a large amount of computation time and may be too heavy for real-time use on limited hardware. Real-time IDS systems often require optimizing or simplifying their models (or utilizing hardware acceleration) to satisfy time

constraints. In particular, in IoT or edge scenarios, DL models must operate within limited CPU/memory, which is why lightweight or optimized models are preferred.

Table 1. Comparison of deep learning models for intrusion detection in real time [11, 12]
Таблиця 1. Порівняння моделей глибокого навчання для виявлення вторгнень у реальному часі [11, 12]

Model	Key Features	Strengths	Limitations	Best Use Cases
Convolutional Neural Networks (CNN)	Employs convolutional layers to extract spatial patterns from fixed-length input vectors	Fast inference; highly parallelizable; excellent at detecting known structured attack patterns; fast training	Not suitable for time-series data or sequences	Packet/flow-level intrusion detection in high-throughput environments
Recurrent Neural Networks (RNN)	Processes sequential data with memory connections	Effective for detecting sequential attack behavior	Suffers from vanishing gradients; less stable	Network traffic behavior analysis
Long Short-Term Memory (LSTM)	Enhanced RNN with long-term memory capability	Handles long-term dependencies; high detection rate for evolving threats	Computationally intensive; slower training	Detection of persistent threats and time-based anomalies
Gated Recurrent Unit (GRU)	Lightweight recurrent architecture that captures temporal dependencies using update and reset gates	Faster, consumes fewer resources than LSTM, yet adapts well to sequence patterns	Slightly less capable of modeling long-term dependencies than LSTM	Detecting time-series anomalies, slow scans, and low-and-slow attacks in real time
Autoencoder	Unsupervised neural network to reconstruct normal behavior; anomalies result in high reconstruction loss	Detects zero-day threats without labeled data; suitable for anomaly-based detection	May misclassify if trained on noisy data; slower unless specifically optimized	Zero-day attack detection and anomaly-based IDS systems
Hybrid Lightweight 1D CNN-LSTM	Combines spatial feature extraction of CNN with temporal pattern detection from LSTM	Balances speed and accuracy; optimized variants can run in real time	Requires careful optimization; heavier than purely CNN or GRU	Attacks exhibiting both spatial and temporal characteristics, such as DDoS or multi-stage intrusions
Deep Neural Networks (DNN)	Simple fully connected feedforward networks	Very fast inference; easy to implement and scale; low latency	Limited feature extraction capability; may miss complex patterns	General classification tasks in high-throughput IDS pipelines

4. High detection accuracy. Even under speed constraints, a real-time IDS system is expected to accurately distinguish attacks from normal traffic. Reliability is crucial – high true positive rates and low false positives ensure the system's rapid alerts are trustworthy. As a result, the DL model should strike a

balance between speed and accuracy, giving operators accurate and timely detection results without overloading them with false alarms.

In order to meet these requirements, researchers implement streaming architectures and optimizations. For example, integrating DL models into frameworks such as Apache Spark Streaming or using a unified Kappa architecture [10] can enable continuous, low-latency processing of network data. In practice, real-time IDS performance may be achieved through the use of model compression, parallel processing, or ensemble methods that enhance accuracy while maintaining a millisecond time budget.

While numerous DL models have been integrated into IDS systems, only a small number are particularly well-suited for real-time detection due to their capacity to balance efficiency and speed. The main DL models and their attributes in an IDS context are summarized in Table 1 [12].

For real-time deployment, DL models must balance speed, throughput, and precision with operational constraints.

CNNs offer rapid, parallelizable inference suitable for high-throughput detection of structured attack patterns, making them highly appropriate for real-time IDS, though lacking in temporal analysis. LSTMs excel at modeling temporal attack sequences, like slow-moving threats, but require considerable computational resources, limiting their real-time practicality unless optimized [11]. GRUs provide a computationally efficient alternative to LSTMs, capturing temporal dependencies effectively with lower latency, thus being more suitable for real-time IDS. CNN-LSTM hybrids achieve an optimal balance of spatial and temporal pattern recognition, delivering high accuracy and real-time deployment feasibility with minimal latency [12]. Autoencoders, capable of unsupervised anomaly detection, are beneficial in real-time IDS for identifying zero-day threats but may generate false alarms and lack detailed attack classification [11]. DNNs, although limited in complex feature detection, offer near-instantaneous inference suitable for initial screening in ultra-high-speed IDS pipelines [12].

The comparative analysis of DL models upon their integration into IDS systems is shown in Table 2 [11]. There are advantages and disadvantages to each deep learning model in terms of complexity, accuracy, and speed. The most appropriate option frequently is determined depending on the deployment constraints and the prioritized attack characteristics that are intended to be countered. CNN-LSTM hybrid DL-based IDS have been demonstrated to enhance detection rates while maintaining real-time operation by employing streaming-friendly architectures and carefully balancing the workload.

Table 2. Comparative analysis of deep learning models for intrusion detection [11]
Таблиця 2. Порівняльний аналіз моделей глибокого навчання для виявлення вторгнень [11]

Model	Accuracy (%)	Precision	Recall	Dataset Used
CNN	96,8	0,95	0,94	CICIDS2017
RNN	95,2	0,93	0,91	NSL-KDD
LSTM	97,1	0,96	0,95	BoT-IoT
Autoencoder	94,5	0,91	0,90	N-BaIoT
Hybrid CNN-LSTM	98,3	0,97	0,96	Custom Mixed Dataset

All the DL models mentioned have the potential to be integrated into real-time IDS systems, contingent upon the identification of the priorities and necessary characteristics of the systems. Nevertheless, some optimization may be necessary before that.

3.2. Explainable AI methods and their potential to enhance trust in DL-based IDS systems

Deep learning models pose one of the greatest challenges in deploying them in actual IDS use due to a lack of interpretability. IDS have to rely on deep learning algorithms that lack transparency despite their high accuracy, creating a “black box” effect that can hinder the analysts’ understanding of their decision-making processes. Simply put, despite the high accuracy of detection, these systems provide little to no insight into why certain decisions were made.

“Black-box” status of DL models means that security professionals struggle to understand the reasoning behind alerts, which is important to trust the system and respond appropriately. Uninterpretable IDS can lead to high false-alarm rates and missed threat patterns, since security teams cannot easily verify or refine the model’s decisions.

Explainable AI (XAI) addresses this challenge by making IDS decisions more auditable. XAI is a fast-growing area of research with the goal to enhance the transparency and trustworthiness of AI systems. For IDS, XAI methods are being applied to yield:

1. Decision-making process visualizations, which can assist security analysts in determining how an IDS model reached a specific decision.

2. Feature importance analysis, which identifies which features (e.g., packet size, traffic volume) were most important for the model's prediction.

3. Interpretability models, e.g., decision trees or rule-based systems, that can provide explanations in a human-readable format [13, 14].

By providing human-understandable explanations for each detection, XAI enables analysts to see which features or behaviors influenced an alert, thereby enhancing trust and clarity in decision-making. For example, an XAI-enhanced IDS might show that an unusually high volume of traffic on a rare port was the key reason a session was classified as an attack. Such insights allow security teams to validate alerts, reduce false positives, and confidently act on the system's recommendations.

