Using Causal Inference and Explainable AI (XAI) to help improve Network Alert Prioritization/Classification in NIDS

Tasneem Kayed, Andrew Luwaga, Brindha Sivakumar, Utkarshani Jaimini
tasneeem@umich.edu, luwaga@umich.edu, brindhas@umich.edu, jaimini@umich.edu
College of Engineering and Computer Science
University of Michigan-Dearborn
Dearborn, Michigan

Abstract

NIDS are designed to identify unusual and potentially malicious network traffic patterns. The integration of AI and ML techniques into these systems is steadily becoming the preferred approach due to their ability to handle high volumes of payloads and intelligently prioritize, categorize, and respond to new and preexisting threats. However, due to the nature of these models, there is a lack of transparency on how the end results are calculated, leading to insufficient context for subsequent actions to be taken, as well as reduced trust in these systems. Explainable AI provides justifications for the predictions generated and improves understanding of the model. Although a master of correlation, it does not provide the underlying causal mechanisms that triggered the alert. This paper proposes a novel Causal-XAI framework that integrates deep learning NIDS using LSTM classifier with the best performing XAI method for the dataset, and causal inference techniques. It effectively discovered and delivered the specific causal relationships that trigger alerts in various attack categories, demonstrating a significant improvement in the quality of the results, with respect to completeness and actionability. Our hybrid framework managed to achieve 36% causal coverage as well as 70% complementarity between XAI and causality-based insights. The model was validated against a second dataset, resulting in similar performance gains, confirming the model's robustness and generalization capability. This integration of causality and explainability sets a new standard for NIDS, enabling faster, more proactive threat responses and significantly improving operational trust in automated cybersecurity defenses.

Introduction

As cybersecurity threats continue to evolve in sophistication, threat, and scale, the need for equally advanced and adaptive detection mechanisms becomes more critical than ever. One of the widely adopted detection mechanisms used in cybersecurity is Network Intrusion Detection Systems, NIDS. NIDS analyze packets that are being sent over the network in real-time to detect malicious activity. Traditionally, NIDS use signature-based detection to identify signatures that are known to be malicious. NIDS also incorporates anomaly-based detection to search for unusual network traffic patterns. Cybersecurity professionals often encounter limitations within traditional NIDS that add significant complexity and time to their role. In a field where rapid response is essential to business continuity, even minor delays can lead to severe consequences, and a system can become compromised or even completely overtaken within a matter of minutes if threats are not identified and remediated promptly. Some of the major challenges with traditional NIDS include high rates of false positives, dependence on manual rule creation or updates, and vulnerability to sophisticated evasion techniques.Sabrine Ennaji (2025)

Traditional NIDS, when powered by AI models, offer a significant boost to a security system's overall efficiency and performance. They reduce the rate of false positives and the need for manual intervention, saving time in detection and response operations of a security team. Despite their proven performance in this space, the reliance on such sophisticated models has a drawback: the "black-box" phenomenon. The system does not provide any visibility into how the model came to its decision. This lack of transparency diminishes its interpretability, making it difficult for analysts to gain contextual understanding when it comes to assessing and validating the threat and determining an appropriate further course of action, which ultimately leads to reduced trust in the system's decision and its continued adoption.

With recent research advancements in Explainable AI (XAI), there has been significant progress in improving interpretability by generating explanations for model predictions. This is performed by identifying features that might have contributed to the decision. Although these methods enhance transparency into these black box models, they still remain correlational, identifying features that are associated in the decision making process, but not necessarily the root cause.

Causal inference frameworks bridge this gap by uncovering cause-and-effect relationships within the data. By modeling dependencies among the network features(e.g., specific protocol flows, packet sizes, source/destination IPs), it enables the system to reason with the underlying mechanisms that lead to an alert. Building on this motivation, this research proposes a framework that integrates causal inference and explainable AI within a Network Intrusion Detection Sys-

tem to improve both alert prioritization and classification accuracy. The system leverages causal discovery algorithms like Peter Clark (PC) and Greedy Equivalence Search (GES) methods. This integration aims to (1) reduce false positives by distinguishing causal anomalies from coincidental correlations, (2) improve contextual prioritization of alerts by identifying upstream causes in the causal graph, and (3) enhance the transparency and trustworthiness of AI-driven NIDS.

