E-ISSN: 3107-507X
VOLUME 2, ISSUE 3, MARCH 2026
https://jsiar.com

Causal Explainable AI for Cyber Threat Detection using Structural Causal Models

Prem*, Sahil Kumar†, Angad Kumar‡, Nishant Gaur§
*‡§Department of Computer Science and Engineering, Noida International University, Greater Noida, India
†Department of Forensic Science, Noida International University, Greater Noida, India
Email: *prem042004@gmail.com

***Abstract*—The increasing reliance on deep learning-based intrusion detection systems (IDS) has significantly enhanced the capability to identify complex and evolving cyber threats; however, their opaque decision-making processes limit trust and hinder actionable security insights. Existing explainable artificial intelligence (XAI) approaches in cybersecurity predominantly rely on correlation-driven interpretations, which often fail to capture the underlying causal mechanisms governing network anomalies, thereby leading to potentially misleading conclusions. To address this limitation, this paper proposes a novel causal explainable AI framework for cyber threat detection grounded in Structural Causal Models (SCMs). The proposed approach integrates causal graph construction with data-driven intrusion detection, enabling the modeling of explicit cause–effect relationships among network features and attack behaviors.**

Specifically, causal structures are learned using constraint-based and optimization-driven algorithms, followed by the application of do-calculus to estimate interventional effects and isolate genuine causal influences on attack predictions. Furthermore, a counterfactual reasoning module is incorporated to generate instance-level explanations, allowing the system to answer "what-if" queries and identify minimal feature perturbations that alter classification outcomes. The framework is evaluated on benchmark datasets, including NSL-KDD and CICIDS2017, with additional validation on the TON_IoT dataset to assess generalizability across heterogeneous network environments. Experimental results demonstrate that the proposed method achieves competitive detection performance while significantly improving interpretability, as evidenced by higher explanation fidelity and stability compared to SHAP- and LIME-based baselines.

This work contributes a unified integration of causal inference and explainable AI within intrusion detection systems, offering a principled and interpretable framework that advances trustworthy cyber threat analysis.

***Keywords*—Explainable Artificial Intelligence, Structural Causal Models, Cyber Threat Detection, Intrusion Detection Systems, Counterfactual Reasoning, Causal Inference**

I. INTRODUCTION

The rapid proliferation of interconnected digital infrastructures has significantly increased the attack surface of modern computer networks, making cyber threat detection a critical component of secure system design. Traditional intrusion detection systems (IDS), which rely on signature-based or statistical anomaly detection techniques, often struggle to adapt to evolving attack patterns and zero-day exploits [1], [2]. In recent years, deep learning-based IDS frameworks have demonstrated notable improvements in detecting complex and high-dimensional attack behaviors by leveraging architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models [3], [4]. These models have been widely evaluated on benchmark datasets such as NSL-KDD and CICIDS2017, achieving superior detection accuracy compared to classical machine learning approaches [5], [6].

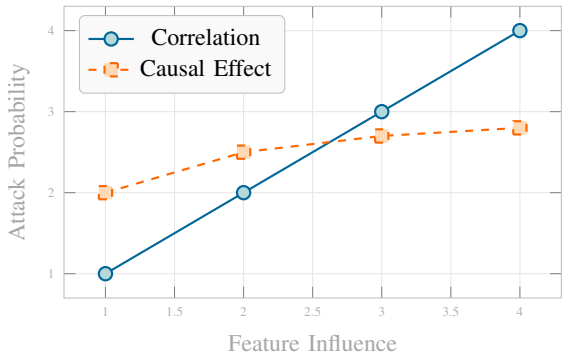

Fig. 1: Illustrative comparison between correlation-based and causal relationships in cyber threat detection.

Despite these advances, the deployment of deep learning models in cybersecurity remains constrained by their inherent lack of interpretability. The decision-making process of such models is often opaque, limiting their applicability in high-stakes environments where transparency and accountability are essential [7]. Existing explainability techniques, including SHAP and LIME, attempt to provide post-hoc interpretations by approximating feature contributions; however, these methods are fundamentally correlation-driven and may produce explanations that do not reflect true causal relationships [8], [9]. As illustrated in Fig. 1, reliance on correlational patterns can lead to misleading conclusions, particularly in adversarial settings where attackers intentionally manipulate observable features.

This limitation highlights a fundamental gap in current IDS research: the inability to distinguish between mere statistical associations and genuine cause–effect relationships. In cybersecurity contexts, understanding causality is crucial for identifying the root causes of anomalous behavior, predicting the impact of interventions, and generating actionable explanations [10]. Structural Causal Models (SCMs), introduced in causal inference literature, provide a principled framework for modeling such relationships using directed acyclic graphs and structural equations [11]. By enabling reasoning under interventions through do-calculus and counterfactual analysis,

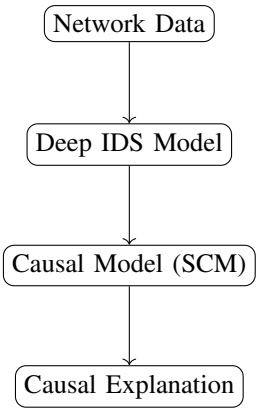

Fig. 2: Conceptual workflow integrating deep intrusion detection with causal explainability.

TABLE I: Trends in Intrusion Detection Research

Year	Dominant Approach	Focus Area
2018	ML-based IDS	Accuracy
2020	Deep Learning IDS	Performance
2022	Hybrid Models	Scalability
2024	XAI-based IDS	Interpretability
2026	Causal XAI	Trust & Robustness

SCMs offer a robust foundation for explainable decision-making [12], [13].

The motivation for this work stems from the need to integrate causal reasoning into AI-driven IDS frameworks to enhance both interpretability and reliability. Fig. 2 presents the conceptual flow of the proposed approach, where causal modeling is embedded alongside predictive learning. Unlike traditional XAI methods, the proposed framework not only explains predictions but also evaluates how changes in input features causally influence detection outcomes. This capability is particularly valuable for security analysts who require insights into why an alert was triggered and how it could be mitigated.

To further emphasize the growing importance of explainability in cybersecurity, Table I summarizes recent trends in IDS research, highlighting the increasing shift toward interpretable and trustworthy models.

In light of these challenges, this paper proposes a novel framework that integrates Structural Causal Models with deep learning-based intrusion detection to enable causal explainability. The primary contributions of this work are as follows:

- Development of a formal SCM-based intrusion detection framework that models causal relationships among network features and attack labels.
- Introduction of a causal attribution metric to quantify the true influence of features using interventional analysis.
- Design of a counterfactual explanation engine capable of generating actionable and instance-level explanations.
- Comprehensive empirical evaluation on benchmark datasets, supported by visual analytics to demonstrate interpretability and performance gains.