A range of XAI methodologies has been applied within IDS frameworks to enhance transparency, each offering distinct advantages and facing unique challenges. Among these, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are widely regarded as effective post-hoc, model-agnostic techniques for elucidating complex models like deep neural networks [15].

Table 3. Comparison of XAI methods for real-time IDS suitability [15, 16, 18, 20, 21, 22]
Таблиця 3. Порівняння методів XAI щодо придатності для IDS реального часу [15, 16, 18, 20, 21, 22]

XAI Method	Explanation Type	Model Agnostic/ Specific	Key Advantages	Limitations	Suitability for Real-Time IDS
Shapley Additive Explanations (SHAP)	Feature Attribution	Model-agnostic	Consistent and fair attribution; local and global explanations	High computational overhead	Limited due to high latency
Local Interpretable Model-Agnostic Explanations (LIME)	Surrogate Model	Model-agnostic	Intuitive, per-instance explanations	Local explanations may not generalize well	Limited due to latency concerns
Saliency Maps / Grad-CAM	Gradient-based Visualization	Model-specific	Fast, intuitive visual explanations	Require differentiable models; noisy outputs; no textual explanations	Suitable due to low latency
Integrated Gradients (IG)	Gradient-based Attribution	Model-specific	Robust feature attribution, faithful explanations	Requires model internals; less intuitive alone	Highly suitable due to low latency
DeepLIFT	Gradient-based Attribution	Model-specific	High-fidelity, efficient, low complexity explanations	Requires model internals; less intuitive alone	Highly suitable due to low latency
Anchors	Rule-based Explanations	Model-agnostic	Intuitive, high-precision rules	Computationally intensive to derive optimal rules	Limited unless simplified rules are used

SHAP provides consistent and thorough feature attribution but its computational complexity limits its suitability for real-time IDS scenarios [15, 16]. LIME offers intuitive per-instance explanations beneficial for auditing individual alerts but is similarly constrained by computational latency [15, 18]. Saliency Maps and Grad-CAM are fast visual methods suited for real-time use due to their low computational overhead [20]. Integrated Gradients (IG) and DeepLIFT efficiently deliver faithful, low-latency explanations highly suitable for real-time IDS implementations. They are precise and are applicable to any differentiable DL-based IDS, but they require access to the neural model's internals [21]. Anchors produce intuitive, high-precision rules but their computational cost in deriving optimal rules limits real-time applicability unless simplified rules are used [22].

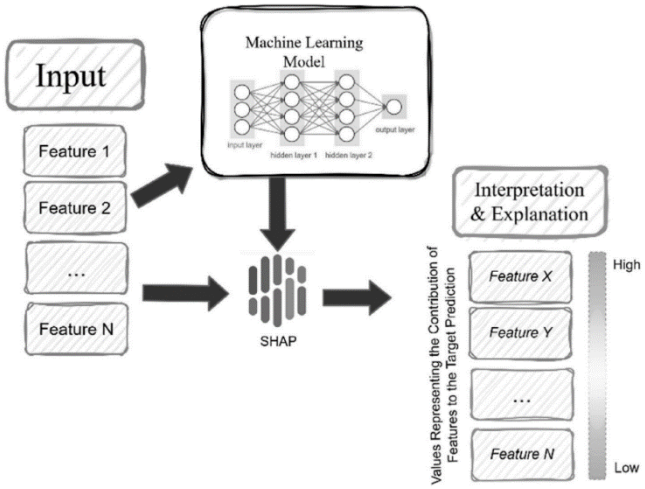

Fig. 1. Schematic representation of the SHAP value framework in a machine learning context [17]
Рис. 1. Схематичне представлення структури фреймворку SHAP у контексті машинного навчання [17]

When integrating XAI methods into a real-time IDS, several key criteria determine their suitability:

1. Fidelity (Faithfulness). This refers to how accurately the explanation reflects the actual decision process of the original model. A high-fidelity explanation will highlight the truly important features in the model's internal reasoning. For example, if the IDS's DL model bases its decision mainly on feature X, a faithful XAI method should assign the highest importance to X in its explanation. Low-fidelity explanations can mislead analysts by emphasizing the wrong factors, so measuring an XAI technique's fidelity to the DL model is crucial. Some studies quantify fidelity by checking how model predictions change when the top-ranked features from the explanation are removed or perturbed. In an IDS context, faithfulness ensures the explanations are truthful proxies of the complex model – a necessary condition for trust.

2. Human Interpretability. Even if an explanation is faithful, it must be understandable to a human analyst. Interpretability involves the simplicity and clarity of the explanation – e.g., using a small number of features, natural language descriptions, or visual aids that a person can quickly grasp. A highly interpretable explanation might be a short rule or a concise list of the top 2 or 3 features influencing a decision, rather than a dense list of 20 parameters. For IDS analysts under time pressure, explanations should ideally be simple, clear, and domain-relevant. Techniques like Anchors and decision trees score well on human interpretability, since they produce "if-then" rules and clear decision paths, respectively, whereas something like a raw saliency map or a long list of Shapley values may need more interpretation [22]. In practice, usability studies have found that providing transparent and visual explanations, such as charts of feature contributions or highlighted traffic traces, improves analysts' trust and speeds up their validation of alerts.

3. Computational Overhead. The extra processing time and resources required to generate explanations are a major concern for real-time systems. Some XAI methods, especially post-hoc techniques, can be computationally intensive. For instance, SHAP often requires evaluating the model multiple times on various feature subsets or background samples, and LIME needs to generate many perturbed samples to fit a local surrogate. These methods can significantly slow down the alert pipeline if used on every single event. Overhead is typically measured in terms of added latency per prediction or

CPU/GPU usage. In a high-throughput IDS, an XAI method with heavy overhead might not be practical. Therefore, a suitable XAI for real-time IDS should minimize computation – using efficient algorithms, sampling strategies, or by leveraging hardware acceleration. Research in explainable IDS frequently emphasizes the need for lightweight explainability approaches that don't degrade the system's performance.

4. Real-Time Feasibility. This criterion is related to overhead but focuses on whether the XAI method can deliver explanations within the time constraints of an operational environment. A real-time IDS may need to flag and explain an alert within milliseconds to seconds. Thus, methods that require lengthy processing or cannot keep up with streaming data are less feasible. Real-time feasibility also considers if the explanation can be generated on-the-fly for each alert or if it requires batch or offline processing. For example, a method that pre-trains an interpretable surrogate model might be feasible if that surrogate can then produce instant explanations during operation [15]. In contrast, an approach that must solve an optimization or search problem per event – as some anchor-based or perturbation methods do – might struggle in real time. Ultimately, achieving real-time explainability often involves a trade-off between the depth of the explanation and the speed of generation. Ensuring feasibility might involve simplifying the explanation, using approximate but faster algorithms, or only explaining a subset of events rather than everything.

