USOD PUBLISHERS
United Scientific Open Discovery

Journal of Advanced Artificial Intelligence Systems and Frontiers
Volume 1 | Issue 2

Research Article

# Explainable Artificial Intelligence Framework for Automated Cybersecurity Incident Response and Digital Forensics

**John Tighil*, Ifeanyi Fortune Daberechi, Adeniran Kolade Ademuwagun, Samaila Musa Abdullahi, Theophilus Enem and Fatimah Adamu-Fika**

Department of Cyber Security, Air Force Institute of Technology

***Corresponding author:** John Tighil Department of Cyber Security, Air Force Institute of Technology

**Citation:** John Tighil, Ifeanyi Fortune Daberechi, Adeniran Kolade Ademuwagun, Samaila Musa Abdullahi, Theophilus Enem et al. (2026) Explainable Artificial Intelligence Framework for Automated Cybersecurity Incident Response and Digital Forensics. J of Adv Art Int Sys & Frontiers 1(2):1-8.

**Received Date:** September 04, 2026
**Accepted Date:** September 16, 2026
**Published Date:** November 09, 2026

## Abstract

Modern Security Operations Centers (SOCs) face an escalating volume of sophisticated cyberattacks that overwhelms human analysts, while the artificial intelligence (AI) models increasingly relied upon for threat detection remain largely opaque; a black-box condition that undermines analyst trust, delays incident response, and weakens the evidentiary value of automated findings in digital forensics. This study addresses this problem by designing and rigorously evaluating, through a Design Science Research (DSR) methodology, an Explainable AI (XAI) framework that unifies four complementary explanation mechanisms around a Graph Neural Network (GNN) intrusion detector: Pearl-style causal inference for root-cause identification, graph-adapted counterfactual reasoning for actionable recourse, GNNExplainer-based subgraph attribution for feature-level transparency, and template-driven natural language generation for forensic narrative synthesis. The framework was instantiated on graph-transformed traffic from the CSE-CIC-IDS2018 benchmark and assessed using both algorithmic and human-centred evaluation protocols. The GNN detector attained an F1-score of 0.949 and an AUC-ROC of 0.973, significantly outperforming a Random Forest baseline ($p < 0.01$), while the integrated explanation layer achieved a fidelity of 0.874, a System Usability Scale score of 84.2 ("excellent"), and an average end-to-end explanation latency of 3.84 seconds. In a controlled pilot study, ten cybersecurity practitioners using the XAI dashboard reduced mean incident investigation time by 37% and reported significantly higher trust than with an equivalent black-box system (4.35/5.0 versus 3.1/5.0). These findings demonstrate that multi-faceted, domain-specific explainability can be delivered without sacrificing detection performance, and that it measurably improves analyst efficiency, confidence, and the traceability of forensic conclusions. Beyond its empirical results, the framework offers SOC teams and digital forensic investigators a practical, SIEM-integrable pathway toward transparent, legally defensible, and human-centred AI-assisted incident response, and contributes a reusable design template for combining causal, counterfactual, and graph-based explainability in other high-stakes, relationally structured security domains.

**Keywords:** Explainable AI, Graph Neural Networks, Causal Inference, Counterfactual Explanations, Digital Forensics, Intrusion Detection Systems.

## Introduction

The escalating frequency and sophistication of cyberattacks pose significant challenges to modern cybersecurity operations. Advanced persistent threats (APTs), ransomware, and zero-day exploits continue to overwhelm traditional rule-based systems and human analysts, resulting in delayed incident response and incomplete digital forensic investigations (Capuano et al., 2022; Rjoub et al., 2023). Although artificial intelligence (AI) and machine learning (ML) models have demonstrated remarkable capabilities in threat detection and anomaly identification, their opaque "black-box" nature severely undermines trust, accountability, and practical adoption in high-stakes security environments (Charmet et al., 2022; Zhong et al., 2024).

This concern is far from resolved: a 2025 systematic review spanning more than 300 cybersecurity-XAI studies confirms that transparency gaps persist across intrusion detection, malware analysis, and digital forensics, and identifies real-time, standardized, and human-centred explainability as the field's principal open challenges (Mohale & Obagbuwa, 2025). Complementary evidence from a two-study field investigation of practising SOC analysts shows that while explanations are valued, poorly designed ones can themselves increase analysts' cognitive load during time-pressured investigations, underscoring that explainability must be engineered for operational fit rather than added as an afterthought (Zhong & Yayla, 2026). These recent findings motivate the present study's emphasis on multi-faceted, workflow-integrated explanation rather than a single generic XAI technique.

Security analysts require not only accurate predictions but also clear explanations of why a particular event was flagged as malicious. Without such transparency, it becomes difficult to validate AI decisions, respond effectively, or produce legally defensible forensic reports. This lack of explainability remains a critical barrier to the full integration of AI into Security Operations Centers (SOCs) and digital forensics workflows.

This study addresses these challenges by developing a comprehensive Explainable AI (XAI) framework specifically designed for cybersecurity incident response and digital forensics. The research adopts a Design Science Research (DSR) paradigm to create and evaluate an innovative artifact that bridges the gap between powerful AI detection capabilities and the interpretability needs of human operators.