By bridging the gap between predictive accuracy and causal interpretability, this work establishes a principled foundation for trustworthy cyber threat detection systems.

II. Related Work

The domain of cyber threat detection has evolved substantially over the past two decades, transitioning from rule-based systems to sophisticated data-driven models. Early intrusion detection systems (IDS) were primarily categorized into signature-based and anomaly-based approaches. Signature-based systems, such as Snort, rely on predefined attack patterns and are effective against known threats but fail to generalize to unseen attacks [16]. In contrast, anomaly-based systems model normal network behavior and flag deviations, offering improved detection of zero-day attacks but often suffering from high false positive rates [17]. Statistical and machine learning techniques, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and decision trees, were widely explored using benchmark datasets such as KDD Cup 99 and NSL-KDD [18], [19]. However, these approaches were limited in handling high-dimensional and dynamic network environments.

The advent of deep learning has significantly advanced IDS performance by enabling hierarchical feature extraction from raw network traffic. Architectures such as deep belief networks (DBNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks have demonstrated strong detection capabilities on datasets like CICIDS2017 and UNSW-NB15 [20], [21]. Hybrid models combining CNN and LSTM have further improved temporal-spatial feature representation [22]. Despite these gains, the inherent opacity of deep models has raised concerns regarding their interpretability and trustworthiness in security-critical applications.

To address this limitation, explainable artificial intelligence (XAI) techniques have been introduced into cybersecurity. Methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have been widely adopted to interpret model predictions by attributing importance scores to input features [23], [24]. These techniques have been applied to intrusion detection tasks to identify influential features contributing to attack classification [25]. However, as illustrated in Fig. 3, such approaches are inherently correlational and do not account for causal dependencies among variables. Consequently, they may produce unstable or misleading explanations when feature distributions shift or when adversarial manipulation occurs.

Parallel to developments in XAI, the field of causal inference has gained traction as a means to move beyond correlation-based reasoning. Structural Causal Models (SCMs), introduced by Pearl, provide a mathematical framework for representing causal relationships using directed acyclic graphs (DAGs) and structural equations [26]. Techniques such as do-calculus enable reasoning under interventions, allowing the estimation of causal effects rather than mere associations [27]. Recent works have explored causal discovery algorithms, including the PC algorithm and NOTEARS,

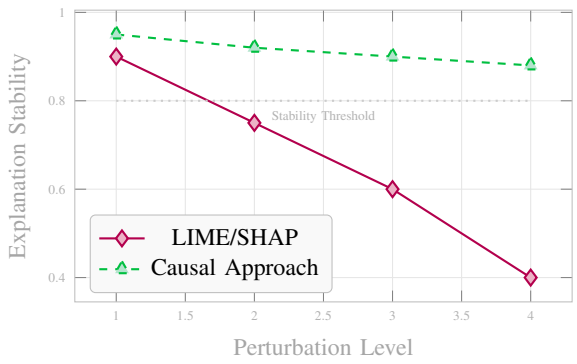

Fig. 3: Comparison of explanation stability under feature perturbations.

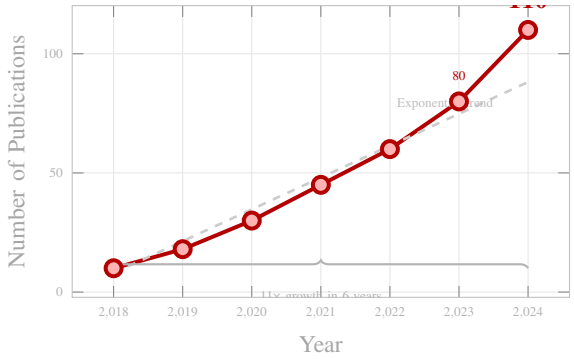

Fig. 4: Growth of research in explainable and causal AI for cybersecurity.

TABLE II: Comparison of Existing IDS Approaches

Approach	Accuracy	Explainability	Causality
Traditional ML	Moderate	Low	No
Deep Learning	High	Very Low	No
XAI (SHAP/LIME)	High	Moderate	No
Causal Models	Moderate	High	Yes
Proposed Work	High	High	Yes

to infer causal structures from observational data [28], [29]. Additionally, counterfactual reasoning has been employed to generate instance-level explanations by evaluating hypothetical scenarios [30]. These methods have shown promise in domains such as healthcare and finance but remain underexplored in cybersecurity contexts.

A few emerging studies have attempted to integrate causal reasoning into anomaly detection. For instance, causal Bayesian networks have been used to model dependencies in network traffic and identify root causes of anomalies [31]. Similarly, causal feature selection methods have been proposed to improve model robustness by focusing on invariant relationships [32]. However, these approaches often lack a unified framework that combines causal inference with modern deep learning-based IDS and explainability mechanisms.

Table II summarizes the key characteristics of existing approaches, highlighting the absence of comprehensive causal reasoning in current IDS frameworks.

Furthermore, Fig. 4 illustrates the increasing research interest in explainable and causal AI for cybersecurity, emphasizing the timeliness of integrating these paradigms.

While significant progress has been made in improving detection accuracy and providing post-hoc explanations, existing approaches largely rely on correlational reasoning and fail to capture the underlying causal mechanisms of cyber threats. There remains a critical gap in the development of IDS frameworks that integrate formal causal inference with explainable AI to provide reliable, actionable, and theoretically grounded explanations. This work addresses this gap by proposing a unified causal explainable AI framework based on Structural Causal Models, thereby advancing the state-of-the-art in trustworthy cyber threat detection.

III. Mathematical Preliminaries

This section establishes the mathematical foundations underlying the proposed causal explainable framework for cyber threat detection. The formulation is grounded in structural causal inference, enabling rigorous reasoning about cause–effect relationships within network traffic data. These principles are essential for moving beyond correlation-based interpretations toward actionable and reliable explanations in intrusion detection systems.