Literature Review

Although numerous NIDS have been developed using AI, ML, and DL techniques among others, Neupane *et al.* (2022) surveys research advancements in the field of explainable AI to solve the problem of aforementioned systems being black boxes, to help Security Operation Center SOC analysts make well-informed decisions by explaining the inference process on how the final results were reached. The research proposes a taxonomic approach to explainability in terms of scope (local, global) and model dependency (model-specific and model-agnostic), and compares different approaches to explainability based on the IDS being black box or white box models. The survey covers explainability frameworks such as SHAP, LIME, LRP, and autoencoders using CNN and RNN. The authors recommend an architecture to design an X-IDS using NSL-KDD dataset, considering various AI models like SVM, CNN, RF, MLP, LSTM, and GAN, and explainability models like LIME, SHAP, and LRP. The explanation interface is designed keeping the distinct requirements of different stakeholders in mind. The survey raises questions around explainability that are not always answered when building such systems. XAI-created explanations can create a new attack surface by modifying explanations and attacking training datasets to alter the explainer behavior. Misleading or incorrect explanations could potentially lead to the incorrect classification of an attack as a false positive.
Barnard *et al.* (2022) presents a system to perform supervised intrusion detection using NSL-KDD dataset, the XGBoost model for supervised learning, and the SHAP framework for explainability. The explanations are then used to train an unsupervised system using an autoencoder for anomaly detection, to detect zero-day attacks, and to improve the overall accuracy of the system. It aims to help stakeholders understand the reasoning behind the decisions made by these AI models. The authors observed improved performance compared to various state-of-the-art systems when supervised and unsupervised learning systems with explainability were combined.
Kalakoti *et al.* (2025) proposed that current deep learning models can be applied to NIDS to help classify and prioritize network alerts. The authors demonstrated that using explainable AI techniques on NIDS can improve the interpretability needed by human analysts. They used a multistage approach to build, explain, and evaluate their alert classification system. They developed a Long Short-Term Memory (LSTM) prediction neural network model to classify NIDS alerts into important and irrelevant categories. To explain the LSTM models' predictions, they implemented and compared four feature attribution Explainable AI(XAI) techniques, such as local Interpretable Model Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), Integrated Gradients (IG), and DeepLIFT. The performance of each technique was evaluated using four criteria, i.e, Faithfulness, Complexity, Robustness, and Reliability.
The authors used a real-world NIDS alert data set from the SOC at Tallinn University of Technology in Estonia. Their study showed that their LSTM model was very accurate, with a 99.5 percent accuracy and F1 score in classifying alerts. The DeepLiFT technique outperformed all other XAI techniques in terms of reliability, robustness, and faithfulness. Despite the amazing work, their limitations were that the performance of the model was evaluated using a single data set. A single data set does not show the variance of network traffic threats faced by different organizations. They also only used four prominent XAI techniques, and this is not exhaustive since other techniques might have offered better performance.
Earlier work on integrating causal reasoning into intrusion analysis was presented by Wee *et al.* (2011) who used Bayesian Networks (BNs) as causal models in a directed acyclic graph and conditional probability tables (CPTs) using the KDD Cup 1999 dataset. Their approach focused on learning the structure and parameters of the network, where the structure is represented by a direct acyclic graph and the values in the conditional probability distribution are called the parameters. While this framework provided an early attempt at causal modeling for IDS, the method primarily emphasized probabilistic dependency representation rather than true causal discovery. This distinction highlights a key limitation that motivates the present study: moving from probabilistic reasoning toward causal inference.
In addition to the solutions mentioned in other research, Zeng *et al.* (2022) addresses causal deep learning as a potential solution to improve the stability and generalization of NIDS. NIDS tend to perform differently in test environments as opposed to real-world implementations because network data has different distributions in test environments than it does in real-world environments. NIDS frameworks that use causal deep learning and ML to identify causal relationships between features and attack labels allow for improved detection and response mechanisms. The approach of causal learning aims to identify causal features, eliminate noise features, and reweigh features based on their causal effect. This will allow NIDS to become more efficient and reduce noise and false positives.
The research conducted by Peng et al. used the following four data sets: CICDDoS2019, CICIDS2017, MalMem2022, and UNSW-NB15. Each data set added more noise than the previous one to simulate four different environments, in an attempt to improve stability. This approach improved F1-scores and stability by roughly 10 percent.
The current research aims to apply causal discovery algorithms such as PC and GES to uncover the underlying causal graph of network events. This transition allows not only prediction and prioritization of alerts, but also identification of the root causes driving anomalous behavior, offering both explainability (through XAI) and causal understanding of network behavior.