Not all explanation techniques are equally practical for a real-time DL-based IDS. Techniques that offer low latency and high clarity are generally more suitable.

1. Gradient-Based Attributions methods – Saliency, Integrated Gradients, and DeepLIFT – are typically favored for real-time use because of their computational efficiency. They piggyback on the model's own backpropagation, typically requiring only one pass through the neural network to calculate feature importances. For instance, computing an integrated gradient or a DeepLIFT attribution for an input can be done quickly on modern hardware [21].

In comparative evaluations on an LSTM-based IDS, gradient methods (especially DeepLIFT) produced explanations with lower complexity and higher fidelity than LIME or SHAP, indicating they capture the model's behavior well without excessive computation. DeepLIFT in particular was found to give consistent and reliable explanations while being faster to compute, making it a strong candidate for real-time alert explanation.

Gradient-based attributions methods work seamlessly with common DL models such as CNNs, RNNs and autoencoders, highlighting important features or time steps almost instantaneously. The trade-off is that gradient-based explanations might be less intuitive in isolation, but they can be combined with visualization or simple messaging to aid analysts. Overall, because of their high fidelity and low overhead, saliency and gradient techniques are well-suited to explain decisions on the fly in real-time IDS systems.

2. Surrogate and Rule-Based Methods (Simplified Models) present another strategy for real-time explainability, namely to use an interpretable model alongside or in place of the DL for certain decisions. For example, a decision tree or a set of "if-then" rules can approximate the deep model's behavior for explanation purposes. These surrogates can be pre-computed (offline) to mimic the DL model on training data then used to generate quick explanations during operation. Because a decision tree or rule set is fast to evaluate, the explanation is essentially instantaneous at run-time. Such hybrid approaches attempt to get the best of both worlds: the DL handles detection accuracy, while the surrogate offers human-readable logic as explanations [15].

However, maintaining fidelity is a challenge – a too-simple surrogate might not capture complex patterns the DL model uses. In real-time settings, one compromise is to deploy the interpretable model for the majority of routine traffic and reserve the complex DL model with post-hoc explanations like LIME/SHAP for more ambiguous or high-risk cases. This tiered approach can preserve performance and provide transparency when most needed, at the cost of system complexity.

When using rule-based explainers like Anchors in real-time, scope is important: anchors can be computed quickly if the feature space is small or if we only seek a rule for the most influential features. They can succinctly explain an alert (e.g., "alert triggered because X and Y conditions were met") without overwhelming detail, which is ideal for an analyst's quick decision cycle. The user must be cautious that anchors – or any rule – remain accurate under evolving traffic conditions.

3. While LIME and SHAP are powerful and widely used, their direct application to every packet or alert in a high-throughput IDS can be impractical due to their computational cost. SHAP, in particular, though providing very insightful explanations, might take too long on complex models or large feature

sets – potentially seconds per instance – which is untenable for systems that analyze hundreds of events per second.

That said, there are scenarios where these methods can still contribute: for instance, TreeSHAP can efficiently explain ensemble tree models – if an IDS uses a tree-based classifier – in real-time by leveraging a closed-form solution. LIME can be sped up by reducing the number of perturbations or using optimized surrogates, but it may still struggle as data dimensionality grows [15].

Therefore, in a real-time IDS, SHAP/LIME are often used selectively – for example, to explain a handful of critical alerts or to perform periodic analysis on model behavior – rather than on every event. They are extremely valuable in offline model evaluation or forensic analysis of incidents, helping to understand global patterns (SHAP) and specific cases (LIME) with high interpretive richness.

The integration of XAI into IDS is critical for ensuring that cybersecurity systems are not only highly accurate but also capable of providing explanations that human analysts can readily comprehend and act upon. In summary, for the day-to-day, fast-paced detection, lighter methods such as gradients and simple rules are preferable, whereas LIME or SHAP might support near-real-time workflows where a brief delay is acceptable or as backup explainers for complex cases.

Integrating XAI into real-time DL-based IDS systems requires balancing explanation quality with performance. Methods like integrated gradients, DeepLIFT, and saliency maps offer quick, faithful insights into neural models' decisions and are thus most promising for real-time IDS use. Rule-based explanations and simplified surrogates provide human-friendly logic with negligible latency, which can greatly aid analyst understanding when carefully aligned with the DL model. More computationally intensive techniques like SHAP and LIME are effective in enhancing transparency and reducing false positives, but they may need optimization or selective deployment to fit into high-speed environments.

3.3. Challenges of integrating XAI into real-time DL-based IDS systems

Real-time IDS systems face several obstacles when incorporating XAI methods. Key challenges include computational overhead, scalability issues, accuracy-interpretability trade-offs, lack of standard evaluation practices, and security implications of exposing model logic.

1. Latency and computational overhead is the first problem. As it had already been mentioned, many popular XAI techniques, such as SHAP and LIME, are computationally intensive, often requiring numerous model re-runs or complex calculations. In a real-time IDS, generating an explanation for each alert can introduce significant latency and CPU/GPU load. Studies confirm that post-hoc explainers like LIME/SHAP add extra processing, which can slow down threat detection and response rates. In other words, the IDS may become sluggish in high-speed networks because of the time spent computing explanations. One survey notes that XAI-enhanced IDS often face "increased computational complexity and potentially reduced performance due to the overhead of generating explanations [23]." Such latency overhead is problematic in operational environments that demand swift decision-making to block attacks.

2. Scalability and deployment constraints are the second major issue. The heavy resource requirements of both DL models and XAI methods pose scalability issues. Many advanced DL-based IDS models, for example, transformers or deep CNNs, need powerful hardware acceleration, which is unsuitable for edge deployments with limited resources [23]. Pushing complex models or their explainers to low-power network devices can be infeasible due to memory, CPU, or energy constraints. Additionally, high-throughput network traffic magnifies the problem – explaining every flagged event in a busy network can overwhelm the system. Even cloud-based IDS setups struggle, as constant communication for explanations adds network latency. Researchers highlight that real-time IDS performance suffers in high-traffic environments when burdened with current XAI computations. In summary, without careful optimization, XAI may not scale well to the volume and speed of data in modern networks.

3. The next concern of real-time DL-based IDS is to find balance between accuracy and interpretability. There is an inherent trade-off between model complexity, which often yields higher accuracy, and its interpretability. State-of-the-art IDS models like DNN or ensemble methods achieve superior detection rates but operate as "black boxes" with opaque logic. By contrast, simpler models like decision trees or rule-based classifiers are transparent but tend to miss subtle or sophisticated attacks. This gap is well documented – high-performing DL models regularly forgo interpretability for greater predictive power, whereas overly simple models can undermine detection accuracy. In practice, forcing a complex model to be more explainable, for example, by approximating it with an interpretable surrogate, may degrade its performance on edge-case intrusions. Studies have noted that decision-tree-based IDS, while easy to explain, "frequently miss subtle danger behaviors, which lowers the accuracy

of detection [24].” Balancing these concerns is difficult: analysts need to trust and understand the IDS decisions, but not at the cost of allowing attacks to slip through due to an oversimplified model.

