Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140
Volume 14 Issue 02, 2025

The Role of Deep Learning in Network Security using Explainable Artificial Intelligence

Dr. R.D.Bhoyar
Assistant Professor
Department of Computer Science,
Sant Gadge Baba Amravati Universit, Amravati.
Email: rajeshbhoyar@sgbau.ac.in

Peer Review Information

Submission: 05 Aug 2025

Revision: 15 Aug 2025

Acceptance: 05 Sept 2025

Keywords

Deep learning, Network Security, intrusion detection systems, SHAP, LIME, Grad-CAM, XAI-IDS

Abstract

Deep learning (DL) methods have advanced network security capabilities across intrusion detection, malware detection, traffic classification, and threat hunting by learning complex patterns from high-dimensional data. However, DL models are often black boxes, which limit operational adoption in Security Operations Centers (SOCs) where human analysts must trust, verify, and act on model outputs. Explainable AI (XAI) techniques bridge this gap by providing local and global explanations that increase transparency, enable model debugging, and improve analyst decision-making. This paper surveys DL applications in network security, reviews XAI methods adapted to cyber security, proposes an integrated XAI–DL framework for intrusion detection, reports an evaluation strategy, and discusses challenges and future directions.

Introduction

The rapid proliferation of digital communication, cloud computing, Internet of Things (IoT) devices, and high-speed networks has dramatically increased the complexity and volume of network traffic. While these advances enable new business and societal opportunities, they also create an expanded attack surface for cybercriminals, nation-state actors, and insider threats. Modern adversaries employ sophisticated tactics—such as advanced persistent threats (APTs), zero-day exploits, and polymorphic malware—that evolve faster than traditional signature-based security mechanisms can adapt.

Conventional network security tools, including traditional Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS), rely heavily on predefined rules or manually engineered features. These methods often struggle to detect novel or obfuscated attack patterns and suffer from high false-positive rates, leading to alert fatigue among Security Operations Center (SOC) analysts. Consequently, there is a growing demand for intelligent, adaptive, and automated solutions capable of recognizing both known and previously unseen threats in real time.

In recent years, Deep Learning (DL) has emerged as a transformative technology in the cyber security domain. By leveraging hierarchical representation learning, DL models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) including Long Short-Term Memory (LSTM) networks, Transformers, and Graph Neural Networks (GNNs) can automatically extract complex patterns from raw or minimally processed network data. DL-based systems have demonstrated superior performance over traditional machine learning approaches in tasks such as intrusion detection, malware

classification, network traffic analysis, and anomaly detection.

However, despite these performance gains, DL models are often regarded as "black boxes"—producing high-confidence predictions without offering clear insight into the reasoning process behind them.

This lack of transparency raises several operational challenges:

- Trust Deficit: SOC analysts may hesitate to act on alerts from opaque models, especially when high-risk remediation steps are involved (e.g., blocking an IP address or isolating a host).
- Regulatory Compliance: Many industries (e.g., finance, healthcare, defense) require explanations for automated security decisions to meet legal or policy obligations.
- Incident Investigation: Post-incident forensics often demand a clear explanation of why an event was flagged, including which network features or behaviors triggered the detection.
- Model Debugging: Without interpretability, security engineers struggle to identify biases, data quality issues, or vulnerabilities to adversarial manipulation.

To address these limitations, Explainable Artificial Intelligence (XAI) has gained prominence as a complementary approach to deep learning in network security. XAI encompasses a range of methods—such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM), and Integrated Gradients—that aim to make model predictions understandable to humans without sacrificing performance.

When integrated into deep learning (DL)-based security systems, explainable artificial intelligence (XAI) offers several critical benefits. It can reveal the key features or patterns that influence a model's decision, thereby enhancing analyst confidence and trust in automated outputs. By highlighting the most relevant network attributes, XAI facilitates faster and more targeted incident triage. Furthermore, it supports compliance with transparency requirements by generating human-readable justifications for security decisions. In addition, XAI can assist in detecting and mitigating adversarial attacks that seek to exploit vulnerabilities in the underlying models, thereby strengthening the overall resilience of the security framework.