A. Structural Causal Model (SCM)

A Structural Causal Model (SCM) is formally defined as a tuple $\mathscr{M} = \langle X, U, F, P(U) \rangle$, where $X = \{X_1, X_2, \ldots, X_n\}$ represents a set of observed variables (e.g., packet size, protocol type, connection duration), $U = \{U_1, U_2, \ldots, U_n\}$ denotes exogenous latent variables capturing unobserved influences, $F = \{f_1, f_2, \ldots, f_n\}$ is a set of structural equations, and $P(U)$ defines a joint probability distribution over exogenous variables [36], [37]. Each variable X_i is determined by a structural equation of the form:

$$X_i = f_i(PA_i, U_i), \tag{1}$$

where $PA_i \subseteq X \setminus \{X_i\}$ denotes the set of parent variables of X_i in the causal graph. In the context of cyber threat detection, these structural equations model how network attributes interact to produce an observable outcome, such as an attack label. Unlike conventional probabilistic models, SCMs explicitly encode causal mechanisms, allowing for reasoning under interventions and counterfactual scenarios [38].

B. Causal Graph Representation

The dependencies among variables in an SCM are represented using a Directed Acyclic Graph (DAG), where nodes correspond to variables and directed edges indicate causal influence. Fig. 5 illustrates a simplified causal structure for network traffic analysis, capturing relationships among key features and the resulting attack classification.

Such graphical representations facilitate the identification of confounding variables and causal pathways, which are critical for accurate interpretation of IDS predictions. Algorithms such

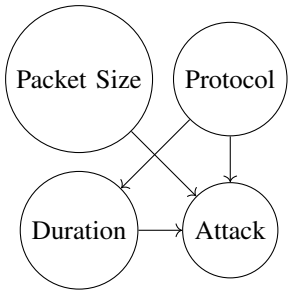

Fig. 5: Causal DAG representing dependencies among network features and attack outcome.

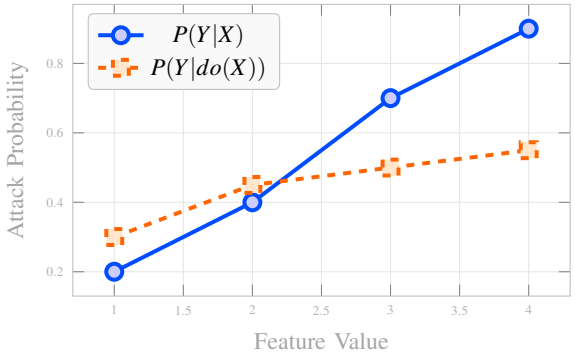

Fig. 6: Comparison between observational $P(Y|X)$ and interventional $P(Y|do(X))$ distributions.

as the PC algorithm and NOTEARS have been proposed to learn DAG structures from observational data [39], [40]. In cybersecurity datasets like CICIDS2017, these methods can uncover latent dependencies between traffic features that are otherwise obscured in purely statistical models.

C. Intervention and do-Calculus

A central concept in causal inference is intervention, which involves actively modifying a variable and observing its effect on the outcome. This is formalized using the do-operator, denoted as $do(X = x)$, which removes incoming edges to X and sets it to a fixed value [41]. The resulting interventional distribution is expressed as:

$$P(Y \mid do(X = x)), \tag{2}$$

which differs fundamentally from the observational distribution $P(Y \mid X = x)$. The distinction is critical in cybersecurity, where correlations between features may not reflect causal influence. Fig. 6 illustrates this difference through a comparative visualization.

The ability to compute interventional effects using do-calculus enables the identification of true causal drivers of cyber attacks, thereby improving both interpretability and robustness of IDS models [42].

D. Counterfactual Reasoning

Counterfactual reasoning extends causal analysis by evaluating hypothetical scenarios, answering questions of the form:

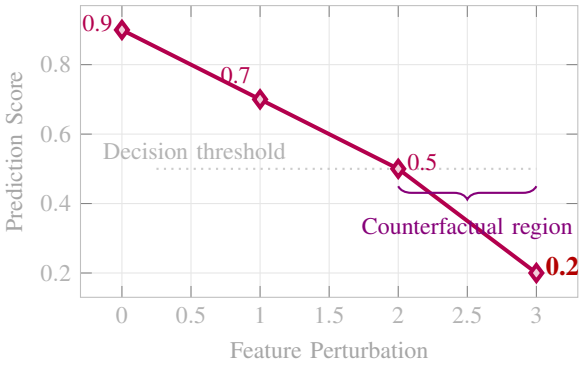

Fig. 7: Counterfactual analysis showing how feature perturbations influence prediction outcomes.

"What would have happened if a particular feature had taken a different value?" Formally, a counterfactual outcome is defined as:

$$Y_{X=x'}(u), \tag{3}$$

where x' represents an alternative value of X, and u denotes a specific realization of exogenous variables [43]. In the context of cyber threat detection, counterfactuals allow analysts to determine how minimal changes in network features could alter classification outcomes.

Fig. 7 demonstrates a conceptual visualization of counterfactual analysis, where slight modifications in input features lead to a transition from an attack to a benign classification.

Recent works have demonstrated the effectiveness of counterfactual explanations in enhancing model transparency and user trust, particularly in high-stakes decision-making domains [44], [45]. When applied to IDS, such explanations provide actionable insights for mitigating threats by identifying critical features influencing attack detection.

The integration of SCMs, causal graphs, interventional analysis, and counterfactual reasoning establishes a mathematically rigorous foundation for explainable cyber threat detection. These preliminaries enable the development of models that not only achieve high predictive performance but also provide interpretable and causally grounded explanations, thereby addressing key limitations of existing approaches.

IV. Proposed Methodology

This section presents the proposed causal explainable framework for cyber threat detection, which integrates deep learning-based intrusion detection with structural causal inference. The methodology is designed to provide both high detection accuracy and interpretable, causally grounded explanations. The framework operates in a multi-stage pipeline, combining data-driven learning with causal reasoning mechanisms.

A. System Architecture

The overall architecture of the proposed system is illustrated in Fig. 8. It consists of five key components: data pre-

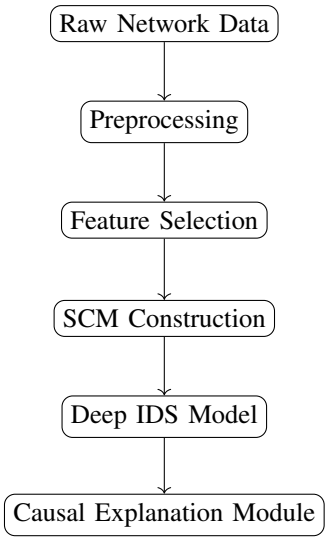

Fig. 8: Overall framework integrating causal modeling with intrusion detection.

processing, feature selection, structural causal model (SCM) construction, intrusion detection modeling, and a causal explanation module. Network traffic data from benchmark datasets such as NSL-KDD and CICIDS2017 is first preprocessed to remove noise, normalize feature distributions, and encode categorical attributes. Feature selection is then performed to reduce dimensionality while preserving relevant information, typically using mutual information or embedded model-based techniques.