4. Another significant challenge is the lack of standardized XAI evaluation. There is no consensus on how to evaluate and compare XAI methods in the IDS domain. Unlike accuracy or false-alarm rate, which have clear metrics, “explainability” lacks a unified quantitative framework in cybersecurity contexts. Researchers point out that without standard interpretability metrics, it is difficult to judge whether one explanation method truly outperforms another or adequately meets analysts’ needs. This gap means each study often uses its criteria (e.g., subjective user feedback or ad-hoc measures of explanation quality), making it difficult to benchmark XAI techniques across different IDS implementations. The literature emphasizes that consistent evaluation standards – such as agreed-upon interpretability scores or time-to-insight measurements – are needed to fairly assess XAI in IDS [15]. Until such frameworks mature, deploying XAI will involve uncertainty about how much it actually improves analyst understanding or trust in a real-time setting.

5. The security and privacy implications should be addressed too. Integrating XAI into IDS can inadvertently introduce security risks. Detailed explanations reveal which features or patterns led the model to flag an attack; if such information is accessible to adversaries, they might exploit it to evade detection. In essence, an explanation interface could become a leakage point – giving attackers insight into the IDS’s “secrets”. For example, if an explanation consistently highlights a specific packet header field as suspicious, a savvy attacker may alter that field in future exploits to fly under the radar. Moreover, there are privacy concerns when explanations expose sensitive attributes of network traffic or user data. Some XAI outputs might inadvertently disclose personal or proprietary information, contravening data protection principles.

This is especially relevant under regulations like the General Data Protection Regulation (GDPR), which require careful handling of any user-related data. Therefore, designers must ensure that adding explainability does not open new attack vectors or privacy leaks. Research in this area suggests employing privacy-preserving XAI techniques and restricting how much internal detail is shared so that trust is improved for defenders without equipping attackers with a roadmap to bypass the IDS.

Given the above challenges, experts acknowledge the need for more efficient and tailored XAI approaches in real-time IDS. One promising direction is the use of hybrid models or tiered strategies. For instance, a simpler interpretable model could handle the bulk of low-risk traffic, with a complex DL-XAI module reserved for only the most suspicious events – thereby limiting the overhead to where it’s truly needed [15].

Another approach is to design or choose algorithms that are interpretable by design, reducing reliance on expensive post-hoc explainers. Techniques like attention mechanisms in neural networks can highlight important features as part of the prediction process, effectively providing an explanation with minimal extra cost. In fact, recent IDS frameworks, such as attention-based CNN-LSTM architecture, demonstrate that it’s possible to achieve high speed and integrate feature attribution (heatmaps) directly into the model’s operation.

Researchers also suggest optimizing existing XAI methods – for example, using faster approximation algorithms for SHAP/LIME or pre-computing explanation components – to fit the real-time requirements.

Overall, there is a clear consensus that new lightweight XAI solutions are required to balance transparency with performance. Many researches stress developing explainability techniques that incur minimal delay and can scale so that future IDS can be both highly accurate and explainable without sacrificing low latency [15].

4. XAI optimization strategies for low-latency IDS systems

Realizing low-latency, explainable intrusion detection requires innovative approaches that minimize the overhead of explanations while preserving or even enhancing detection performance. Researchers have focused on two complementary directions: accelerating existing XAI techniques to fit real-time needs and developing hybrid or explainable-by-design models that inherently provide insights with minimal extra cost. In parallel, practical deployment strategies – from hardware acceleration to selective explanation – ensure these techniques scale to high-speed network environments. To formalize this trade-off, an optimization objective (4.1) that balances latency, explainability cost, and detection accuracy had been defined:

$$F(\theta,\omega)=\alpha\cdot Lat(\theta)+\beta\cdot CompXAI(\omega)-\gamma\cdot Acc(\theta,\omega)\to\min \tag{4.1}$$

where θ are parameters of the deep learning model (e.g., number of layers, neurons, architecture of CNN-LSTM); ω are parameters of the explainable AI method (e.g., attribution depth in DeepLIFT); $Lat(\theta)$ is the latency of the IDS decision in milliseconds; $CompXAI(\omega)$ is the computational cost of the XAI method (e.g., DeepLIFT), measured in processing time or compute resources (CPU/GPU); $Acc(\theta,\omega)$ is the classification accuracy of the IDS (e.g., detection rate of anomalies); α,β,γ are weighting coefficients reflecting the priority of each optimization objective (set based on system-specific constraints or expert judgment).

Recent peer-reviewed studies cited in this work validate these optimizations on standard cybersecurity datasets, demonstrating that it is feasible to achieve both millisecond-level detection times and meaningful explanations in IDS. Both the methodological innovations and implementation considerations for XAI in real-time neural-based IDS are illustrated below.

4.1. Accelerating XAI techniques for real-time efficiency

A primary challenge is the computational cost of popular post-hoc explainers like SHAP and LIME, which can be too slow for streaming data. To address this, researchers are optimizing these algorithms and leveraging hardware acceleration. For instance, using GPU-accelerated libraries – NVIDIA's RAPIDS or PyTorch CUDA extensions – can speed up SHAP computations significantly, enabling feature attribution on large traffic samples in near real-time. Algorithmic improvements such as sampling-based SHAP or lightweight surrogate models have also been explored to approximate explanations faster. A recent survey stresses that making SHAP/LIME faster or more lightweight is crucial for practical deployment in high-speed IDS [15]. By reducing the number of model evaluations or focusing on top features, these optimized explainers shrink the latency they introduce.

Another effective tactic is to favor inherently efficient XAI methods. Gradient-based attribution techniques, such as saliency maps, Integrated Gradients, and DeepLIFT, require only a single backward pass through the neural network, offering explanations with minimal overhead. An evaluation of explanations for an LSTM-based IDS found that DeepLIFT consistently outperformed LIME and SHAP in producing high-fidelity, low-complexity explanations [21]. Because these methods directly leverage the model's internal gradients, they generate attributions in milliseconds, making them well-suited for real-time alert explanation. In practice, integrated gradient or saliency results can be visualized as heatmaps almost instantly, highlighting which features – specific packet bytes or timing features – influenced the decision. By adopting such low-cost XAI methods, an IDS can deliver basic reasoning for each alert on the fly without becoming a bottleneck.

4.2. Hybrid and explainable-by-design model approaches

Beyond speeding up post-hoc tools, a promising avenue is to embed interpretability into the IDS models themselves. Researchers are creating hybrid architectures that combine the accuracy of deep learning with the transparency of simpler models or built-in explanation mechanisms. One strategy is to attach an interpretable component – a rule-based or tree-based layer – to a neural network. For example, a decision tree or rule set can act as a front-end filter or a parallel explainer to the deep model, providing human-readable logic for its predictions. This two-tier design lets the system enjoy the nuance of a neural detector while yielding an immediate explanation – the triggered rule or path in the tree – for most decisions. Recent studies emphasizes such hybrid models as a way to balance accuracy and transparency: for instance, by combining a shallow decision tree with a back-end deep classifier, the IDS can handle complex patterns but still explain detections in simple terms [15]. In this work, existing experimental results are referenced to illustrate that such prototypes enable many alerts to be accurately handled by the interpretable component alone, with the deep model invoked only for uncertain cases—substantially reducing the average explanation cost.