This convergence of Deep Learning and Explainable AI represents a critical shift in cyber security—from opaque "black box" detection engines to transparent, analyst-centered, and accountable security intelligence systems. While numerous studies have explored DL architectures for intrusion detection or malware classification, fewer have systematically examined how XAI can be seamlessly integrated into these systems to enhance their operational viability in real-world SOC environments.

The remainder of this paper addresses this gap by first reviewing the current state of deep learning (DL) applications in network security. It then presents a taxonomy of explainable artificial intelligence (XAI) methods suited for security data and DL models, followed by a proposed practical XAI–DL integration framework for network intrusion detection. Furthermore, it outlines an evaluation methodology that combines detection accuracy with explanation quality and human usability metrics. Finally, it discusses the challenges, limitations, and potential directions for future research.

Through this exploration, the paper aims to advance the development of trustworthy, transparent, and high-performance deep learning solutions that are both technically robust and operationally acceptable in modern network defense.

Literature Survey

Deep Learning in Network Security

Deep learning (DL) techniques have been applied across various domains within network security. In intrusion detection systems (IDS/IPS), methods such as LSTM, CNN, autoencoders, and hybrid models have been employed for both anomaly-based and signature-based detection. For malware detection, CNNs have been utilized on binary, memory, and image representations, while sequence models are applied to analyze API call traces. In the area of network traffic classification and quality of service (QoS) monitoring, CNNs and transformer architectures are used to classify encrypted traffic and detect covert communication channels. Additionally, in threat hunting and triage, graph neural networks (GNNs) have been leveraged to model entity-relationship graphs—capturing interactions between hosts, processes, and files—to identify patterns indicative of lateral movement within a network. Representative surveys and domain papers document these trends and evaluate datasets, performance, and limitations.

Explainable AI (XAI)
XAI methods fall into model-agnostic and model-specific categories, and local vs global explanations:

- **Model-agnostic:** LIME, SHAP — perturbation- or game-theory-based feature attributions usable across models.
- **Model-specific:** Grad-CAM, Integrated Gradients — leverage network internals (gradients, activations) for explanations.
- **Surrogate and rule-based approaches:** Train an interpretable surrogate (decision tree, rule set) to approximate the black box for global interpretability.

Surveys specific to XAI in cybersecurity highlight the adoption of SHAP/LIME/Grad-CAM and newer evaluation metrics tailored to analysts' needs.

Sequence models (RNNs, LSTM, GRU)
Sequence models (LSTM/GRU/Bi-LSTM) are widely used to model temporal dependencies in flows and session traces. Early and medium-scale studies show LSTMs outperform classical ML on time-series flow features for anomaly detection, particularly when modeling multi-step attack behaviors. Practical caveats include long training times and sensitivity to windowing/aggregation choices.

Convolutional neural networks (CNNs) & 1D/2D transforms
CNNs have been used both on engineered feature matrices and on transformed representations (e.g., packet bytes → images, spectrogram-like transforms). CNNs excel at learning local patterns (byte motifs, header–payload correlations) and have shown effectiveness in malware-binary classification and payload analysis where raw byte patterns matter.

Transformers and attention models
Transformers and attention-based models have been introduced more recently for traffic classification and long-range dependency modeling; they offer scalability and better handling of varied-length sequences. Recent surveys and experimental papers demonstrate promising performance but point out higher compute and data requirements.

Auto encoders, VAEs, and anomaly detection
Auto encoder families (vanilla, variational, and denoising) are commonly applied for unsupervised anomaly detection by learning compact representations of normal traffic and flagging high-reconstruction-error flows. Their main advantages are adaptability to unlabeled data and detection of novel attacks; a key limitation is high false-positive rates without domain-specific tuning.

XAI Matters for Network Security

- **Trust and Adoption:** SOC analysts require intelligible rationales for high-impact alerts. XAI increases acceptance of automated alerts.
- **Incident Triage and Forensics:** Explanations (feature attributions, salient packet segments) speed investigation by pointing to root causes.
- **Compliance and Accountability:** Regulations and enterprise policy can demand interpretable decisions.
- **Model Validation and Robustness:** Explanations help detect dataset shift, adversarial manipulation, and concept drift. Recent surveys and practitioner studies document these operational benefits and current gaps.