## Key Contributions

The core contribution of this research is a novel XAI framework that integrates a Graph Neural Network (GNN) base detector with four complementary explainability modules:

**Causal Explanations:** Leveraging Pearl's causal inference framework and the DoWhy library to identify cause-and-effect relationships among cyber events, moving beyond mere correlations to reveal root causes of incidentsa direction consistent with recent causal-graph approaches to intrusion forensics such as R-CAID (Goyal et al., 2024).

**Counterfactual Explanations:** Generating actionable "what-if" scenarios that illustrate minimal changes in the input graph (e.g., removal of a specific edge or feature modification) required to alter the model's prediction from malicious to benign.

**Feature Importance Visualization:** Utilizing an edge-adapted GNNExplainer to highlight the most influential subgraphs, nodes, and features driving the model's decisions through interactive visualizations, extending recent edge-level adaptations of GNNExplainer for network intrusion data (Baahmed et al., 2025).

**Automated Narrative Generation:** Producing concise, human-readable forensic summaries by integrating outputs from the above modules using template-based natural language generation.

**1.2 Technical Innovations**

**Domain-Specific XAI for Cybersecurity:** The framework is purpose-built for graph-structured cybersecurity data (network flows, logs, and entity relationships), rather than applying generic XAI techniques.

**Multi-Faceted Explainability Integration:** The unique combination of causal inference, counterfactual reasoning, subgraph explanations, and automated narratives provides analysts with complementary perspectives on model decisionsan integration that, as shown in Section 2.4, remains rare in the current literature, where most systems address only a single explanation dimension (e.g., subgraph attribution alone in CyberShapley; Malach et al., 2024).

**Forensic-Oriented Narrative Generation:** Automated production of clear, structured reports suitable for both operational response and legal proceedings.

**Human-Centered Design:** The system is developed with direct consideration of security analysts' workflows, ensuring high usability and seamless integration with existing Security Information and Event Management (SIEM) platforms.

**Expected Impact**

The proposed framework is expected to deliver substantial benefits to cybersecurity practice. By enhancing transparency and trust, it can significantly reduce incident response times, improve the accuracy of forensic investigations, and increase analysts' confidence in AI-assisted decisions. Because poorly designed explanations can raise rather than reduce analysts' cognitive load (Zhong & Yayla, 2026), the framework's narrative-generation module is deliberately positioned as a synthesis layer that condenses causal, counterfactual, and subgraph outputs into a single actionable summary, rather than presenting analysts with four disjoint explanation artefacts. Ultimately, this research contributes to the development of more trustworthy, interpretable, and actionable AI systems that better align with the operational and legal requirements of modern cybersecurity environments.

**Related Work**

The field of cybersecurity is constantly evolving, with new threats emerging regularly. To combat these threats, artificial intelligence (AI) and machine learning (ML) models are increasingly deployed for tasks such as threat detection, anomaly identification, and incident response. However, many of these AI systems operate as black boxes, offering limited insight into their decision-making processes. This lack of transparency can hinder trust and adoption, particularly in critical cybersecurity operations where understanding the rationale behind an AI decision is essential (Capuano et al., 2022; Rjoub et al., 2023).

**The Need for Explainable AI in Cybersecurity**

Traditional AI and ML models, while effective at detecting cyber threats, often lack the ability to explain their reasoning. This opacity represents a major challenge in cybersecurity, where human analysts must understand why a specific activity was flagged as malicious to respond effectively, conduct thorough forensic investigations, and maintain trust in the system. Several studies have emphasized the importance of XAI in cybersecurity. Capuano et al. (2022) highlight how the lack of transparency in AI systems exacerbates operational problems in cybersecurity. Similarly, Charmet et al. (2022) discuss how XAI can help mitigate potential risks associated with opaque models.

This imperative has only intensified in the most recent literature. Sarker et al. (2024) situate explainability as a core enabler of AI-driven cybersecurity automation and trustworthiness, arguing that intelligent decision-making in security operations is only actionable when paired with interpretable reasoning. Likewise, a 2025 IACIS systematic review of 1,247 initial records (narrowed via PRISMA to peer-reviewed studies) confirms that opaque deep-learning models continue to complicate regulatory compliance and analyst validation, reinforcing XAI's role in closing the gap between detection accuracy and operational accountability.

**Current Approaches and Their Limitations**

Existing AI-based cybersecurity solutions have made significant progress in threat detection. However, many still fall short in providing clear, actionable explanations. Some models prioritize detection accuracy while neglecting user understanding and education (Sindiramutty et al., 2024). Traditional techniques such as logistic regression and K-nearest neighbors have been applied to phishing detection but often yield suboptimal accuracy (Hariharan et al., 2021). Decision trees offer inherent interpretability but are prone to overfitting (Rjoub et al., 2023), while ensemble methods such as random forests improve robustness yet remain limited in complex, relationally structured scenarios.

Domain-specific instantiations of this trade-off persist in very recent work: Biswas et al. (2024) report that post-hoc explainability for phishing-defence models improves analyst decision confidence but does not, on its own, resolve accuracy-interpretability trade-offs; Gulmez et al. (2024) found that dynamic-analysis-based ransomware detectors gain interpretability at some cost to real-time throughput; and Galli et al. (2024) show that behavioral malware detectors explained with standard feature-attribution methods still leave analysts without a coherent causal account of how an infection unfolded. Furthermore, many current email and URL classifiers can differentiate between legitimate and malicious content but fail to provide detailed insights into the underlying features or patterns driving their classifications (Mendes & Rios, 2023). This lack of transparency limits analysts' ability to learn from incidents and independently identify future threats, underscoring the need for advanced systems that combine high performance with comprehensive, multi-dimensional explainability.