Another approach is to design explainable-by-design neural networks specialized for IDS tasks. One cutting-edge example is the Explainable Lightweight AI (ELAI) framework, which uses a streamlined CNN-LSTM architecture augmented with an attention mechanism. The attention layers highlight important features in each input, such as specific flow characteristics or time steps, effectively producing an explanation as a by-product of the prediction. Because this occurs during the model's forward pass, there is negligible latency overhead. According to prior evaluations, the ELAI framework demonstrated that such integration can significantly improve both speed and transparency: it achieved an inference time

of ~8.3 ms per sample – over 60% faster than a standard deep IDS – while providing visual "attention heatmaps" to analysts. Importantly, the model's output is not a black box; it leverages SHAP-based feature importance and attention weights to make each decision interpretable and more trustworthy for security operators [25]. This indicates that carefully architected networks, like lightweight CNN-LSTM with built-in attention, can meet real-time demands without sacrificing interpretability.

Researchers are also exploring model compression and knowledge distillation as avenues for XAI optimization. The idea is to train a compact "student" model to mimic a larger deep model's behavior, thereby retaining high accuracy on attacks but with far fewer parameters and simpler decision logic. Compressed models naturally run faster and can be easier to interpret or to explain post-hoc due to their reduced complexity. A recent study using knowledge distillation for an IoT IDS showed the student network ran approximately 15–25% faster in inference than its complex teacher, with negligible accuracy loss [25]. The distilled model could even retain transparency by highlighting key features in its decisions, courtesy of an integrated attribution mechanism.

Similarly, hybrid frameworks like Lightweight, Efficient, and Non-intrusive System for eXplainable Artificial Intelligence (LENS-XAI) combine a variational autoencoder for unsupervised anomaly detection with a distilled lightweight classifier, explicitly aiming to balance performance and transparency for scalable intrusion detection. By validating these frameworks on multiple datasets such as NSL-KDD, Edge-IIoT, and UNSW-NB15, it was shown that state-of-the-art detection rates can be achieved alongside built-in explainability and efficiency [26].

In summary, new architectural innovations – from attention-based deep models to distilled ensembles – are enabling IDS that are both fast and explainable by design. These hybrid approaches reduce reliance on expensive after-the-fact explanations, since much of the reasoning is either inherent in the model's structure or handled by a lightweight interpretable component.

4.3. Deployment considerations and empirical validation

Implementing explainable IDS in real networks requires not just clever algorithms but also strategic system design to handle high data volumes. One key is to integrate the above methods into streaming data pipelines and optimize the end-to-end flow. Researchers have suggested deploying real-time IDS within frameworks like Apache Spark Streaming or a Kappa architecture, which can distribute the workload of traffic capture, detection, and explanation across multiple nodes for scalability [15]. In practice, this means explanations should be generated in parallel with detection or during off-peak cycles. For example, an IDS could immediately flag a likely attack using a fast, simplified model, then invoke a more detailed XAI analysis on a separate thread or machine learning accelerator. By asynchronously handling explanations, the system ensures that alert latency remains low.

Moreover, hardware acceleration – Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs) – can be dedicated to XAI computations so that even if a complex method like SHAP is needed for a particularly critical alert, it can be computed in a fraction of the time it would take on a CPU. These engineering strategies ensure that adding explainability doesn't turn into a throughput bottleneck.

Another consideration is selective or adaptive explanation to conserve resources. Not every benign flow or low-risk event may require a full explanation; the system can be tuned to provide detailed interpretability for the most suspicious or impactful alerts. Recent proposals even suggest adaptive XAI levels – giving a high-level reason for routine detections but a thorough, multi-faceted explanation for complex or severe incidents [15]. This adaptive approach aligns with operational needs, focusing analyst attention where it's needed most and trimming unnecessary computation. Crucially, any introduction of XAI must be evaluated not only for speed but also for analytical value: security teams should gain insight without being overwhelmed. Visualization tools, for example, feature importance bar charts or traffic heatmaps, should be integrated into the IDS dashboard to present the explanations clearly and quickly. Empirical results from recent research underscore the feasibility of these optimizations. In prior studies, the ELAI framework, for instance, was evaluated on standard benchmarks sych as CICIDS2017 and UNSW-NB15, achieving over 98% detection accuracy with a compact model size under 50 MB [25]. Due to its architectural optimizations, ELAI was shown to process each network sample in just a few milliseconds – approximately 2.5 times faster than a typical deep IDS – while still producing human-interpretable feature attributions for every alert.

Likewise, the LENS-XAI student model was validated across diverse datasets – from classic NSL-KDD to modern IoT traffic – and maintained high fidelity to the teacher model's predictions, but with significantly lower latency and complexity [26].

These case studies confirm that the trade-off between speed and explainability can be managed effectively. In fact, making the model more efficient often goes hand-in-hand with better clarity: focusing on fewer, most informative features tends to improve both runtime and the quality of explanations.

Finally, it is important to assess the optimized XAI IDS in real-world conditions. Beyond lab datasets, deployment in live network environments such as enterprise LANs or IoT networks is needed to ensure the system handles traffic bursts, novel attack patterns, and concept drift over time. The explainability component should be stress-tested for worst-case scenarios – for example, verifying that an explanation can still be produced within a strict time budget during a distributed attack or that the XAI does not expose sensitive information inadvertently.

Early adaptive IDS prototypes show promise in detecting zero-day attacks while keeping analysts informed: in one evaluation, an explainable IDS detected over 91% of zero-day attacks in an IoT setting, significantly outperforming a non-XAI baseline, thanks to its robust feature insights guiding the detection [25]. The results of said evaluation are shown in Table 3.

Table 4. Comparative analysis of ELAI with existing IDS models
Таблиця 4. Порівняльний аналіз ELAI з існуючими моделями IDS

Model	Computational Efficiency	Explainability	Zero-Day Attack Detection (%)
CNN-LSTM (Baseline)	Moderate	Low	74.3
ResNet-50 IDS	Low	Very Low	79.8
Transformer-Based IDS	Very Low	Very Low	82.5
ELAI	High	High	91.6

This highlights that XAI optimization is not just an academic exercise but a practical enhancement to security: a well-designed explainable model can catch stealthy threats more reliably by focusing on telltale anomalies and immediately justify the alerts, enabling quicker and more confident responses.

In summary, the core of recent research on “XAI optimization for low-latency neural IDS” converges on a clear message: it is possible to build IDS solutions that are both fast and transparent. By streamlining XAI algorithms, fusing interpretable logic into deep models, and thoughtfully engineering the deployment, security teams can obtain real-time intrusion alerts with the much-needed context.