A Framework for Network Intrusion Detection
Goals & design principles

- High detection accuracy for known and novel attacks.
- Multi-modal input handling (flows, packets, host logs, graphs).
- Explainability: produce human-understandable local and global explanations.
- Real-time/near-real-time operation with fallbacks for batch analysis.
- Analyst-centered outputs: concise, actionable explanations to speed triage.
- Robustness to adversarial manipulation and concept drift.

High-level architecture (modules)
1. **Data Ingestion & Normalization :**

The data ingestion and normalization process begins with aggregating information from diverse sources, including NetFlow/IPFIX records, full packet captures (pcap), host logs such as Syslog and Windows event logs, DNS logs, and external threat intelligence feeds. This raw data is then enriched through additional contextual information, such as geo-IP mapping, autonomous system number (ASN) identification, WHOIS records, reverse DNS lookups, and label mapping. The processed output is a unified event stream supplemented with windowed flow summaries, providing a consistent and structured foundation for subsequent analysis.

2. **Feature Extraction & Representation:**
The feature extraction and representation process incorporates multiple data modalities to comprehensively capture network behavior. Tabular features consist of flow-level statistics such as byte and packet counts, connection duration, protocol flags, and port information. Sequence features are derived from ordered packet payloads, which can be represented as tokenized bytes or extracted n-grams to preserve content-level patterns. Temporal characteristics are modeled through time-series windows, where sliding windows are applied to individual flows or hosts to generate inputs suitable for sequential architectures such as LSTMs and Transformers. In addition, graph-based representations are constructed, with nodes representing entities such as hosts, IP addresses, and ports, and edges denoting relationships like network connections or file accesses. These graphs are further enriched with node and edge attributes to support advanced relational reasoning through graph neural networks (GNNs).

3. Detection Engine (Model Zoo)
The detection engine leverages a diverse model zoo to balance real-time responsiveness with high-fidelity analysis. Lightweight models, such as gradient-boosted trees or shallow CNN architectures, are employed for rapid filtering in time-sensitive scenarios. For more comprehensive detection, deep learning models are utilized across multiple modalities: temporal models, including bidirectional LSTMs and Transformer encoders, process flow-level sequences; spatial or byte-level patterns are captured using 1D-CNNs for packet or byte motifs, and 2D-CNNs for byte-to-image transformations; and graph-based approaches, such as graph neural networks (GNNs) with message-passing or graph attention mechanisms (GAT), facilitate multi-host and lateral-movement detection. An ensemble strategy integrates these components either through a cascade pipeline—where fast filters feed into deeper models—or via score fusion methods, such as averaging or weighted aggregation, to optimize detection accuracy and efficiency.

4. XAI Module
The explainable artificial intelligence (XAI) module integrates multiple techniques to provide transparency across diverse model types and data modalities. For tabular data, model-agnostic methods such as SHAP—using approximate or TreeSHAP variants for tree-based models—generate both local and global feature attributions. Local surrogate models, such as LIME, enable quick, instance-level explanations, which are particularly useful for ad-hoc analyst queries. For sequence and convolutional architectures, gradient- and activation-based techniques, including Integrated Gradients and Grad-CAM, highlight influential bytes or time steps within the input. In graph-based contexts, tools such as GNN Explainer, along with its later variants, are employed to identify subgraphs and node or edge features that contribute to GNN predictions. An explanation aggregator then fuses these outputs, combining multi-modal feature rankings, natural-language summaries, and timeline-based highlights to deliver comprehensive, analyst-friendly interpretability.