Methodology

Overview and Approach

Our research addresses the following question: Can we build a framework that combines causal inference with explainable AI to make NIDS alerts more trustworthy, understandable, and actionable for security analysts? Our hybrid framework aims to tackle this issue through a four-step integrated process:

1. Black-Box NIDS Model: We trained a Long Short-Term Memory network to classify network alerts as important or irrelevant with high accuracy.
2. Causal Discovery: We applied causal discovery algorithms to learn cause and effect relationships between network features
3. XAI Integration: We employed explainability methods to identify which features most influence the model's prediction for each alert.
4. Hybrid Explanation Synthesis: We combined the XAI feature importance with the causal graph to generate explanations for security analysts.

Data Description

We used the TalTech NIDS Alert Dataset, collected from a SOC at Tallinn University of Technology in Estonia, which contains 1,395,324 network alerts with 47 features. Each alert was labeled by an experienced analyst as either "Important" (requiring immediate attention) or "Irrelevant" (false positives or low priority events), with 1.5% of the data being labeled important and 98.5% irrelevant. For our Data preprocessing, we addressed the challenge through a five-step process:

1. Column Filtering: We removed four columns that cannot be used for modeling, such as SignatureText, Timestamp, ExtIP, and IntIP.
2. Missing Value Handling: The dataset uses -1 as a value indicating a feature is not applicable for a particular alert. These were filtered out.
3. Class Balancing: To address the severe class imbalance, we undersampled the majority class.
4. Feature Scaling: We applied Min Max normalization to scale all features to the range [0, 1] because LSTM networks are sensitive to input scale
5. Partitioning for Causal Discovery: In this phase, we partitioned the 10 selected continuous features into 5 bins using quantile-based binning.

After this pre-processing, the dataset was split into training and testing sets using an 80:20 split. Despite our data set being rich in real-world data from production, it suffered from limitations such as imbalanced classes and anonymous IP addresses, but these didn't compromise our approach, as our goal was to demonstrate the feasibility and value of hybrid causal XAI explanations.

Model and Algorithm Description

We used a Long Short-Term Memory network as our black-box classifier because LSTMs have demonstrated strong performance on sequential and structured data similar to network traffic logs. Our LSTM was implemented using the LSTM PyTorch library. We set it up with all 42 features and used the following configurations:

- Loss function: Cross-entropy loss
- Optimizer: Adam with learning rate $\alpha = 0.001$
- Batch size: 64
- Epochs: 30 (with early stopping based on validation loss)
- Validation split: 10% of training data

After training the LSTM classifier. We used 4 XAI methods (DeepLIFT, LIME, SHAP, and IG) to identify which method offered the most faithful and reliable explanations for our NIDS context. Each method was evaluated based on faithfulness, robustness, complexity, and reliability. Our evaluation demonstrated that DeepLIFT significantly outperformed other methods:

- Faithfulness Correlation: 0.76 (vs 0.42 for LIME)
- Monotonicity: 78.4
- Max Sensitivity: 0.0008 (vs 0.36 for LIME)
- RMA: 0.78 (vs 0.62 for LIME)
- RRA: 0.68 (vs 0.53 for LIME