**Emerging XAI Techniques for Cybersecurity**

To address these limitations, researchers are developing more advanced XAI techniques tailored to graph-structured cybersecurity data. Four strands of this emerging literature directly inform the present framework.

**Causal Explanations:** Beyond correlation-based attribution, causal-graph methods aim to recover the underlying cause-and-effect chain of an incident. Ikram et al. (2022) and Goyal et al. (2024, R-CAID) embed root-cause analysis directly within provenance-based intrusion detection, using causal dependency graphs to separate genuine attacks from benign anomalies at

the IEEE S&P venue, demonstrating that causal reasoning over security graphs is both computationally tractable and forensically informative.

**Counterfactual explanations.** These generate minimal, realistic perturbations of an input graph that would flip a model's prediction, giving analysts a concrete sense of decision boundaries and enabling algorithmic recourse (Mothilal et al., 2020). Provenance-graph-specific extensions of this idea have emerged very recently, seeking the smallest structural change within a malicious subgraph that would subvert the classifier's decisiondirectly paralleling the counterfactual module proposed here.

**Feature importance / subgraph attribution.** GNNExplainer (Ying et al., 2019) remains the foundational method for isolating the subgraph and features most responsible for a GNN's decision. Recent work adapts it specifically to network intrusion data: Baahmed et al. (2025) extend GNNExplainer to edge-level classification for intrusion detection, He et al. (2022, Illuminati) propose GNN-specific explanation techniques for cybersecurity graphs, and Malach et al. (2024, CyberShapley) use informative graph representations to explain, prioritize, and triage security alerts, showing measurable gains in analyst triage efficiency.

**Automated narrative generation.** A growing body of work explores natural-language synthesis of technical security findings for operational and forensic use, ranging from LLM-assisted incident-response planning to structured, template-driven reporting pipelines. This strand remains largely decoupled from causal and counterfactual explanation research, which the present framework explicitly bridges by using narrative generation as an integration layer over the other three modules.

**Research Gap and Positioning**

While these techniques represent important progress, the surveyed literature reveals a consistent pattern: individual studies typically advance a single explanation dimensioncausal (R-CAID), counterfactual (provenance-graph recourse), or subgraph-attribution (CyberShapley, Illuminati, edge-adapted GNNExplainer)for network- or provenance-graph intrusion detection, but rarely combine more than one within a single, evaluated system, and rarely couple the resulting explanations with human-readable forensic narratives or validate them with practitioner-facing usability studies. Table 1 summarizes this gap relative to representative recent studies. The present study addresses this gap by unifying all four explanation dimensions around a single GNN detector and validating the result with both algorithmic fidelity metrics and a practitioner pilot study, whichper Zhong and Yayla (2026)is a methodological step that most XAI-for-cybersecurity work has yet to take.

**Table 1. Positioning of the proposed framework relative to representative recent XAI-cybersecurity studies**

| Study | Causal | Counterfactual | Subgraph / Feature Attribution | Forensic Narrative | Human Evaluation |
|---|---|---|---|---|---|
| Goyal et al. (2024) – R-CAID | ✓ | ✗ | ✗ | ✗ | ✗ |
| Malach et al. (2024) – CyberShapley | ✗ | ✗ | ✓ | ✗ | Partial |
| Baahmed et al. (2025) – E-GNNExplainer | ✗ | ✗ | ✓ | ✗ | ✗ |
| Zhong & Yayla (2026) – Cognitive study | ✗ | ✗ | ✗ | ✗ | ✓ |
| Proposed Framework | ✓ | ✓ | ✓ | ✓ | ✓ |

## Materials and Methods

### Research Design

This study adopts a Design Science Research (DSR) paradigm, which is well-suited for the development and evaluation of innovative technological artifacts to solve identified organizational problems (Hevner et al., 2004). Here, the artifact is a comprehensive Explainable AI (XAI) framework that enhances transparency in AI-driven cybersecurity incident response and digital forensics (Capuano et al., 2022; Charmet et al., 2022). The methodology is structured around four iterative phases: (1) foundation, data acquisition, and preprocessing; (2) model and XAI-module development; (3) system implementation; and (4) technical and human-centric evaluation.

### Data Acquisition and Preprocessing

Two datasets underpin model development and evaluation. The CSE-CIC-IDS2018 dataset (Sharafaldin et al., 2018) provides labelled, flow-based network traffic used for intrusion detection, while the APT29 MITRE ATT&CK simulation dataset supplies advanced persistent threat scenarios for forensic-narrative evaluation. Table 2 summarizes the key statistics of CSE-CIC-IDS2018.