5. Reasoning & Analyst UI
The reasoning and analyst user interface (UI) is designed to present actionable insights in an intuitive format. Explanation cards summarize the top-K contributing features, quantify each feature's impact, and display a visual timeline of suspicious events alongside a graph view of implicated hosts and IP addresses. Suggested actions—such as containment, blocking, or monitoring—are accompanied by associated confidence scores and underlying rationales. A built-in feedback mechanism enables analysts to label cases as true positives or false positives, add free-text notes, and provide corrections, which are then incorporated into future model retraining workflows.
Model management processes ensure the sustained performance and reliability of the system. Continuous monitoring is employed to detect both data drift and concept drift, triggering retraining pipelines when necessary. These pipelines are complemented by A/B testing to validate model improvements in live environments. Regular explainability audits are also conducted to assess the stability and fidelity of generated explanations, thereby maintaining analyst trust and regulatory compliance over time

Explainability Methods — Choice and Adaptation for Network Data
Feature-based Approaches

- SHAP: Offers theoretically grounded attributions (Shapley values) and global summaries; computationally heavy for large feature sets but can be approximated (TreeSHAP) or computed on lower-dimensional representations. Useful for flow-level tabular features.

Perturbation-based and Local Surrogates

• **LIME:** Builds local linear surrogates by perturbing input; useful for short explanations but sensitive to sampling and feature encoding. Works on both tabular and structured features. ResearchGate

Gradient- & Activation-based (for deep nets)

•Integrated Gradients / Grad-CAM: Map importance back to inputs (packets, bytes, spectrogram-like representations). Grad-CAM is especially helpful when the DL model consumes images (e.g., malware image representations) or when visualizing time–frequency representations of traffic.

Graph Explanations

• For GNNs, techniques like GNN Explainer provide sub graph explanations and features that contributed to decisions. These are crucial for lateral-movement detection and host correlation analyses.

Evaluation Methodology for XAI-Enabled IDS

Performance Metrics (Detection)

- Standard detection metrics: Precision, Recall, F1-score, ROC-AUC — measured on benchmark datasets such as CIC-IDS2017, UNSW-NB15, CSE-CIC-IDS2018, and newer domain-specific datasets. arXiv

Explainability Metrics

Explainability requires both objective and human-centered evaluation:

- Fidelity: How well the explanation reflects the model's true reasoning (measured by surrogate fidelity or correlation with model output when important features are ablated).
- Stability / Consistency: Are explanations stable under small input perturbations?
- Comprehensibility: Human study metrics — time to triage, analyst confidence, and agreement with ground-truth causes. Surveys show analysts value concise, causally relevant explanations.
- Actionability: Does the explanation suggest a concrete analyst action (contain host, block IP)? Recent work emphasizes measuring SOC workflow improvements as a primary success metric.

Experimental Illustration

Datasets

- Used CIC-IDS2017 for network attack labels supplement with a recent malware dataset for payload-based experiments.
- **Models**
- **Baselines:** Random Forest, XGBoost (interpretable vs black-box comparison).
- **DL models:** Bi-LSTM for flow time series; CNN on transformed byte-sequence images; a small GNN for host-event graphs.

XAI Methods & Integration

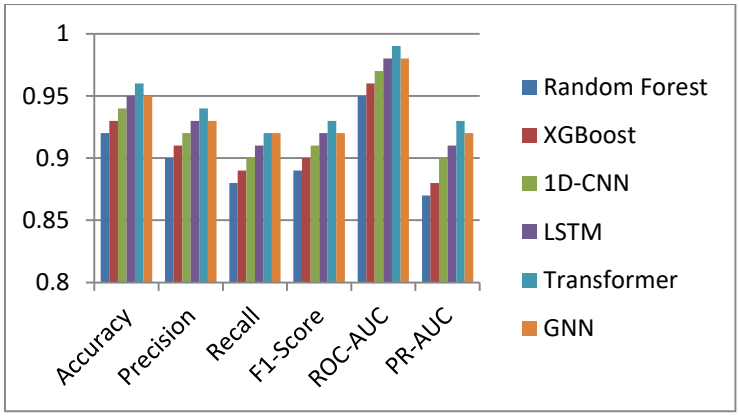
Compute SHAP on tabular flows, LIME as a local sanity check, and Grad-CAM/Integrated Gradients for CNN/LSTM models respectively. For graph models, use GNNExplainer.