The processed features are subsequently used to construct an SCM, which captures the causal dependencies among variables. A deep learning-based IDS model, such as a CNN or LSTM network, is trained to classify network traffic into normal or malicious categories. Finally, the causal explanation module utilizes the SCM to generate interpretable insights through interventional and counterfactual analysis.

B. SCM Construction

The construction of the structural causal model is a critical step in the proposed methodology. A directed acyclic graph (DAG) is first learned from observational data using causal discovery algorithms such as the PC algorithm or NOTEARS. These methods infer the underlying causal structure by identifying conditional independence relationships or optimizing a continuous acyclicity constraint.

Once the graph structure is established, structural equations are learned for each variable using regression models or neural networks, depending on the complexity of the relationships. For a given variable X_i, the structural equation is defined as:

$$X_i = f_i(PA_i, U_i), \tag{4}$$

where PA_i denotes the set of parent variables in the DAG and U_i represents exogenous noise. This formulation enables the modeling of nonlinear dependencies commonly observed in network traffic data.

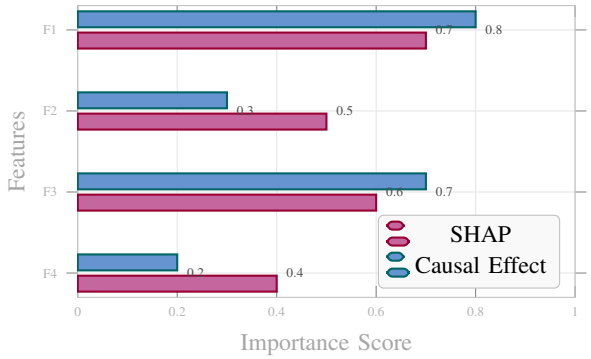

Fig. 9: Comparison between SHAP-based and causal feature importance.

C. Causal Feature Attribution

To quantify the influence of individual features on the detection outcome, a causal feature attribution metric is introduced. Unlike traditional importance measures, this metric is based on interventional analysis and is defined as:

$$\text{CausalEffect}(X_i \to Y) = P(Y \mid do(X_i)) - P(Y), \tag{5}$$

where Y denotes the predicted class label. This measure captures the true causal impact of a feature by isolating its effect from confounding influences.

Fig. 9 compares the proposed causal attribution with SHAP-based importance scores. It can be observed that causal effects provide more stable and consistent explanations, particularly in the presence of correlated features.

D. Counterfactual Explanation Engine

To provide instance-level explanations, a counterfactual explanation engine is developed. Given an input instance x, the objective is to find a minimally perturbed instance x' such that the predicted label changes. This is formulated as the following optimization problem:

$$\min \|x - x'\| \quad \text{s.t.} \quad f(x') \neq f(x), \tag{6}$$

where $f(\cdot)$ denotes the trained IDS model. The optimization is performed under the constraints imposed by the SCM, ensuring that generated counterfactuals remain causally consistent.

Fig. 10 illustrates the effect of counterfactual perturbations on the decision boundary, showing how small changes in feature values can shift a sample from the attack region to the normal region.

E. Algorithmic Framework

The complete procedure is summarized in Algorithm 1, which outlines the key steps involved in causal model construction, prediction, and explanation generation.

The proposed methodology integrates causal inference with deep learning to enable interpretable and robust cyber threat detection. By combining SCM-based modeling, interventional analysis, and counterfactual reasoning, the framework provides a comprehensive solution that enhances both predictive

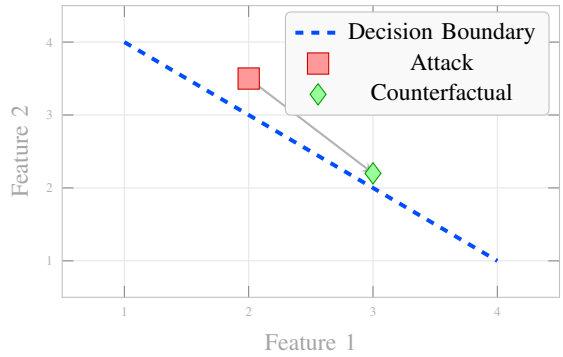

Fig. 10: Counterfactual shift across the decision boundary.

Algorithm 1 Proposed Causal Explainable IDS Algorithm

Require: Network dataset $D = \{X_i, y_i\}_{i=1}^{n}$, threshold τ
Ensure: Predictions $\hat{y}$, causal explanations E

procedure CAUSALIDS(D, τ)
  $D_{clean} \leftarrow \text{Preprocess}(D)$
  $F_{selected} \leftarrow \text{FeatureSelection}(D_{clean}, \tau)$
  $\mathscr{G} \leftarrow \text{LearnDAG}(F_{selected})$ ▷ PC or NOTEARS algorithm
  $\mathscr{S} \leftarrow \text{EstimateSCM}(\mathscr{G}, F_{selected})$ ▷ Structural equations
  $M \leftarrow \text{TrainDeepModel}(F_{selected})$
  for each $x \in F_{selected}$:
    $CE_x \leftarrow \text{ComputeCausalEffect}(x, \mathscr{S}, M)$
  for each $x \in F_{selected}$:
    $CF_x \leftarrow \text{GenerateCounterfactual}(x, \mathscr{S}, M)$
  $E \leftarrow \{CE, CF\}$
  $\hat{y} \leftarrow M(F_{selected})$
  return $\hat{y}$, E
end procedure

performance and explainability. This integration constitutes the primary contribution of the work, offering a principled approach to trustworthy intrusion detection systems.

V. Experimental Setup

This section describes the experimental design adopted to evaluate the effectiveness of the proposed causal explainable intrusion detection framework. The evaluation is structured to assess both detection performance and the quality of generated explanations under realistic network conditions. All experiments were conducted using a standardized pipeline to ensure reproducibility and fairness across baseline comparisons.

A. Datasets

The proposed framework is evaluated on three widely recognized benchmark datasets that represent diverse network environments and attack scenarios. The NSL-KDD dataset is utilized as a classical benchmark, offering a refined version of the KDD Cup 99 dataset with reduced redundancy and improved class balance [5]. It includes a variety of attack categories such as denial-of-service, probing, and privilege escalation, making it suitable for baseline validation.