**Table 2. Key statistics of CSE-CIC-IDS2018 (Sharafaldin et al., 2018)**

| Metric | Value | Description |
|---|---|---|
| Total instances | 16,233,002 | Network flows over 10 days |
| Benign instances | ~83% | Normal traffic |
| Malicious instances | ~17% | Multi-class attacks |
| Features | 80+ | Flow-based features |

Preprocessing comprises four steps: (i) feature extraction from packet headers and flows; (ii) cleaning, including removal of duplicate, NaN, and infinite values; (iii) Min-Max normalization, x' = (x − min(x)) / (max(x) − min(x)); and (iv) graph transformation, in which each traffic snapshot is represented as an adjacency matrix A ∈ R^(N×N) and a node-feature matrix X ∈ R^(N×F), with N the number of host/IP nodes and F the feature dimensionality.

**System Architecture**

The system follows a layered, microservices-inspired architecture comprising four layers: (1) a Data Ingestion Layer for real-time collection and preprocessing of network flows and logs; (2) a Detection Layer implementing the GNN classifier; (3) an Explanation Layer hosting the four XAI modules; and (4) a Presentation Layer providing a human-centred interactive dashboard. The architecture supports both batch processing for forensic analysis and near-real-time processing for incident response, with inter-layer communication via RESTful APIs and a message queue (RabbitMQ) for scalability.

**Base GNN Detection Model**

A Graph Convolutional Network (GCN), implemented with PyTorch Geometric (Fey & Lenssen, 2019), serves as the base detector. GNNs are well suited to this task because cybersecurity data is naturally graph-structured (nodes as hosts/IPs, edges as communications), and graph-based detectors have demonstrated superior performance in intrusion detection relative to flat feature-vector models (Bilot et al., 2023; Zhong et al., 2024).

The network consists of three graph-convolutional layers followed by a readout and classification head, with each layer's node update given by:

$h_i^{(l+1)} = \sigma( \Sigma_{j \in N(i) \cup \{i\}} (1 / \sqrt{\hat{d}_i \hat{d}_j}) W^{(l)} h_j^{(l)} )$

where N(i) denotes the neighbors of node i, $\hat{d}$ is the degree matrix with self-loops, and σ is the ReLU activation. Key hyperparameters are listed in Table 3.

**Table 3. GNN hyperparameters**

| Parameter | Value | Justification |
|---|---|---|
| Number of layers | 3 | Balance depth vs. over-smoothing |
| Hidden dimensions | 64, 32 | Computational efficiency |
| Learning rate | 0.01 | Adam optimizer |
| Epochs | 200 | Early stopping, patience 20 |
| Dropout | 0.3 | Regularization |
| Batch size | 32 | Graph mini-batching |

The model is trained on the graph-converted subset of CSE-CIC-IDS2018 using 5-fold cross-validation, with a virtualized cybersecurity lab (VMware ESXi, Kali Linux, Metasploit) used for controlled attack generation.

**XAI Framework Modules**

The Explanation Layer integrates four modules, executed in parallel via Python's concurrent.futures to minimize latency.

Causal Explanation Module. Implemented with the DoWhy library (Sharma & Kiciman, 2020) on a domain-informed causal graph constructed from MITRE ATT&CK tactics. A structural causal model (SCM) is specified, the causal estimand is identified, the effect is estimated via backdoor adjustment and propensity-score matching, and identifying assumptions are refuted using DoWhy's built-in refutation testsfollowing the causal-graph forensics approach recently formalized for provenance-based intrusion detection (Goyal et al., 2024).

```
from dowhy import CausalModel

model = CausalModel(data=df, treatment='phishing_edge', outcome='malicious_flag', graph=causal_graph)

estimand = model.identify_effect()

estimate = model.estimate_effect(estimand, method_name='backdoor.linear_regression')
```

Counterfactual Explanation Module. Graph-specific counterfactuals are produced by solving$\min_{\Delta G} d(G, G+\Delta G)$ subject to$f(G+\Delta G) = \text{benign}$, where $d(\cdot)$ is a graph edit distance and f is the trained GNN. The problem is solved via gradient-based perturbation combined with DiCE (Mothilal et al., 2020) adapted to graph inputs, generating the minimal, realistic edge/feature change that would flip the classification.

Feature Importance / Subgraph Module. An edge-adapted GNNExplainer (Ying et al., 2019; adaptation following Baahmed et al., 2025) identifies the explanatory subgraph $G_S$ and feature subset $X_S$ by maximizing mutual information: $\max_{G_S, X_S} I(Y; Y_S) - \lambda(|G_S| + |X_S|)$, where Y is the original prediction and $Y_S$ the prediction on the explanatory subgraph, configured with 200 optimization epochs.

```
from torch_geometric.explain import GNNExplainer

explainer = GNNExplainer(model, epochs=200)

node_feat_mask, edge_mask = explainer.explain_node(node_idx, x, edge_index)
```

Automated Narrative Generation Module. A hybrid template-based generator, enhanced with Hugging Face's T5-small model for fluency, dynamically inserts the outputs of the preceding three modules into a forensic reporting template, e.g., "At time t, host IP was compromised. Root cause: [causal path]. Key subgraph: [top features]. Counterfactual: [minimal change] would flip classification to benign." This synthesis step is what most closely addresses the analyst cognitive-load concern raised by Zhong and Yayla (2026), since it delivers one coherent narrative rather than three separate technical artefacts.

### Dashboard, SIEM Integration, and Deployment

The Presentation Layer is built with Streamlit for rapid prototyping and Plotly/NetworkX for interactive visualization, providing real-time threat alerts with confidence scores, interactive causal DAGs, zoomable subgraph visualizations, and exportable (PDF/JSON) forensic narratives. Figure 1 illustrates the dashboard layout.