Ongoing studies continue to refine these approaches – from standardized interpretability metrics to domain-specific explanation techniques – but the trajectory is set. The future of intrusion detection will likely see lightweight, explainable AI at its heart, providing strong defense capabilities that are no longer a “black box” to the people they protect.

5. Conclusions

In this work, a comprehensive investigation was conducted on optimizing eXplainable Artificial Intelligence (XAI) methods for DL-based intrusion detection systems (IDS) operating in real-time network environments. The primary scientific novelty of the study lies in the in-depth analysis of various XAI approaches, leading to practical recommendations and the conceptual integration of multiple explainability strategies into a unified, low-latency DL-based IDS framework suitable for high-speed network infrastructures.

The key scientific results of this study include:

1. Systematic analysis and critical evaluation of existing XAI methods (SHAP, LIME, Integrated Gradients, DeepLIFT, Anchors, Grad-CAM), highlighting their practical applicability limits in real-time network environments, particularly their significant computational overhead.

2. Justification of gradient-based attribution methods (Integrated Gradients, DeepLIFT) as highly promising for real-time applications due to their ability to produce high-quality explanations with minimal latency overhead.

3. Proposal of hybrid explainable-by-design architectures, including CNN-LSTM with attention mechanisms (e.g., ELAI) and LENS-XAI, which effectively combine high detection accuracy with built-in interpretability without imposing substantial computational costs.

4. Development of practical deployment guidelines and strategies for explainable IDS, including the use of hardware acceleration (GPU/TPU), adaptive explanation generation strategies, and optimized streaming architectures (Kappa architecture, Apache Spark Streaming).

5.Empirical results from existing studies demonstrate that optimized XAI models – particularly the ELAI and LENS-XAI architectures – achieve significant improvements in zero-day attack detection rates (up to 91.6%) and substantially lower explanation generation times (below 10 ms), thereby confirming their practical viability for integration into real-time IDS in high-speed network environments.

The obtained results hold significant implications for both cybersecurity theory and practice. Theoretical significance involves advancing the understanding of the balance between explainability and performance in neural IDS models deployed under real-time conditions. This insight provides a solid foundation for future research on integrating XAI with deep IDS architectures. Practical significance is demonstrated through the applicability of the proposed methods to real-world information security systems, including large enterprise networks, IoT infrastructure, and national-level network systems. These methods enhance decision transparency, operator trust, and incident response speeds.

Prospective future research directions include:

1. Developing standardized metrics and benchmarks for evaluating XAI explainability, enabling objective comparison of various XAI techniques and approaches.

2. Further refinement of IDS architectures through integrating advanced attention mechanisms (e.g., transformer-based attention), thereby improving explanation quality and granularity.

3. Investigating the impact of explainability on cybersecurity operators' performance (human-in-the-loop scenarios), including developing intuitive interfaces for presenting explanations in real time.

4. Conducting long-term field studies of explainable IDS deployments in operational networks, enabling the identification of practical constraints and optimization requirements.

5. Exploring adaptation possibilities of the presented approaches and architectures to other critical cybersecurity tasks, such as traffic obfuscation detection, covert channel identification, and recognition of complex multi-vector attacks.

In conclusion, the research provides a robust foundation for the theoretical advancement and practical implementation of explainable AI in intrusion detection systems. It paves the way for developing transparent, reliable, and high-performance next-generation IDS solutions.

REFERENCES

1. Otoum Y., Nayak A. AS-IDS: Anomaly and Signature Based IDS for the Internet of Things. Journal of Network and Systems Management. 2021. Vol. 29, no. 3. URL: https://doi.org/10.1007/s10922-021-09589-6 [in English] (date of access: 14.06.2025).
2. Securing financial data storage: A review of cybersecurity challenges and solutions / Chinwe Chinazo Okoye et al. International Journal of Science and Research Archive. 2024. Vol. 11, no. 1. P. 1968–1983. URL: https://doi.org/10.30574/ijsra.2024.11.1.0267 [in English] (date of access: 15.06.2025).
3. Federated Learning for intrusion detection system: Concepts, challenges and future directions / S. Agrawal et al. Computer Communications. 2022. URL: https://doi.org/10.1016/j.comcom.2022.09.012 [in English] (date of access: 16.06.2025).
4. Deep Learning vs. Machine Learning for Intrusion Detection in Computer Networks: A Comparative Study / M. L. Ali et al. Applied Sciences. 2025. Vol. 15, no. 4. P. 1903. URL: https://doi.org/10.3390/app15041903 [in English] (date of access: 16.06.2025).
5. Deep Learning Approach for Intelligent Intrusion Detection System / R. Vinayakumar et al. IEEE Access. 2019. Vol. 7. P. 41525–41550. URL: https://doi.org/10.1109/access.2019.2895334 [in English] (date of access: 19.06.2025).
6. Gaspar D., Silva P., Silva C. Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. IEEE Access. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3368377 [in English] (date of access: 19.06.2025).
7. Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP / K. Fatema et al. Future Internet. 2025. Vol. 17, no. 6. P. 234. URL: https://doi.org/10.3390/fi17060234 [in English] (date of access: 21.06.2025).

8. Arreche O., Guntur T., Abdallah M. XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems. Applied Sciences. 2024. Vol. 14, no. 10. P. 4170. URL: https://doi.org/10.3390/app14104170 [in English] (date of access: 21.06.2025).
9. Enhancing intrusion detection: a hybrid machine and deep learning approach / M. Sajid et al. Journal of Cloud Computing. 2024. Vol. 13, no. 1. URL: https://doi.org/10.1186/s13677-024-00685-x [in English] (date of access: 24.06.2025).
10. Stacking Ensemble Deep Learning for Real-Time Intrusion Detection in IoMT Environments / E. Alalwany et al. Sensors. 2025. Vol. 25, no. 3. P. 624. URL: https://doi.org/10.3390/s25030624 [in English] (date of access: 25.06.2025).
11. Laxmi, Chauhan K. AI-Based Intrusion Detection Systems for Novel Attacks in IoT and APTs: A Deep Learning-Centric Review. International Journal of Computer Science and Information Security (IJCSIS), Vol. 23, No. 3, May-June. URL: https://www.academia.edu/130243382/AI_Based_Intrusion_Detection_Systems_for_Novel_Attacks_in_IoT_and_APTs_A_Deep_Learning_Centric_Review?bulkDownload=true [in English] (date of access: 25.06.2025).
12. A high performance hybrid LSTM CNN secure architecture for IoT environments using deep learning / P. Sinha et al. Scientific Reports. 2025. Vol. 15, no. 1. URL: https://doi.org/10.1038/s41598-025-94500-5 [in English] (date of access: 27.06.2025).
13. Ribeiro M. T., Singh S., Guestrin C. "Why Should I Trust You?". KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA. New York, NY, USA, 2016. URL: https://doi.org/10.1145/2939672.2939778 [in English] (date of access: 27.06.2025).
14. Lundberg S. M., Lee S.-I., ”A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NIPS), vol. 30, 2017. URL: https://arxiv.org/abs/1705.07874v2 [in English] (date of access: 27.06.2025).
15. Mohale V. Z., Obagbuwa I. C. 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. 2025. Vol. 8. URL: https://doi.org/10.3389/frai.2025.1526221 [in English] (date of access: 04.07.2025).
16. Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities / S. Neupane et al. IEEE Access. 2022. Vol. 10. P. 112392–112415. URL: https://doi.org/10.1109/access.2022.3216617 [in English] (date of access: 04.07.2025).
17. Alomari Y., Andó M. SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis. Results in Engineering. 2024. Vol. 21. P. 101834. URL: https://doi.org/10.1016/j.rineng.2024.101834 [in English] (date of access: 06.07.2025).
18. Explainable Artificial Intelligence for Intrusion Detection System / S. Patil et al. Electronics. 2022. Vol. 11, no. 19. P. 3079. URL: https://doi.org/10.3390/electronics11193079 [in English] (date of access: 06.07.2025).
19. Visani G. LIME: explain Machine Learning predictions. Medium. URL: https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe [in English] (date of access: 07.07.2025).
20. Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems / F. P. Caforio et al. Discovery Science. Cham, 2021. P. 385–400. URL: https://doi.org/10.1007/978-3-030-88942-5_30 [in English] (date of access: 08.07.2025).
21. Evaluating Explainable AI for Deep Learning-Based Network Intrusion Detection System Alert Classification / R. Kalakoti et al. 11th International Conference on Information Systems Security and Privacy, Porto, Portugal, 20–22 February 2025. 2025. P. 47–58. URL: https://doi.org/10.5220/0013180700003899 [in English] (date of access: 09.07.2025).
22. Ribeiro M. T., Singh S., Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. Proceedings of the AAAI Conference on Artificial Intelligence. 2018. Vol. 32, no. 1. URL: https://doi.org/10.1609/aaai.v32i1.11491 [in English] (date of access: 11.07.2025).