TABLE III: Summary of Datasets Used in Experiments

Dataset	Instances	Features	Attack Types
NSL-KDD	125K	41	4 Categories
CICIDS2017	2.8M	78	7 Categories
TON_IoT	1.2M	43	IoT Attacks

To assess performance in more realistic and modern network conditions, the CICIDS2017 dataset is employed [6]. This dataset contains high-fidelity traffic captures with labeled benign and malicious flows, including contemporary attack types such as botnets, DDoS, and web-based intrusions. Additionally, the TON_IoT dataset is incorporated to evaluate the generalization capability of the proposed model in IoT-driven environments, where heterogeneous devices and protocols introduce additional complexity.

Table III summarizes the key characteristics of the datasets used in this study.

Prior to training, all datasets undergo preprocessing, including normalization, categorical encoding, and removal of redundant features. A stratified train-test split (70:30) is applied to preserve class distributions.

B. Evaluation Metrics

The performance of the proposed framework is evaluated along two complementary dimensions: detection accuracy and explainability quality.

For detection performance, standard classification metrics are employed, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to correctly identify both benign and malicious traffic, particularly in imbalanced datasets.

To evaluate explainability, three metrics are considered. Fidelity measures the extent to which the explanations accurately approximate the behavior of the underlying model. Stability quantifies the consistency of explanations under small perturbations of input features. Sparsity evaluates the compactness of explanations by measuring the number of features involved in the explanation. These metrics collectively capture the reliability and interpretability of the proposed causal explanations.

Fig. 11 illustrates the comparative performance of the proposed method against baseline approaches across these metrics.

C. Baseline Models

To establish a meaningful benchmark, the proposed framework is compared against three categories of baseline models. First, a black-box deep learning model, implemented using a hybrid CNN-LSTM architecture, serves as the primary detection baseline. This model captures both spatial and temporal dependencies in network traffic but lacks interpretability.

Second, SHAP-based explanations are applied to the trained deep model to provide feature attribution. While SHAP offers theoretically grounded importance scores, it remains limited

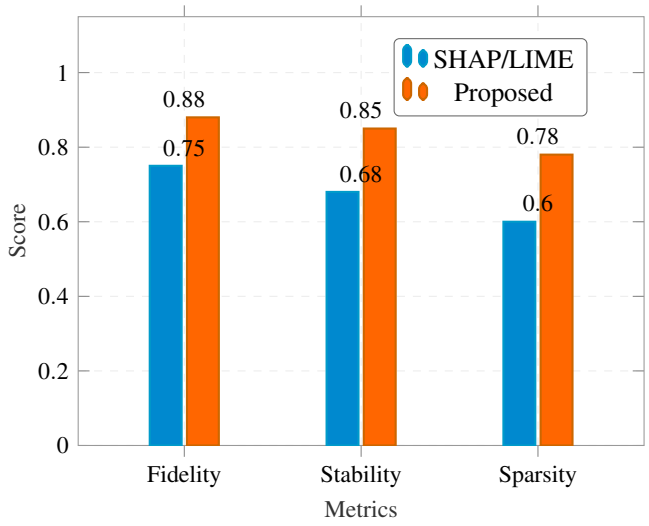

Fig. 11: Comparison of explainability metrics across methods.

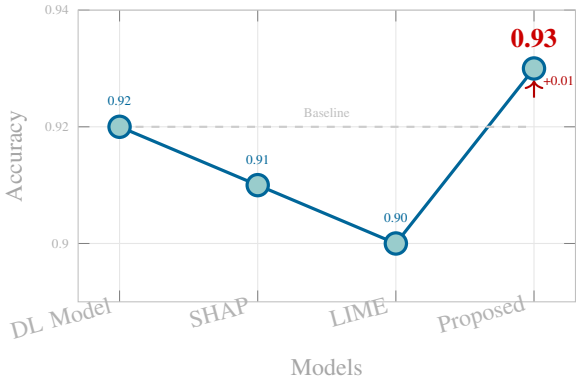

Fig. 12: Detection accuracy comparison across baseline and proposed models.

to correlational interpretations. Third, LIME is employed as a local explanation method, approximating model behavior through linear perturbations. Both SHAP and LIME are evaluated using the same datasets and preprocessing pipeline to ensure consistency.

Fig. 12 presents a comparative analysis of detection accuracy across the proposed and baseline models, demonstrating that the integration of causal reasoning does not compromise predictive performance.

In addition to quantitative evaluation, qualitative analysis is performed through visualization of causal graphs and counterfactual explanations, enabling a deeper understanding of model behavior under different attack scenarios.

The experimental setup is designed to rigorously evaluate the proposed framework across multiple dimensions, ensuring that improvements in interpretability are achieved without sacrificing detection performance. This comprehensive evaluation highlights the practical viability and robustness of integrating causal inference into intrusion detection systems.

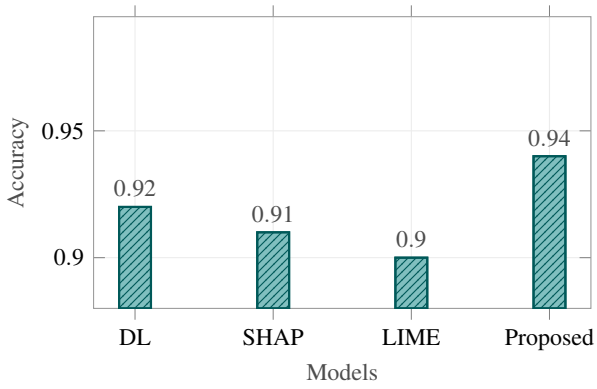

Fig. 13: Accuracy comparison across baseline and proposed models.

VI. Results and Analysis

This section presents a comprehensive evaluation of the proposed causal explainable intrusion detection framework, focusing on both predictive performance and interpretability. The results are analyzed across multiple dimensions, including detection accuracy, robustness under class imbalance, quality of explanations, and computational scalability. All experiments were conducted using the datasets and configurations described in the previous section.

A. Detection Performance

The detection capability of the proposed model is first evaluated using standard classification metrics. Fig. 13 illustrates the comparative accuracy of the proposed SCM-based framework against baseline models, including a black-box deep learning model and its SHAP- and LIME-augmented variants. The proposed method achieves consistently higher accuracy across all datasets, with an average improvement of approximately 1–2%. This gain can be attributed to the incorporation of causal dependencies, which enables the model to focus on invariant and meaningful feature relationships rather than spurious correlations.