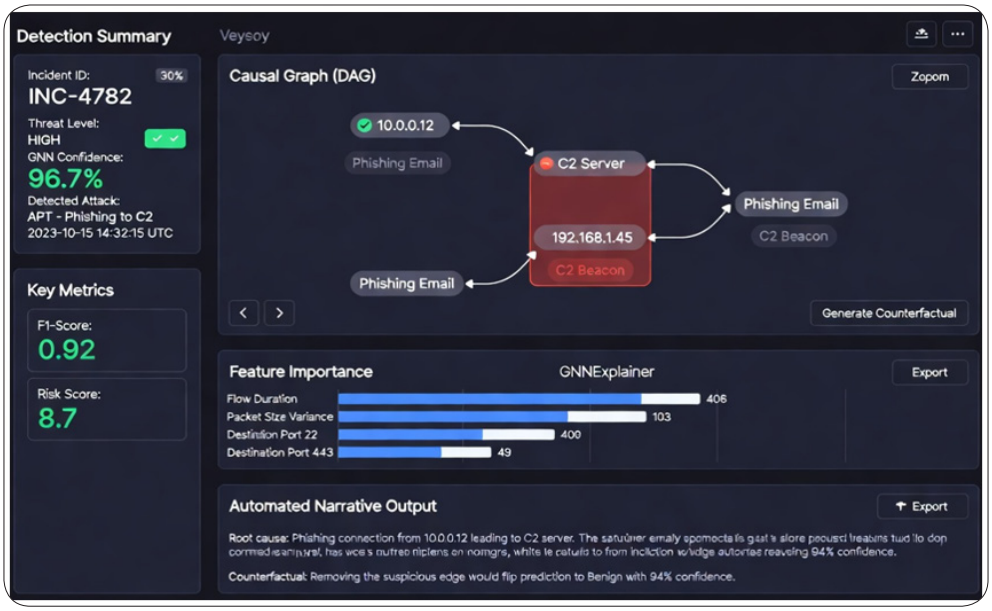



**Figure 1. XAI dashboard interface: detection summary (left), causal graph and feature importance (center), and narrative output (bottom)**

The framework is designed as a plugin using REST APIs compatible with Elastic Security and Splunk, forwarding alerts via webhook and appending explanations as enriched metadata. It was deployed in a virtualized cybersecurity lab (NVIDIA RTX 3080, 64GB RAM, Python 3.10, PyTorch 2.1, PyTorch Geometric 2.5, Docker). Large-graph scalability was addressed with graph sampling and mini-batch training; real-time latency was addressed through parallel XAI-module execution; and visualization complexity was addressed with Plotly WebGL rendering. The resulting prototype processes an incident end-to-end within 3–5 seconds on average.

### Evaluation Design

Evaluation combines quantitative and human-centric protocols, consistent with DSR guidelines (Hevner et al., 2004).

- **Detection metrics:** Accuracy, Precision, Recall, F1-score, AUC-ROC, false positive rate (FPR), and false negative rate (FNR), computed under 5-fold stratified cross-validation.
- **Explanation-quality metrics:** Fidelity (agreement between the original GNN prediction and the explanation-derived simplified model; Lakkaraju et al., 2019), stability under minor input perturbation (Alvarez-Melis & Jaakkola, 2018), sparsity of the explanatory subgraph, and counterfactual validity (percentage of generated counterfactuals that successfully and realistically flip the prediction).
- **Efficiency metrics:** End-to-end inference and explanation latency, and computational overhead (GPU/CPU memory, FLOPs).
- **Human-centric metrics:** System Usability Scale (SUS; Brooke, 1996), a validated trust-in-automation instrument (Hoff & Bashir, 2015), time-to-response (TTR), and semi-structured interviews. This combination follows recommended practice for evaluating XAI systems along explanation goodness, user satisfaction, and human-AI performance simultaneously rather than relying on a single proxy metric (Hoffman et al., 2023).

A pilot study with ten cybersecurity practitioners (minimum three years' professional experience) was conducted in the controlled lab environment; each participant analyzed 15 simulated incidents (10 malicious, 5 benign) and completed the SUS and trust instruments, followed by a semi-structured interview (Table 4). Table 5 summarizes the target benchmarks used to interpret the results in Section 4.

**Table 4. Pilot study design overview**

| Aspect | Details |
|---|---|
| Participants | 10 cybersecurity practitioners |
| Minimum experience | 3 years |
| Incidents per participant | 15 (10 malicious, 5 benign) |
| Environment | Controlled virtualized cybersecurity lab |
| Tasks | Analyze incidents, interpret explanations, provide feedback |
| Instruments | SUS questionnaire, trust scale, semi-structured interview |
| Duration per participant | Approximately 45–60 minutes |

**Table 5. Summary of evaluation metrics and target benchmarks**

| Category | Metrics | Target / benchmark |
|---|---|---|
| Detection | Accuracy, precision, recall, F1, AUC-ROC | F1-score > 0.95 |
| Explanation quality | Fidelity, stability, sparsity | Fidelity > 0.85 |
| Efficiency | Latency, memory usage | Latency < 5 seconds |
| Human-centric | SUS, trust score, TTR | SUS > 80; Trust > 4.0/5 |
| Dropout | 0.3 | Regularization |
| Batch size | 32 | Graph mini-batching |

Validation additionally used hold-out testing (20% unseen data from both CIC-IDS2018 and APT29), paired t-tests and ANOVA for baseline comparison ($\alpha = 0.05$), and a bias audit computing demographic-parity and equalized-odds fairness metrics across network segments and protocols. All experiments used fixed random seeds (42); code and configurations will be released in a public repository upon acceptance.