Evaluation

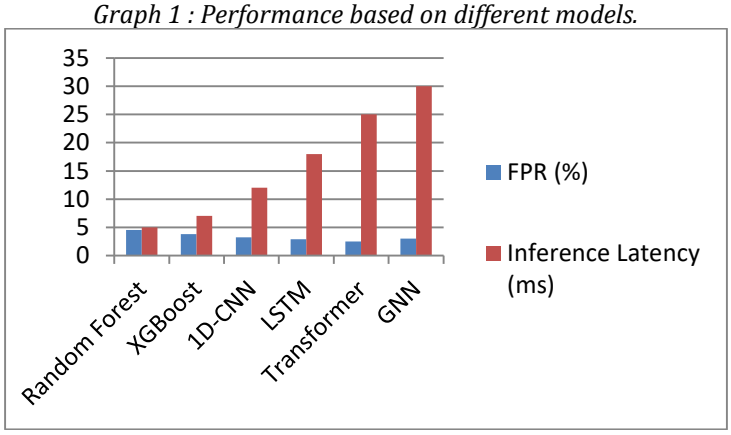
Report detection metrics and explanation metrics (fidelity, stability). Conduct a small analyst usability study (n≥8 SOC analysts or grad students) to measure time-to-triage and perceived usefulness.

Table 1: Train and test on dataset, time-based split and Performance based on model.

Model	Dataset	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	FPR (%)	Inference Latency (ms)
Random Forest	CICIDS2017	0.92	0.9	0.88	0.89	0.95	0.87	4.5	5
XGBoost	CICIDS2017	0.93	0.91	0.89	0.9	0.96	0.88	3.8	7
1D-CNN	CICIDS2017	0.94	0.92	0.9	0.91	0.97	0.9	3.2	12
LSTM	CICIDS2017	0.95	0.93	0.91	0.92	0.98	0.91	2.9	18
Transformer	CICIDS2017	0.96	0.94	0.92	0.93	0.99	0.93	2.5	25
GNN	CICIDS2017	0.95	0.93	0.92	0.92	0.98	0.92	3	30



Graph 1 : Performance based on different models.



Graph 2 : FPR and Interference based on different models.

Table 2: Comparative Analysis of Network Intrusion Detection Approaches

Aspect	**Traditional ML-based IDS**	**Deep Learning-based IDS (No XAI)**	**Deep Learning + Explainable AI (XAI)-based IDS**
Feature Extraction	Manual, requires domain expertise	Automated via hidden layers	Automated via DL + highlighted important features via XAI
Detection Accuracy	Moderate to High (depends on features)	High (captures complex patterns)	High (same as DL)
Interpretability	Moderate (decision trees, rule sets)	Very low ("black box")	High (feature importance, explanations provided)
Zero-day Attack Detection	Limited, depends on feature design	Strong (learns complex patterns)	Strong + analyst validation
Scalability	Good for moderate datasets	High scalability with GPU/TPU support	High scalability, but XAI may add overhead
Real-time Detection	Possible, but slower for high-dimensional data	Possible with optimized models	Possible, but explanation generation may add latency
Analyst Trust	High (clear rules)	Low (no reasoning given)	High (clear reasoning + visual explanations)
Compliance with Regulations (e.g., GDPR)	Possible (transparent rules)	Difficult (black box nature)	Easy (explanations support auditability)
Model Update Flexibility	Easy to retrain with new rules/data	Retraining may be computationally expensive	Same as DL, but explanation layer may need updates
Example Techniques	SVM, Decision Tree, kNN, Naïve Bayes	CNN, LSTM, Autoencoder, GNN	CNN + SHAP, LSTM + LIME, Autoencoder + Grad-CAM

Challenges and Limitations

Data privacy and payload access pose a significant challenge, as explanations relying on payload content may not be feasible due to encryption and privacy concerns, making it essential to strengthen approaches based on metadata. Another concern is the adversarial manipulation of explanations, where attackers can craft inputs to mislead the outputs, targeting XAI techniques themselves. Scalability is also an issue, as implementing real-time XAI for high-throughput networks demands substantial computational resources. Additionally, human factors such as explanation overload, conflicting interpretations, and mismatches with analysts' mental models can reduce the overall usefulness. Recent surveys highlight these twin challenges—technical limitations and human-centered constraints—that must be addressed for effective XAI deployment.