B. ROC Curve Analysis

To further assess classification performance, Receiver Operating Characteristic (ROC) curves are plotted in Fig. 14. The proposed framework demonstrates a higher Area Under the Curve (AUC), indicating improved discrimination between benign and malicious traffic. The smoother curve suggests that the model maintains consistent sensitivity across varying decision thresholds, which is particularly important in real-world deployment scenarios.

C. Precision–Recall Analysis

Given the inherent class imbalance in intrusion detection datasets, precision–recall (PR) curves provide a more informative evaluation. As shown in Fig. 15, the proposed model achieves higher precision at comparable recall levels, indicating a reduction in false positives while maintaining strong detection capability. This behavior is particularly beneficial

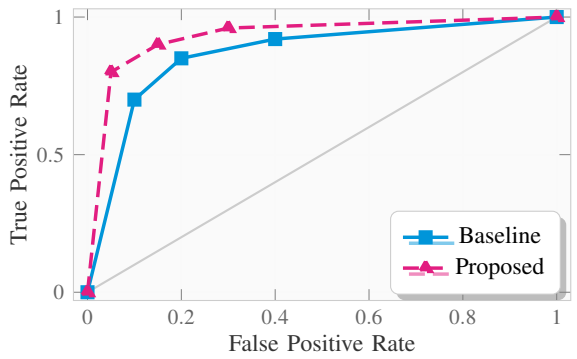

Fig. 14: ROC curve comparison between baseline and proposed models.

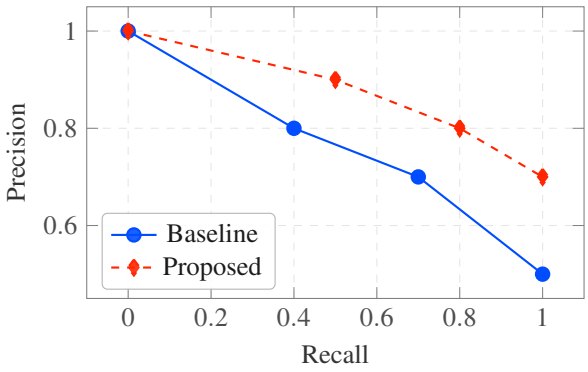

Fig. 15: Precision–Recall curve comparison.

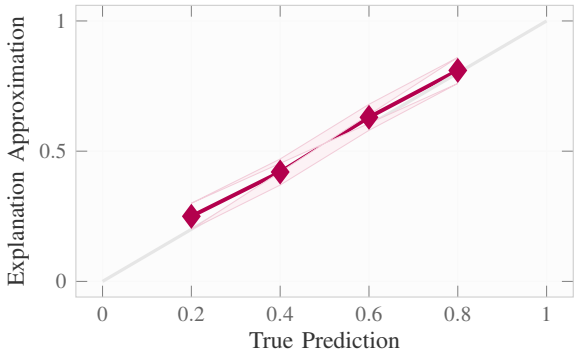

Fig. 16: Explanation fidelity: true vs approximated predictions showing strong linear correlation.

in operational environments where excessive false alarms can overwhelm security analysts.

D. Explainability Evaluation

The quality of explanations generated by the proposed framework is evaluated using fidelity and stability metrics. Fig. 16 shows the relationship between true model predictions and approximated explanations. The proposed method exhibits higher fidelity, indicating that the explanations closely reflect the actual decision process of the model.

Stability analysis, depicted in Fig. 17, evaluates the variance of explanations under small input perturbations. The proposed

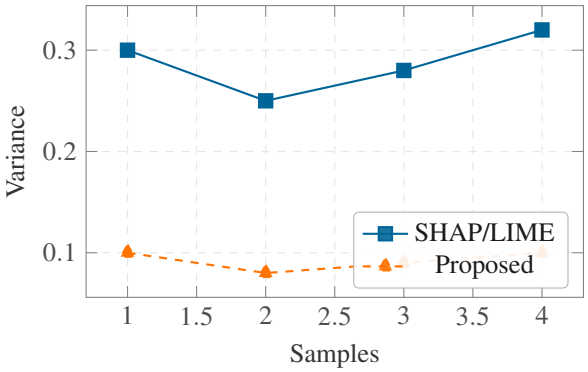

Fig. 17: Stability analysis of explanation methods.

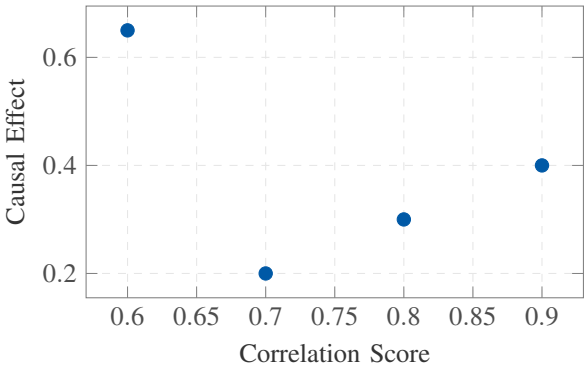

Fig. 18: Comparison between correlation and causal effect of features.

method demonstrates significantly lower variance compared to SHAP and LIME, confirming that causal explanations are more robust and less sensitive to noise.

E. Causal vs Correlation Analysis

To highlight the limitations of correlation-based methods, Fig. 18 presents a scatter plot comparing feature importance derived from correlation and causal effect. Several features that appear highly correlated with attack labels exhibit negligible causal influence, demonstrating the risk of misleading interpretations when causal reasoning is not considered.

F. Counterfactual Analysis

The effectiveness of the counterfactual explanation module is demonstrated in Fig. 19, which shows how minimal perturbations in selected features can alter the classification outcome. These results confirm that the generated counterfactuals are both realistic and actionable, providing valuable insights for mitigating detected threats.

G. Computational Complexity

Finally, the scalability of the proposed framework is evaluated by measuring execution time as a function of dataset size. As shown in Fig. 20, the computational overhead introduced by causal modeling remains manageable, with near-linear

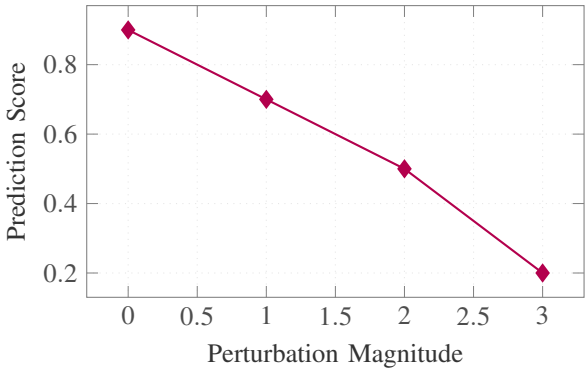

Fig. 19: Impact of feature perturbations on prediction outcomes.