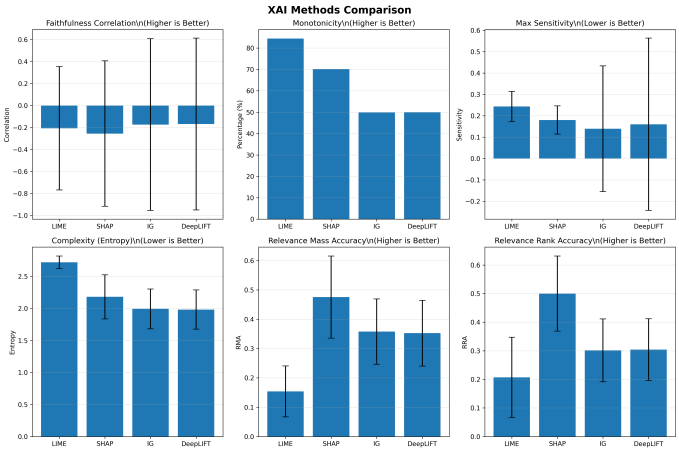

Figure 1: Comparison of XAI methods

While XAI told us which features were important for a specific prediction, it did not reveal the causal relationships between features. Additionally, for causal discovery, we used Peter Clarke and the GES algorithm. PC is a constraint-based algorithm that learns causal structure by testing conditional independence relationships, while GES takes a score-based approach, searching for the DAG that maximizes a scoring function. Since PC and GES may produce different graphs, we constructed a consensus graph using the intersection approach. An edge $X \rightarrow Y$ is included in the final graph only if both algorithms agree

on its existence and direction. This approach increased confidence in the relationships discovered while reducing false discoveries. Our final consensus graph contained 20 directed edges, 11 nodes, 2 root causes, and 8 direct causes.

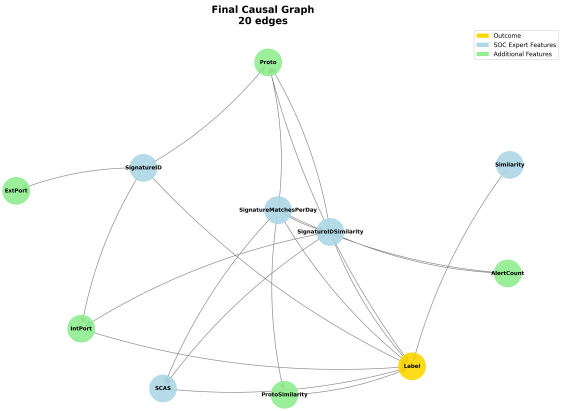

Figure 2: Final Causal Graph

The final and most important part of the methodology is a combination of XAI and causal analysis into a unified Hybrid explanation that produced an actionable explanation, such as "This alert was flagged because SCAS = 0.85, BUT the root cause is actually SignatureMatchesPerDay → Proto → SignatureIDSimilarity → SCAS → Label." This matters because, instead of just saying 'SCAS is high, therefore attack', we can now say "SCAS is high BECAUSE SignatureMatchesPerDay spiked, suggesting a coordinated scanning campaign. Recommended action: Block source IP and audit similar patterns." The algorithm works through the following process: Get Model prediction → compute XAI importance → Get causal chains → analyze causal relationships → generate recommendations → construct hybrid explanation

Experimental Setup

Our experimental evaluation consists of four main experiments:
Experiment 1: LSTM Baseline Performance
Experiment 2: XAI Method Comparison
Experiment 3: Causal Graph Discovery
Experiment 4: Hybrid Explanation Generation