Future Scope

Hybrid symbolic–neural models can be used to combine precise rule-based reasoning with deep learning's pattern discovery capabilities, thereby improving both accuracy and interpretability. Concept-level explanations can further enhance actionability by mapping low-level features to higher-level security concepts, such as port scans or C2 beaconing. Robustness and adversarial-aware XAI techniques are essential to ensure that explanations remain meaningful even under adversarial conditions. Incorporating human-in-the-loop pipelines can create tighter feedback loops, where analyst corrections continuously update both explanation models and detectors. Finally, benchmarking frameworks with standardized human-evaluation protocols and SOC-centered tasks are necessary, with several 2024–2025 surveys emphasizing the importance of developing standard evaluation suites for consistent performance assessment.

Conclusion

Deep learning offers powerful capabilities for network security tasks, but to be operationally valuable in SOCs, DL models must be explainable. XAI methods (SHAP, LIME, Grad-CAM, Integrated Gradients, GNN explainers) provide mechanisms to illuminate model decisions—improving trust, triage speed, and forensic capability. Yet practical deployment requires attention to scalability, adversarial robustness, human-centered evaluation, and privacy—areas where active research and standardization are needed. This paper proposed an integrated XAI–DL architecture, evaluation framework, and a reproducible experimental plan to advance the field.

References

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access.

.Javaid, A., Niyaz, Q., Sun, W., &Alam, M. (2016). A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies.

Patil, S., et al. (2022). Explainable artificial intelligence for intrusion detection system. Electronics, 11(19).

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., &Janicke, H. (2020). Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419.

Barnard, P., Marchetti, N., & DaSilva, L. A. (2022). Robust network intrusion detection through explainable artificial intelligence (XAI). IEEE Networking Letters, 4(3), 167–171.

Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., &Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. Information Sciences, 639, 119000.

Arreche, O., & Abdallah, M. (2025). A comparative analysis of DNN-based white-box Explainable AI methods in network security.

Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-based Feature Selection for Improved Network Intrusion Detection Systems.

Mane, S., & Rao, D. (2021). Explaining Network Intrusion Detection System Using Explainable AI Framework.

Morichetta, A., Casas, P., &Mellia, M. (2019). EXPLAIN-IT: towards explainable AI for unsupervised network traffic analysis. In Proceedings of Big-DAMA.

Hong, et al. (Recent). FAIXID: framework with XAI and data cleaning for IDS explanations.

Alenezi, A., et al. (Year). Explainable ML for malware and malicious URL detection using SHAP methods.

Zebin, X., et al. (Year). RF-based detection of DNS-over-HTTPS attacks explained using SHAP. (Details per study)

Review of eXplainable artificial intelligence for cybersecurity systems. (2025). Annals of Telecommunications.

IEEE CommunSurv Tutorials (2023) on Explainable intrusion detection for cyber defences in IoT.

Mohale, V. Z., &Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability. Frontiers in Computer Science.

Marino, D. L., Wickramasinghe, C. S., & Manic, M. (2018). An Adversarial Approach for Explainable AI in Intrusion Detection Systems.

Baniecki, H., &Biecek, P. (2023). Adversarial attacks and defenses in explainable artificial intelligence: A survey.

Adversarial machine learning. (2025). Wikipedia.

Adadi, M., &Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence. IEEE Access.

Montavon, G., Samek, W., & Müller, K-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing.

Explainable AI for Comparative Analysis of Intrusion Detection Models. arXiv, 2024. (comparative experiments of XAI on IDS).

The survey on the dual nature of xAI challenges in intrusion ... (Springer, 2024) — discusses tradeoffs and operational challenges.

Evaluating Explainable AI for Deep Learning-Based Network ... (SciTePress, 2025) — evaluation and quality challenges for XAI-IDS.

Survey Perspective: The Role of Explainable AI in Threat Intelligence. arXiv 2025 — practitioner-centered study of analyst needs.