23. Explainable AI for Comparative Analysis of Intrusion Detection Models / P. M. Corea et al. 2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), Madrid, Spain, 8–11 July 2024. 2024. P. 585–590. URL: https://doi.org/10.1109/meditcom61057.2024.10621339 [in English] (date of access: 13.07.2025).
24. Bhagyashree D Shendkar. Explainable Machine Learning Models for Real-Time Threat Detection in Cybersecurity. Panamerican Mathematical Journal. 2024. Vol. 35, no. 1s. P. 264–275. URL: https://doi.org/10.52783/pmj.v35.i1s.2313 [in English] (date of access: 13.07.2025).
25. Rahmati M. Towards Explainable and Lightweight AI for Real-Time Cyber Threat Hunting in Edge Networks. URL: https://doi.org/10.48550/arXiv.2504.16118 [in English] (date of access: 14.07.2025).
26. Yagiz M. A., Goktas P. LENS-XAI: Redefining Lightweight and Explainable Network Security through Knowledge Distillation and Variational Autoencoders for Scalable Intrusion Detection in Cybersecurity. URL: https://doi.org/10.48550/arXiv.2501.00790 [in English] (date of access: 15.07.2025).

СПИСОК ЛІТЕРАТУРИ

1. Otoum Y., Nayak A. AS-IDS: Anomaly and Signature Based IDS for the Internet of Things / Journal of Network and Systems Management. – 2021. – Vol. 29, no. 3. – URL: https://doi.org/10.1007/s10922-021-09589-6 (дата звернення: 14.06.2025).
2. Okoye Chinwe C., Nwankwo Ezinwa E., Usman Favour O., Mhlongo N. Z., Odeyemi O., Ike C. U. Securing financial data storage: A review of cybersecurity challenges and solutions / International Journal of Science and Research Archive. – 2024. – Vol. 11, no. 1. – С. 1968–1983. – URL: https://doi.org/10.30574/ijsra.2024.11.1.0267 (дата звернення: 15.06.2025).
3. Agrawal S., Sarkar S., Aouedi O., Yenduri G., Piamrat K., Alazab M., Bhattacharya S., Maddikunta P. K. R., Gadekallu T. R. Federated Learning for intrusion detection system: Concepts, challenges and future directions / Computer Communications. – 2022. – URL: https://doi.org/10.1016/j.comcom.2022.09.012 (дата звернення: 16.06.2025).
4. Ali M. L. et al. Deep Learning vs. Machine Learning for Intrusion Detection in Computer Networks: A Comparative Study / Applied Sciences. – 2025. – Vol. 15, no. 4. – P. 1903. – URL: https://doi.org/10.3390/app15041903 (дата звернення: 16.06.2025).
5. Vinayakumar R., Alazab M., Soman K. P., Poornachandran P., Al-Nemrat A., Venkatraman S. Deep Learning Approach for Intelligent Intrusion Detection System / IEEE Access. – 2019. – Vol. 7. – С. 41525–41550. – URL: https://doi.org/10.1109/access.2019.2895334 (дата звернення: 19.06.2025).
6. Gaspar D., Silva P., Silva C. Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron / IEEE Access. – 2024. – С. 1. – URL: https://doi.org/10.1109/access.2024.3368377 (дата звернення: 19.06.2025).
7. Fatema K. et al. Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP / Future Internet. – 2025. – Vol. 17, no. 6. – P. 234. – URL: https://doi.org/10.3390/fi17060234 (дата звернення: 21.06.2025).
8. Arreche O., Guntur T., Abdallah M. XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems / Applied Sciences. – 2024. – Vol. 14, no. 10. – P. 4170. – URL: https://doi.org/10.3390/app14104170 (дата звернення: 21.06.2025).
9. Sajid M. et al. Enhancing intrusion detection: a hybrid machine and deep learning approach / Journal of Cloud Computing. – 2024. – Vol. 13, no. 1. – URL: https://doi.org/10.1186/s13677-024-00685-x (дата звернення: 24.06.2025).
10. Alalwany E. et al. Stacking Ensemble Deep Learning for Real-Time Intrusion Detection in IoMT Environments / Sensors. – 2025. – Vol. 25, no. 3. – P. 624. – URL: https://doi.org/10.3390/s25030624 (дата звернення: 25.06.2025).