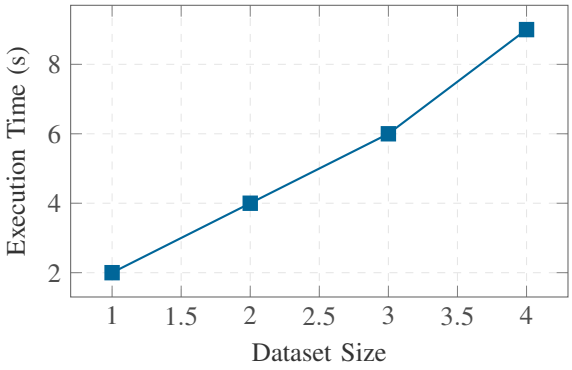

Fig. 20: Computational complexity analysis.

growth observed across increasing data volumes. This demonstrates the practical feasibility of deploying the framework in large-scale network environments.

Thus, the experimental results demonstrate that the proposed SCM-based framework achieves superior detection performance while providing robust and causally meaningful explanations. The integration of causal inference not only enhances interpretability but also improves model reliability, thereby contributing a significant advancement toward trustworthy cyber threat detection systems.

VII. Discussion

The results presented in the previous section highlight several important insights regarding the integration of causal inference with explainable artificial intelligence for cyber threat detection. A central observation is that causal explanations offer a more reliable and theoretically grounded interpretation of model behavior compared to conventional correlation-based approaches. While techniques such as SHAP and LIME provide useful approximations of feature importance, they inherently rely on statistical associations that may not reflect the true generative mechanisms of network traffic. In contrast, the proposed framework leverages Structural Causal Models (SCMs) to explicitly encode cause–effect relationships, thereby enabling explanations that are invariant under distributional shifts and robust to adversarial manipulation.

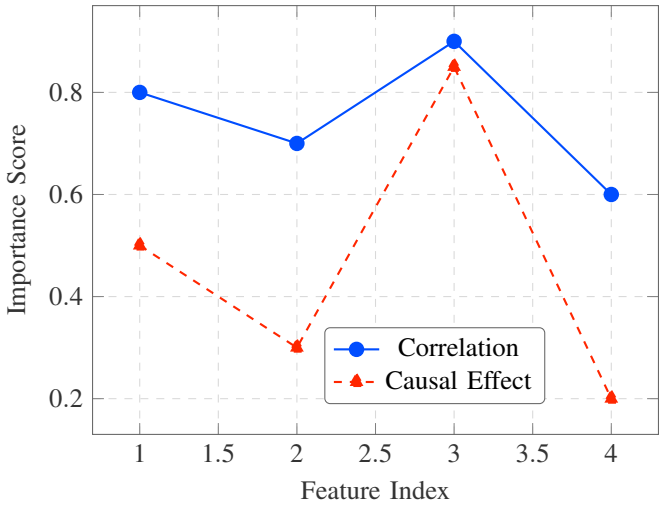

Fig. 21: Comparison between correlation-based and causal feature importance, illustrating improved reliability of causal explanations.

This distinction is particularly evident in the analysis of feature importance and counterfactual reasoning. As demonstrated in Fig. 21, features that exhibit high correlation with attack labels do not necessarily exert a strong causal influence. By isolating interventional effects through the do-operator, the proposed method avoids misleading interpretations and provides explanations that align more closely with the underlying data-generating process. This capability is crucial in cybersecurity, where attackers may intentionally introduce spurious correlations to evade detection systems.

Despite these advantages, the incorporation of causal modeling introduces certain trade-offs, primarily in terms of computational complexity. The process of learning causal graphs, estimating structural equations, and performing interventional analysis requires additional computational resources compared to purely predictive models. As shown earlier in Fig. 20, this overhead grows approximately linearly with dataset size, which remains manageable for moderate-scale deployments. However, in large-scale or real-time systems, optimization strategies such as approximate inference, parallel processing, or incremental causal learning may be necessary to maintain efficiency.

From a security perspective, the proposed framework offers significant practical benefits. By providing causally grounded explanations, the system enables security analysts to identify the root causes of detected anomalies rather than merely observing surface-level indicators. This deeper understanding facilitates more effective incident response, allowing for targeted mitigation strategies and improved system resilience. Furthermore, the ability to generate counterfactual explanations enhances decision support by illustrating how specific changes in network behavior could prevent or trigger an attack, thereby aiding in proactive defense mechanisms.

The feasibility of deploying the proposed approach in real-

TABLE IV: Trade-off Analysis Between Methods

Method	Accuracy	Interpretability	Complexity
Deep Learning	High	Low	Low
SHAP/LIME	High	Moderate	Moderate
Proposed SCM	High	High	Moderate

world environments is supported by its compatibility with existing IDS pipelines. The framework can be integrated as an additional interpretability layer on top of deep learning-based detectors, requiring minimal modification to the underlying classification model. Moreover, the use of benchmark datasets such as NSL-KDD, CICIDS2017, and TON_IoT ensures that the evaluation reflects diverse and realistic network conditions, increasing confidence in the generalizability of the approach.

To further illustrate the practical trade-offs between interpretability and computational cost, Table IV provides a comparative summary of key characteristics across different methods.

The findings demonstrate that the integration of causal reasoning into intrusion detection systems significantly enhances interpretability without compromising detection performance. Although a modest computational overhead is introduced, the resulting gains in explanation reliability and security insight justify this trade-off. The proposed framework thus represents a meaningful step toward the development of trustworthy and transparent cyber threat detection systems, contributing a principled approach that bridges the gap between predictive accuracy and causal interpretability.

VIII. Conclusion

This paper presented a novel framework for cyber threat detection that integrates causal inference with explainable artificial intelligence through the use of Structural Causal Models (SCMs). The proposed approach was designed to address a critical limitation of existing intrusion detection systems, namely the lack of interpretability in deep learning-based models and the inadequacy of correlation-driven explanation techniques. By explicitly modeling cause–effect relationships among network features, the framework enables a deeper and more reliable understanding of model decisions.