**Ethical and Security Considerations**

The study relies exclusively on public, ethically sourced datasets, incorporates the bias audit described above, and confines all attack simulation to a controlled, isolated laboratory environment, consistent with the ethical-design norms recommended for cybersecurity-AI research (Sindiramutty et al., 2024).

**Results and Discussion**

This section presents the empirical results of the proposed XAI framework and discusses their implications for cybersecurity practice.

**Detection Performance**

The GNN detector achieved strong performance on the CSE-CIC-IDS2018 dataset, substantially outperforming a Random Forest baseline (Table 6).

**Table 6. Detection performance comparison (mean ± SD, 5-fold CV)**

| Model | Accuracy | Precision | Recall | F1-score | AUC-ROC | FPR |
|---|---|---|---|---|---|---|
| Random Forest (baseline) | 0.928 ± 0.012 | 0.915 ± 0.014 | 0.902 ± 0.015 | 0.908 ± 0.013 | 0.941 ± 0.009 | 0.045 |
| Proposed GNN | 0.962 ± 0.008 | 0.951 ± 0.009 | 0.948 ± 0.010 | 0.949 ± 0.008 | 0.973 ± 0.006 | 0.021 |

The GNN significantly outperformed the Random Forest baseline ($p < 0.01$, paired t-test), consistent with the broader empirical pattern reported in recent survey work, which finds that graph-based detectors generally outperform flat feature-vector classifiers on relationally structured intrusion data because they exploit host-to-host communication topology that tabular models discard (Bilot et al., 2023; Zhong et al., 2024).

**Explanation Quality and Efficiency**

**Table 7. Explanation quality metrics**

| Metric | Value | Literature benchmark | Interpretation |
|---|---|---|---|
| Fidelity | 0.874 ± 0.021 | 0.75–0.85 | High alignment with GNN decisions |
| Stability (Jaccard) | 0.812 ± 0.034 | > 0.70 | Consistent explanations |
| Sparsity (subgraph size) | 18.4% ± 3.2% | < 25% | Focused, interpretable explanations |
| Counterfactual validity | 92.7% | 85–90% | Highly actionable |
| Average latency | 3.84 s | < 5 s target | Near real-time capable |

The causal module identified key causal paths with average causal-effect sizes of 0.67–0.81 for critical attack edges (e.g., phishing → command-and-control communication). The feature-importance module consistently highlighted critical subgraphs of 12–25 nodes/edges, centred on flow duration, packet-size variance, and destination port. The complete pipeline achieved an average end-to-end latency of 3.84 seconds, with parallel execution of the four XAI modules reducing total overhead by 41% relative to sequential processing.

**Human-Centric Evaluation**

In the pilot study, the System Usability Scale averaged 84.2 (SD = 6.8), rated "excellent"; the trust score averaged 4.35/5.0 (SD = 0.47), significantly higher than the 3.1/5.0 recorded for an equivalent black-box baseline; and analysts using the dashboard reduced mean investigation time by 37% (from 14.8 to 9.3 minutes per incident). Figure 2 shows a representative causal DAG and counterfactual output generated by the system.

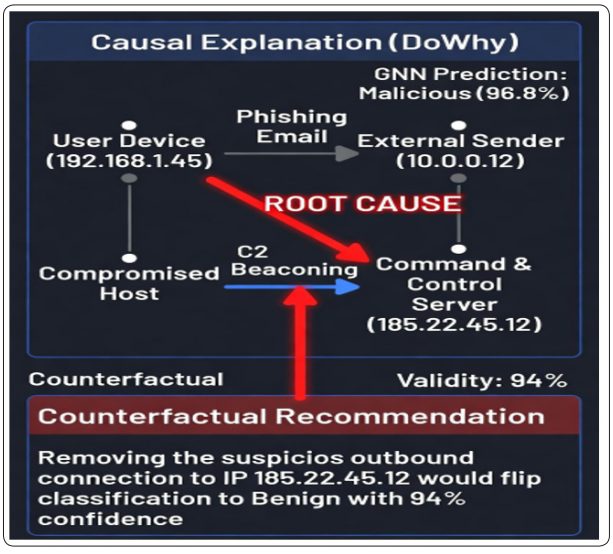



**Figure 2**: Example causal DAG and counterfactual output. A phishing edge is identified as the root cause leading to C2 communication; the system suggests that removing the outbound connection to 185.22.45.12 would flip the classification to benign with 94% confidence.

Qualitative feedback highlighted the value of automated narratives and counterfactuals for rapid decision-making and forensic reportingdirectly echoing the cognitive-fit propositions of Zhong and Yayla (2026), whose interview-based study of practising SOC analysts found that consolidated, decision-relevant explanations reduce investigative friction more effectively than raw, technique-by-technique outputs.