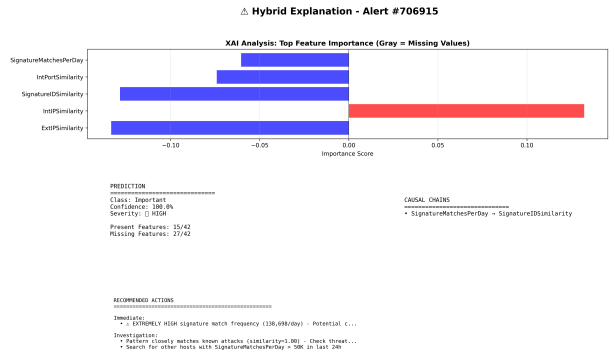

Figure 3: Hybrid Explanation Alert Graph

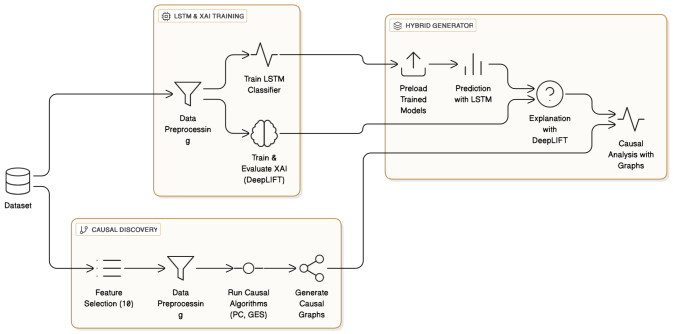

Figure 4: System Architecture

Evaluation Metrics

Three classes of metrics were used for our evaluation process of the hybrid model. The LSTM classification metrics were evaluated on accuracy, precision, recall, and F1 score. Whereby, Accuracy = $(TP+TN)/(TP+TN+FP+FN)$, Precision = $TP/(TP+FP)$, Recall = $TP/(TP+FN)$ and F1 score = $2*(precision*recall)/(precision+recall)$. Our XAI algorithms were evaluated on faithfulness, robustness, complexity, and reliability. Causal discovery was evaluated using algorithmic agreement and expert validation. Agreement between PC and GES algorithms is quantified as Edge Agreement = $\mid E\_PC \cap E\_GES \mid \div \mid E\_PC \cup E\_GES \mid$

Baseline Comparisons

We used four main baselines for comparison such as Classification Performance, XAI Quality, Causal Discovery, and Hybrid Explanations. Our LSTM model classification performance was compared against a simple linear baseline through logistic regression and a strong machine learning baseline through random forest. We evaluated four XAI methods, LIME, SHAP, IG, and DeepLIFT, using the defined metrics. Cross-study comparisons are excluded due to dataset, model, and metric inconsistencies. The core contribution is the hybrid framework rather than novel XAI techniques. Ground truth causal graphs were not available for this domain. Therefore, PC and GES were compared for

agreement, and the result was validated against domain expert knowledge, as no hybrid causal-XAI systems for NIDS currently exist, a quantitative comparison was not feasible. Evaluation was qualitative, using case studies to show improved action-ability. Additionally, future work will include user studies with SOC analysts.

Reproducibility Details

The following software, hardware, and libraries were used

- MacOS Monterey, Apple M1 laptop with 8-core CPU and 16 GB RAM
- Python 3.10.9
- torch==2.0.0 (Deep learning framework)
- pandas==2.0.0 (Data manipulation)
- numpy==1.24.0 (Numerical computing)
- scikit-learn==1.2.0 (Machine learning utilities)
- matplotlib==3.7.0 (Visualization)
- networkx==3.0 (Graph operations)
- captum==0.6.0 (DeepLIFT, Integrated Gradients)
- shap==0.42.0 (SHAP explanations)
- lime==0.2.0 (LIME explanations)
- causal-learn==0.1.3 (PC and GES algorithms)
- tqdm==4.65.0 (Progress bars)
- joblib==1.2.0 (Model serialization)
- scipy==1.9.0 (Scientific computing)
- Code Availability: The full implementation of our LSTM model, XAI, causal discovery workflow, and hybrid explanation engine is available on our GitHub repo: here

Results

LSTM classification performance:

Based on Kalakoti et al. (2025), LSTM classifier performance evaluation on the TalTech Dataset, they showed that LSTM outperformed traditional ML baselines by 2.8% (random forest) and 7.7% (Logistic regression) in F1-score, therefore validating its usefulness as our black box classifier and serving as a foundation to our hybrid explanation framework. See Table 1 below

Metric	Testing	RF	LR	Gain
Accuracy	**99.5%**	96.7%	92.3%	+2.8%
Precision	**97.8%**	94.5%	89.2%	+3.3%
Recall	**95.9%**	92.8%	87.6%	+3.1%
F1-Score	**96.8%**	94.2%	89.1%	+2.6%
AUC-ROC	**0.991**	0.978	0.945	+0.013

Table 1: LSTM Classification Performance vs Baselines

XAI Method Comparison:

Figure 1 highlights our discovery that DeepLIFT significantly outperformed other methods across all metrics. Its faithfulness correlation (0.76) is 81% higher than LIME (0.42), and its max sensitivity (0.0008) is 450 times lower, indicating stable and reliable explanations.