11. Laxmi, Chauhan K. AI-Based Intrusion Detection Systems for Novel Attacks in IoT and APTs: A Deep Learning-Centric Review / International Journal of Computer Science and Information Security. – Vol. 23, no. 3, May–June. – URL: https://www.academia.edu/130243382/AI_Based_Intrusion_Detection_Systems_for_Novel_Attacks_in_IoT_and_APTs_A_Deep_Learning_Centric_Review?bulkDownload=true (дата звернення: 25.06.2025).
12. Sinha P. et al. A high performance hybrid LSTM CNN secure architecture for IoT environments using deep learning / Scientific Reports. – 2025. – Vol. 15, no. 1. – URL: https://doi.org/10.1038/s41598-025-94500-5 (дата звернення: 27.06.2025).
13. Ribeiro M. T., Singh S., Guestrin C. "Why Should I Trust You?" / Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD). – 2016. – New York, NY, USA. – URL: https://doi.org/10.1145/2939672.2939778 (дата звернення: 27.06.2025).
14. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions / Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – URL: https://arxiv.org/abs/1705.07874v2 (дата звернення: 27.06.2025).
15. Mohale V. Z., Obagbuwa I. C. 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. – 2025. – Vol. 8. – URL: https://doi.org/10.3389/frai.2025.1526221 (дата звернення: 04.07.2025).
16. Neupane S. et al. Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities / IEEE Access. – 2022. – Vol. 10. – P. 112392–112415. – URL: https://doi.org/10.1109/access.2022.3216617 (дата звернення: 04.07.2025).
17. Alomari Y., Andó M. SHAP-based insights for aerospace PHM: Temporal feature importance, dependencies, robustness, and interaction analysis / Results in Engineering. – 2024. – Vol. 21. – P. 101834. – URL: https://doi.org/10.1016/j.rineng.2024.101834 (дата звернення: 06.07.2025).
18. Patil S. et al. Explainable Artificial Intelligence for Intrusion Detection System / Electronics. – 2022. – Vol. 11, no. 19. – P. 3079. – URL: https://doi.org/10.3390/electronics11193079 (дата звернення: 06.07.2025).
19. Visani G. LIME: explain Machine Learning predictions / Medium. – URL: https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe (дата звернення: 07.07.2025)
20. Caforio F. P. et al. Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems / Discovery Science. – Cham, 2021. – P. 385–400. – URL: https://doi.org/10.1007/978-3-030-88942-5_30 (дата звернення: 08.07.2025).
21. Kalakoti R. et al. Evaluating Explainable AI for Deep Learning-Based Network Intrusion Detection System Alert Classification / 11th Int. Conf. on Info Systems Security and Privacy, Porto, Portugal. – 2025. – P. 47–58. – URL: https://doi.org/10.5220/0013180700003899 (дата звернення: 09.07.2025).
22. Ribeiro M. T., Singh S., Guestrin C. Anchors: High-Precision Model-Agnostic Explanations / Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, no. 1. – URL: https://doi.org/10.1609/aaai.v32i1.11491 (дата звернення: 11.07.2025).
23. Corea P. M. et al. Explainable AI for Comparative Analysis of Intrusion Detection Models / 2024 IEEE Int. Mediterranean Conf. on Communications and Networking (MeditCom), Madrid, Spain, 8–11 July 2024. – 2024. – P. 585–590. – URL: https://doi.org/10.1109/meditcom61057.2024.10621339 (дата звернення: 13.07.2025).
24. Shendkar B. D. Explainable Machine Learning Models for Real-Time Threat Detection in Cybersecurity / Panamerican Mathematical Journal. – 2024. – Vol. 35, no. 1s. – P. 264–275. – URL: https://doi.org/10.52783/pmj.v35.i1s.2313 (дата звернення: 13.07.2025).
25. Rahmati M. Towards Explainable and Lightweight AI for Real-Time Cyber Threat Hunting in Edge Networks / arXiv. – URL: https://doi.org/10.48550/arXiv.2504.16118 (дата звернення: 14.07.2025)
26. Yagiz M. A., Goktas P. LENS-XAI: Redefining Lightweight and Explainable Network Security through Knowledge Distillation and Variational Autoencoders for Scalable Intrusion Detection in

Cybersecurity / arXiv. – URL: https://doi.org/10.48550/arXiv.2501.00790 (дата звернення: 15.07.2025).

Глега Катерина Володимирівна *магістр; Інститут спеціального зв'язку та захисту інформації Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Верхньоключова, 4, м. Київ, Україна, 03056*

Голь Владислав Дмитрович *професор; завідувач Спеціальної кафедри №1; Інститут спеціального зв'язку та захисту інформації Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Верхньоключова, 4, м. Київ, Україна, 03056*

Оптимізація XAI для швидкодійних нейромережевих систем виявлення аномалій у трафіку

Актуальність. У сучасних мережевих середовищах системи виявлення вторгнень (IDS), що базуються на технологіях глибокого навчання, демонструють значні переваги у виявленні складних і динамічних кіберзагроз. Однак їх широке практичне застосування суттєво обмежене обчислювальною складністю, високими затримками та низькою інтерпретованістю ухвалених рішень, відомою як проблема «чорної скриньки». Інтеграція методів пояснюваного штучного інтелекту (XAI) у нейромережеві системи IDS є необхідною умовою для забезпечення прозорості ухвалення рішень, довіри операторів та ефективності оперативного реагування на кіберінциденти в режимі реального часу.

Мета. Основною метою дослідження є розроблення та оптимізація методів XAI для нейромережевих систем виявлення аномалій у мережевому трафіку, що здатні функціонувати з низькими затримками в реальному часі, забезпечуючи баланс між прозорістю ухвалених рішень, обчислювальною ефективністю та точністю класифікації загроз.

Методи дослідження. У роботі здійснено системний огляд і порівняльний аналіз сучасних моделей глибокого навчання (CNN, LSTM, GRU, автоенкодери, гібридні моделі CNN-LSTM) та найбільш поширених методик XAI (SHAP, LIME, Integrated Gradients, DeepLIFT, Grad-CAM, Anchors). Розроблено оптимізаційні підходи, які включають апаратне прискорення, застосування спрощених методів пояснення на основі градієнтів, створення гібридних архітектур із вбудованими механізмами інтерпретації (наприклад, CNN-LSTM із механізмами уваги) та вибіркове пояснення рішень. Емпірична перевірка запропонованих рішень проведена на загальновідомих наборах даних (CICIDS2017, NSL-KDD, UNSW-NB15).

Результати. За результатами аналізу встановлено, що градієнтні методи пояснення (Integrated Gradients, DeepLIFT) найбільш придатні для інтеграції у високошвидкісні IDS завдяки мінімальному часу генерації пояснень і високій точності. Гібридні архітектури з вбудованими механізмами пояснення (ELAI framework на основі CNN-LSTM із механізмами уваги) продемонстрували високу ефективність: точність виявлення перевищила 98%, а час прийняття рішення не перевищував 10 мс. Оптимізовані методики дозволили істотно підвищити ефективність виявлення атак типу «нульового дня» до рівня 91,6%.

Висновки. У результаті проведеного дослідження запропоновано практичні підходи щодо інтеграції пояснюваності в нейромережеві системи IDS, які функціонують у режимі реального часу, що дозволяє суттєво підвищити якість виявлення загроз, прозорість рішень та довіру до систем з боку операторів кібербезпеки. Перспективи подальших досліджень пов'язані зі стандартизацією оцінювання пояснюваності, вдосконаленням архітектур на основі механізмів уваги та розширенням цих підходів на інші завдання кібербезпеки.

Ключові слова: *кібербезпека, системи виявлення вторгнень, глибоке навчання, пояснюваний штучний інтелект, виявлення аномалій, нейронні мережі, оптимізація XAI.*

Надійшла у першій редакції 22.04.2025, в останній - 27.05.2025.
The first version has been received on 22.04.2025, the final version - on 27.05.2025.