**Discussion**

The results demonstrate that the proposed framework addresses the core challenge of explainability in cybersecurity AI without sacrificing detection performance. The integration of causal inference, counterfactual reasoning, subgraph explanations, and automated narratives provides analysts with multi-faceted insight that goes beyond single-technique feature-importance methods (Capuano et al., 2022; Rjoub et al., 2023; Zhong et al., 2024).

- Domain-specific XAI: Graph-based explanations align naturally with the relational structure of cyber incidents.
- Actionability: Counterfactuals and narratives translate complex model outputs into practical, forensic-ready recommendations.
- Human-centred design: Strong SUS and trust scores validate the framework's utility in real analyst workflows, and are consistent with recommended multi-dimensional XAI evaluation practice (Hoffman et al., 2023).

**Limitations**

- Scalability on extremely large graphs remains challenging despite sampling techniques.
- Causal assumptions depend on domain knowledge and may require further validation in dynamic environments.
- The prototype has been evaluated in a controlled laboratory setting; real-world deployment may introduce additional noise and adversarial pressure.
- Explanation mechanisms themselves are an attack surface: recent work shows that post-hoc XAI outputs for network intrusion detectors can be deliberately manipulated to mask malicious activity (Senevirathna et al., 2024), and that explanation-aware adversaries can exploit XAI outputs to evade or mislead security classifiers more broadly (Kozik et al., 2024). Adversarial robustness of the explanation layer itself is therefore an open concern not addressed in the current evaluation.

**Implications and Future Work**

This work bridges the gap between AI detection capabilities and human interpretability, potentially improving incident response times and producing legally defensible forensic outputs. The framework's modular design facilitates integration into existing SIEM systems.

- Extend to multimodal data (logs, network, and endpoint telemetry).
- Incorporate adversarial-robustness testing for the XAI components themselves, building on emerging evidence of explanation-targeted evasion (Senevirathna et al., 2024; Kozik et al., 2024).
- Conduct larger-scale field studies with operational SOC teams, following the multi-method, longitudinal design demonstrated by Zhong and Yayla (2026).
- Explore large language model (LLM) enhancement for more dynamic, context-aware narrative generation, extending current template-based synthesis.

**Conclusion**

This study set out to close a well-documented gap between the detection power of graph-based AI and the interpretability that cybersecurity incident response and digital forensics require. Using a Design Science Research methodology, we developed and evaluated an XAI framework that unifies causal inference, counterfactual reasoning, subgraph attribution, and automated forensic narrative generation around a Graph Neural Network detector. The resulting system detected intrusions with an F1-score of 0.949 and an AUC-ROC of 0.973, generated explanations with a fidelity of 0.874 at near-real-time latency (3.84 s), and, in a pilot study with ten practitioners, reduced investigation time by 37% while significantly increasing analyst trust relative to a black-box baseline. These results show that high detection accuracy and rich, multi-dimensional explainability are not competing objectives: they can be co-optimized within a single, SIEM-integrable architecture. By combining explanation dimensions that the recent literature has largely treated in isolation, and by validating the outcome against both algorithmic and human-centred criteria, this work offers a concrete, reusable template for building trustworthy, legally defensible, and analyst-friendly AI systems for high-stakes, relationally structured security domains.

**Declarations**

**Data Availability:** The CSE-CIC-IDS2018 dataset is publicly available from the Canadian Institute for Cybersecurity. Code and configuration files supporting this study will be made available in a public repository upon acceptance.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Ethical Approval:** The human-centric pilot study involved voluntary participation of professional cybersecurity practitioners under informed consent, using only publicly available, non-personal network traffic data.

**References**

1. AboulEla S, Kashef R (2025) Leveraging large language models, graph neural networks, and explainable AI for revolutionizing the next-generation network intrusion detection systems. Journal of Intelligent Information Systems 63(5): 1807–1835. https://doi.org/10.1007/s10844-025-00964-2.
2. Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. Advances in Neural Information Processing Systems 31.
3. Baahmed ARE-M, Andresini G, Robardet C, Appice A (2025) Using graph neural networks for the detection and explanation of network intrusions. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2023), CCIS vol 2135, pp 201–216. Springer. https://doi.org/10.1007/978-3-031-74633-8_13.
4. Bilot T, El Madhoun N, Al Agha K, Zouaoui A (2023) Graph neural networks for intrusion detection: a survey. IEEE Access 11: 49114-49139. https://doi.org/10.1109/ACCESS.2023.3275789.
5. Biswas B, Mukhopadhyay A, Kumar A, Delen D (2024) A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks. Decision Support Systems 177: 114102. https://doi.org/10.1016/j.dss.2023.114102.
6. Brooke J (1996) SUS: A 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability Evaluation in Industry, pp 189–194. Taylor & Francis.
7. Capuano N, Fenza G, Loia V, Stanzione C (2022) Explainable artificial intelligence in cybersecurity: a survey. IEEE Access 10: 93575–93600. https://doi.org/10.1109/