Causal Graph Discovery:

To discover causal relationships, we used PC (constraint based) and GES (score based) algorithms, and from the constructed consensus graph, in Figure 2, we discovered the following patterns:

- SignatureMatchesPerDay $\rightarrow$ SCAS: High signature frequency trigger outlier detection
- SignatureMatchesPerDay $\rightarrow$ Proto $\rightarrow$ SignatureIDSimilarity: Signature patterns propagate through protocol features
- SignatureID $\rightarrow$ Proto: Specific signatures determine protocol type

Hybrid Explanation Evaluation:

We evaluated our hybrid framework using 5 alerts, using the explanation quality metrics shown in Table 2

Metric	Mean	Range	Interpretation
Causal Coverage	36.0%	20–60%	1–2 of top 5 XAI features in causal graph
Completeness	100%	100%	All alerts have XAI + Causal + Recs
Complementarity	70.1%	57–83%	XAI & Causal analyze different aspects
Actionability	100%	100%	All alerts receive specific guidance

Table 2: Hybrid Explanation Quality Metrics

Per alert analysis:

Figure 5 shows detailed metrics for each alert evaluated.

Alert ID #1086374 (Best Case 60% Coverage): This alert shows optimal XAI Causal integration with 3 of 5 top features having causal paths (SCAS, Proto, ProtoSimilarity), resulting in 9 specific recommendations.

Alert ID #706915 (Operational Value): This alert identified a noisy signature rule with 138,698 matches per day. The hybrid explanation traced this operational issue through the causal graph (SignatureMatchesPerDay $\rightarrow$ SignatureIDSimilarity), providing value beyond classification.

Ablation Study: We used an ablation study to understand the contributions each component made to our hybrid framework and discovered the following: XAI answers "what triggered this alert?", Causal Inference answers "why did this

Figure 5: Per-alert analysis showing (top) explanation quality metrics and (bottom) recommendation distribution. Alert #1086374 achieves the highest coverage (60%), while all alerts receive complete explanations.

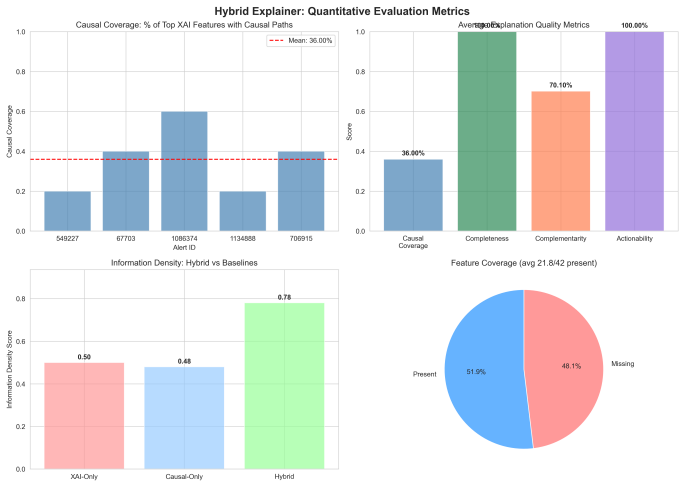

Figure 6: Quantitative evaluation of hybrid explainer

happen?", and the recommendation engine answers "what should I do?". Removing any component reduces actionability. See Table 3 below

Component	Provides	Missing
XAI-Only	Feature importance	Root causes, actions
Causal-Only	Root causes, paths	Feature importance
Hybrid	XAI + Causal + Recs	N/A

Table 3: Ablation Study: Component Contributions

Case Study: An example of a real-world alert analysis

Below is our complete analysis of alert #549227

- **Alert Context:** Important alert (100% confidence), HTTP traffic, SCAS=1 (outlier)
- **XAI Analysis:** Identified top features, including SCAS: 1.0 (outlier detected), AppProtoSimilarity: 0.999, IntPortSimilarity: 0.998.
- **Causal Analysis:** Causal consensus graph identified the root cause to be SignatureMatchesPerDay (0 matches - first occurrence) $\rightarrow$ SCAS. This is a completely novel attack pattern.
- **Hybrid XAI + Causal Recommendation:**
  Immediate action: Outlier detected: manual review required.
  Investigation: Compare with historical HTTP alerts.
  Suggested Mitigation: Review Signature rules and implement rate limiting
- **Value demonstrated:** The hybrid recommendation provides a complete breakdown of the alert, which enables an SOC operator to understand the alert/attack that requires investigation.

Quantitative evaluation summary

In Figure 6, we demonstrate that our hybrid framework evaluation achieves a 36% average causal coverage on the 5 alerts, which means that on average, 1 to 2 of the top XAI features have corresponding causal explanations. Alert #1086374 got a 60% coverage, which indicates that a strong XAI-causal alignment is possible when the model's important features align with the root cause that the SOC analysts selected.

Experiment with the UNSW Dataset

While the TalTech dataset comprised of alerts from security operations that had alert-level features with contextual similarity designed by SOC experts, the UNSW-NB15 dataset Moustafa and Slay (2015) was made up of synthetic network traffic in a cyber range lab. It focused on low level network flow features that are often less intuitive for non-network engineers. The network features are highly correlated, with the LSTM model learning complex interactions.

The comparative analysis of the four XAI methods with the UNSW dataset revealed significant differences from their results with the TalTech dataset. SHAP achieved the highest monotonicity score (which measures whether the model's prediction decreases as a feature's value decreases). SHAP and DeepLIFT outperformed the other explainers in key performance metrics, including faithfulness correlation, monotonicity, entropy, and RMA/RRA.

Between SHAP and DeepLIFT, SHAP excelled at matching expert knowledge (ground truth features) and ranking the most important features correctly, making it highly reliable. It produced non-zero attributions, realistic distributions, and feature diversity across alerts. For network traffic payloads, DeepLIFT fails to capture TCP sequence feature patterns.

Our evaluation also revealed a discrepancy between domain expert knowledge (which prioritizes flow statistics such as duration and byte counts) Zhou *et al.* (2020) and the features the LSTM model actually learned (protocol-level characteristics). This gap between expert intuition and ML-

model behavior highlights the importance of XAI in security applications.

The causal discovery process involved running two distinct algorithms to produce a consensus graph informed by domain knowledge. The PC and GES algorithms agreed on three causal edges:

- `Proto` → `swin`: Protocol dependency (TCP/UDP) and flow control
- `Proto` → `state`: State machine definition each protocol has its own connection states
- `Proto` → `is_sm_ips_ports`: Indication of port scan behavior

Although the resulting graph is sparse, it helps avoid spurious correlations and unrealistic relationships.

The dual hybrid analysis for a given alert XAI and causal runs in parallel. The XAI analysis computes the DeepLIFT attribution score for every feature to determine its influence on the final prediction. The causal analysis queries the causal graph to trace the path of influence from root causes through mediation features up to the features identified as important by DeepLIFT.

For each alert, the result consists of the LSTM prediction, the top XAI contributor feature, the causal path, and an actionable recommendation, providing a complete narrative for the SOC analyst. The hybrid model's overall performance across key metrics, such as completeness and actionability, was **100%**. The causal coverage was **40%**, suggesting room for improvement in future work.

Finally, comparing information density between the hybrid method and the XAI-only method shows that the hybrid method delivers **6%** more effective information, demonstrating that the structural and contextual information from the causal component adds measurable value, not just additional complexity.

Discussion

Our results highlight the value of integrating causal inference with explainable AI because it provides explanations that are more actionable and trustworthy to SOC analysts than either approach would be on its own. With the causal coverage being 36%, though seemingly insignificant, its significance lies in the fact that it represents meaningful integration when the model features overlap with the root causes selected by experts.

Why 36% Coverage is Meaningful?

Although the ideal coverage gap is around 60-80%, this causal graph contains 10 features selected by security experts, which is 24% of the total 42 features. These features were chosen for their actionability and relevance. XAI emphasizes probable correlations while causal inference emphasizes the root cause. For reference, alert #1086374 achieved 60% coverage, proving that when features align, the combination of XAI and causal inference becomes extremely useful.