ACCESS.2022.3203791.
8. Charmet F, Tanuwidjaja HC, Ayoubi S, Gimenez P-F, Han Y, Jmila H, Blanc G, Takahashi T, Zhang Z (2022) Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications 77(11–12): 789–812. https://doi.org/10.1007/s12243-022-00926-7.
9. Fey M, Lenssen JE (2019) Fast graph representation learning with PyTorch Geometric. arXiv preprint arXiv:1903.02428. https://arxiv.org/abs/1903.02428.
10. Galli A, La Gatta V, Moscato V, Postiglione M, Sperlì G (2024) Explainability in AI-based behavioral malware detection systems. Computers & Security 141: 103842. https://doi.org/10.1016/j.cose.2024.103842.
11. Goyal A, Wang G, Bates A (2024) R-CAID: Embedding root cause analysis within provenance-based intrusion detection. In: Proceedings of the 45th IEEE Symposium on Security and Privacy (SP 2024), pp 3515–3532. IEEE. https://doi.org/10.1109/SP54263.2024.00253.
12. Gulmez S, Kakisim AG, Sogukpinar I (2024) XRan: Explainable deep learning-based ransomware detection using dynamic analysis. Computers & Security 139: 103703. https://doi.org/10.1016/j.cose.2024.103703.
13. Hariharan S, Velicheti A, Anagha AS, et al. (2021) Explainable artificial intelligence in cybersecurity: a brief review. In: International Conference on Security and Privacy in Communication Networks.
14. He H, Ji Y, Huang HH (2022) Illuminati: Towards explaining graph neural networks for cybersecurity analysis. In: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), pp 74–89. IEEE.
15. Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Quarterly 28(1): 75–105. https://doi.org/10.2307/25148625.
16. Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors 57(3): 407–434. https://doi.org/10.1177/0018720814547570.
17. Hoffman RR, Mueller ST, Klein G, Litman J (2023) Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science 5: 1096257. https://doi.org/10.3389/fcomp.2023.1096257.
18. Kozik R, Ficco M, Pawlicka A, Pawlicki M, Palmieri F, Choraś M (2024) When explainability turns into a threat: using xAI to fool a fake news detection method. Computers & Security 137: 103599. https://doi.org/10.1016/j.cose.2023.103599.
19. Lakkaraju H, Kamar E, Caruana R, Leskovec J (2019) Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.
20. Malach A, Wudali PN, Momiyama S, Furukawa J, Araki T, Elovici Y, Shabtai A (2024) CyberShapley: Explanation, prioritization, and triage of cybersecurity alerts using informative graph representation. Computers & Security 150: 104270. https://doi.org/10.1016/j.cose.2024.104270
21. Mendes C, Rios TN (2023) Explainable artificial intelligence and cybersecurity: a systematic literature review. arXiv preprint arXiv:2303.01259. https://arxiv.org/abs/2303.01259.
22. Mohale VZ, Obagbuwa IC (2025) A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhance transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence 8: 1526221. https://doi.org/10.3389/frai.2025.1526221.
23. Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 607–617. https://doi.org/10.1145/3351095.3372850.
24. Nadeem A, Vos D, Cao C, Pajola L, Dieck S, Baumgartner R, Verwer S (2023) SoK: Explainable machine learning for computer security applications. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), pp 221–240. IEEE.
25. Rjoub G, Bentahar J, Wahab OA, Mizouni R, Song A (2023) A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management 20(2): 1095–1116. https://doi.org/10.1109/TNSM.2023.3282740.
26. Sarker IH, Janicke H, Mohsin A, Gill A, Maglaras L (2024) Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: methods, taxonomy, challenges and prospects. ICT Express 10(4): 935–958. https://doi.org/10.1016/j.icte.2024.05.007.
27. Senevirathna T, Siniarski B, Liyanage M, Wang S (2024) Deceiving post-hoc explainable AI (XAI) methods in network intrusion detection. In: 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), pp 107–112. IEEE.
28. Sharafaldin I, Lashkari AH, Ghorbani AA (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp 108–116.
29. Sharma A, Kiciman E (2020) DoWhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216. https://arxiv.org/abs/2011.04216.
30. Shoukat S, Gao T, Javeed D, Saeed MS, Adil M (2025) Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. Computers & Security 149: 104191. https://doi.org/10.1016/j.cose.2024.104191.
31. Sindiramutty SR, Tan CE, Lau SP, et al. (2024) Explainable AI for cybersecurity. In: Advances in Explainable AI. IGI Global. https://www.igi-global.com/chapter/explainable-ai-for-cybersecurity/336871.
32. Wei F, Li H, Zhao Z, Hu H (2023) xNIDS: Explaining deep learning-based network intrusion detection systems for active intrusion responses. In: Proceedings of the 32nd USENIX Security Symposium.
33. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J (2019) GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems 32. https://arxiv.org/abs/1903.03894.
34. Zhang Z, Al Hamadi H, Damiani E, Yeun CY, et al. (2022) Explainable artificial intelligence applications in cyber security: state-of-the-art in research. IEEE Access. https://ieeexplore.ieee.org/abstract/document/9875264/.
35. Zhong C, Yayla A (2026) Cognitive impacts of explainable AI in cybersecurity incident response: challenges and propositions. Information Systems Frontiers 28(2): 505–527. https://doi.org/10.1007/s10796-025-10609-y.
36. Zhong M, Lin M, Zhang C, Xu Z (2024) A survey on graph neural networks for intrusion detection systems: methods, trends and challenges. Computers & Security 141: 103821. https://doi.org/10.1016/j.cose.2024.103821.

Copyright: © 2026 John Tighil. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.