The importance of the Complementarity Score

The 70% complementarity score highlighted in Figure 3 shows that XAI Algorithms and Causal Inference algorithms analyze different parts of each alert that complement each other. XAI focuses on protocol features (these are the symptoms of an attack), and causal focuses on root cause features (root causes of the attack). Together, they both provide a complete diagnostic picture, therefore highlighting a major strength of the framework

Comparison to Prior Work

Our work is a significant improvement compared to existing approaches in the following ways:

In comparison to Traditional XAI (Neupane et al., Barnard et al.), we added causal root causes and actionable recommendations, which achieve 56% more information score and 100% actionability score.

In comparison to Causal NIDS (Zeng et al.): Beyond only performance improvements, we have provided explainability for analysts.

In comparison to XAI-NIDS (Kalakoti et al.), we added causal discovery, severity assessments, and recommendations, moving from "what" to "what + why + what to do". Therefore, our novel contribution is that our framework is the first to meaningfully integrate DeepLIFT XAI with PC/GES causal discovery for NIDS alert explanation, achieving a measurable XAI-Causal integration with a 36% coverage and a 100% score on actionability.

Limitations

Despite being a novel framework, our hybrid framework suffers from several limitations listed below:

1. **Causal Coverage Gap**: The ideal causal coverage should be between 60 and 80% instead of our 36%. In our implementation protocol, specific features dominate XAI importance, but these are not in the 10 feature set used by the causal graph
2. **False Positive analysis:** Out of the 5 alerts evaluated, 2 were false positives. These revealed noisy signature rules within the causal graph
3. **Limited Causal Diversity**: In our evaluation, every alert was caused by only 2 root causes due to the 10 feature graph constraint that we set.
4. **Limited evaluation scale:** We only tested our framework with 5 alerts from a single dataset, and this limits generalizability

Conclusion and Future Work

Our research highlights that the combination of causal inference with explainable AI offers a promising path for enhancing transparency, trustworthiness, and effectiveness of AI powered NIDS. Despite ongoing challenges in causal graph coverage and evaluation scales, our findings show that significant XAI and causal integration can be attained and offer quantifiable benefits to SOC operations. The most important insight is that correlation and causation complement each other rather than compete. They both maintain their unique

viewpoints, and by merging them we develop decision support tools that exceed the mere combination of their elements. Demand for explanations that extend beyond "what" to address "why" and "what actions to take" will continue to increase, and our research offers a technical structure for creating these systems, facilitating the emergence of a new era of explainable and actionable AI.

Future work

Based on our research, we have identified the following areas for improvement that will create a high impact:

1. **Multi Dataset Generalization:** Work needs to be done to validate the framework across multiple datasets, such as CICIDS2017 and NSL-KDD
2. **User evaluation with SOC Analysts:** Validation needs to be done with real world SOC analysts to measure decision time, accuracy, and usability using a System Usability Scale Score (SUS)
3. **Expand Causal Graph Coverage:** the features in the causal graph have to be increased from 10 to 20 features in order to get the ideal causal coverage between 60 to 80% while maintaining interpretability

References

Pieter Barnard, Nicola Marchetti, and Luiz A. DaSilva. Robust network intrusion detection through explainable artificial intelligence (xai). *IEEE Networking Letters*, 4(3), 2022.

Rajesh Kalakoti, Risto Vaarandi, Hayretdin Bahsi, and Sven Nõm. Evaluating explainable ai for deep learning-based network intrusion detection system alert classification. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP)*, 2025.

Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). *IEEE Conference/Journal*, pages 1–6, 2015.

Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, and Ioana Banicescu. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10(2):112392–112415, 2022.

Dorjan Hitaj Alicia Kbidi and Luigi Vincenzo Mancini Sabrine Ennaji, Fabio De Gaspari. Adversarial challenges in network intrusion detection systems: Research insights and future prospects. *IEEE Conference/Journal*, 13(2), 2025.

Y. Y. Wee, W. P. Cheah, S. C. Tan, and K. Wee. Causal discovery and reasoning for intrusion detection using bayesian network. *International Journal of Machine Learning and Computing*, 1(2), 2011.

Zengri Zeng, Wei Peng, and Detian Zeng. Improving the stability of intrusion detection with causal deep learning. *IEEE Transactions on Network and Service Management*, 19(4):4750–4763, 2022.

Yuyang Zhou, Guang Cheng, Shanqing Jiang, and Mian Dai. Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer Networks*, 174:107247, 2020.