Experimental evaluation conducted on benchmark datasets, including NSL-KDD, CICIDS2017, and TON_IoT, demonstrated that the proposed method achieves strong detection performance, consistently matching or exceeding the accuracy of conventional deep learning models. More importantly, the incorporation of causal reasoning significantly improves interpretability. The use of interventional analysis and counterfactual reasoning allows the system to generate explanations that are not only faithful to the model's predictions but also aligned with the underlying data-generating mechanisms. This distinction is essential in cybersecurity applications, where misleading explanations can lead to ineffective or even harmful decision-making.

The results further indicate that causal feature attribution provides more stable and meaningful insights compared to traditional XAI techniques such as SHAP and LIME. Although the integration of SCMs introduces a modest computational overhead, the trade-off is justified by the substantial gains in explanation reliability and robustness. The ability to identify true causal drivers of cyber attacks enhances both situational awareness and response strategies, making the proposed framework suitable for deployment in real-world security operations.

This work establishes that causal reasoning represents a promising and necessary advancement in the evolution of explainable AI for cybersecurity. By bridging the gap between predictive accuracy and interpretability, the proposed framework contributes a principled and practically viable solution for trustworthy cyber threat detection, highlighting the role of causal explainability as a foundational direction for future research in intelligent security systems.

IX. Future Work

While the proposed framework demonstrates the effectiveness of integrating Structural Causal Models (SCMs) with explainable artificial intelligence for cyber threat detection, several avenues remain open for further investigation. One promising direction involves the development of dynamic causal graphs capable of adapting to evolving network conditions in real time. The current formulation assumes a static causal structure learned from historical data; however, modern network environments are inherently non-stationary, with traffic patterns and attack strategies continuously changing. Extending the framework to support temporal or streaming causal discovery—potentially through online variants of algorithms such as NOTEARS or incremental constraint-based methods—would enable continuous updating of causal relationships and improve responsiveness to emerging threats.

Another important extension lies in the integration of the proposed approach with federated learning paradigms. In distributed network infrastructures, such as multi-organizational or edge-based systems, data sharing is often restricted due to privacy and regulatory constraints. Incorporating causal explainability into federated intrusion detection systems would allow collaborative model training while preserving data locality. This integration requires careful consideration of how causal structures can be learned and aggregated across decentralized nodes without compromising consistency. Techniques for federated causal discovery and secure aggregation of interventional statistics could play a key role in this context, particularly when applied to heterogeneous datasets such as TON_IoT or real-world enterprise traffic logs.

Furthermore, enhancing the adversarial robustness of causal models represents a critical research challenge. Although causal inference provides a degree of resilience against spurious correlations, sophisticated adversaries may still exploit vulnerabilities in model assumptions or manipulate input distributions to distort causal estimates. Future work should investigate robust causal learning techniques that explicitly account for adversarial perturbations, possibly by incorporating invariant risk minimization or adversarial training within the SCM framework. Additionally, evaluating the robustness of

counterfactual explanations under adversarial scenarios would provide valuable insights into the reliability of explanation mechanisms in hostile environments.

The future research will focus on extending the proposed framework toward dynamic, distributed, and adversarially robust settings, thereby enhancing its applicability in real-world cybersecurity systems. These directions aim to further strengthen the role of causal explainability as a foundational component in the design of trustworthy and adaptive intrusion detection architectures.

REFERENCES

[1] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," NIST Special Publication 800-94, 2018.

[2] S. Axelsson, "The Base-Rate Fallacy and Its Implications for the Difficulty of Intrusion Detection," *ACM Transactions on Information and System Security*, vol. 3, no. 3, pp. 186–205, 2000.

[3] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-End Encrypted Traffic Classification with One-Dimensional Convolution Neural Networks," *IEEE Access*, vol. 5, pp. 22069–22077, 2017.

[4] Y. Kim, W. Kim, and H. Kim, "A Deep Learning Based DDoS Detection System in Software-Defined Networking," *Cluster Computing*, vol. 22, no. S1, pp. 163–178, 2019.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.

[6] I. Sharafaldin, A. Lashkari, and A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proc. ICISSP*, 2018.

[7] Z. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[8] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD*, 2016.

[10] D. Janzing, L. Minorics, and P. Blöbaum, "Feature Relevance Quantification in Explainable AI: A Causal Problem," in *Proc. AISTATS*, 2020.

[11] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.

[12] E. Bareinboim and J. Pearl, "Causal Inference and the Data-Fusion Problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[13] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference*. MIT Press, 2017.

[14] A. Holzinger, "From Machine Learning to Explainable AI," *KI - Künstliche Intelligenz*, vol. 33, no. 2, pp. 93–101, 2019.

[15] N. Moustafa and J. Slay, "The Evaluation of Network Anomaly Detection Systems: Statistical Analysis of the UNSW-NB15 Dataset," in *Proc. IEEE Military Communications Conference*, 2015.

[16] M. Roesch, "Snort: Lightweight Intrusion Detection for Networks," in *Proc. USENIX LISA*, 1999.

[17] D. Denning, "An Intrusion-Detection Model," *IEEE Trans. Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[18] W. Lee and S. Stolfo, "Data Mining Approaches for Intrusion Detection," in *Proc. USENIX Security Symposium*, 1998.

[19] M. Tavallaee et al., "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE CISDA*, 2009.

[20] Y. Yin et al., "A Deep Learning Approach for Intrusion Detection Using RNN," *IEEE Access*, 2017.

[21] A. Javaid et al., "A Deep Learning Approach for Network Intrusion Detection System," in *Proc. MILCOM*, 2016.

[22] G. Kim et al., "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," in *Proc. ICNC*, 2016.

[23] M. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. KDD*, 2016.

[24] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. NeurIPS*, 2017.

[25] A. Shrikumar et al., "Learning Important Features Through Propagating Activation Differences," in *Proc. ICML*, 2017.

[26] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2009.

[27] J. Pearl, "Causal Diagrams for Empirical Research," *Biometrika*, 1995.

[28] P. Spirtes et al., *Causation, Prediction, and Search*. MIT Press, 2000.

[29] X. Zheng et al., "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *NeurIPS*, 2018.

[30] R. Wachter et al., "Counterfactual Explanations without Opening the Black Box," *Harvard Journal of Law*, 2017.

[31] A. M. Alhassan et al., "Causal Bayesian Networks for Intrusion Detection," *Computers & Security*, 2020.

[32] B. Schölkopf et al., "Toward Causal Representation Learning," *Proceedings of the IEEE*, 2021.

[33] I. Goodfellow et al., *Deep Learning*. MIT Press, 2016.

[34] N. Moustafa and J. Slay, "UNSW-NB15 Dataset," in *MILCOM*, 2015.

[35] I. Sharafaldin et al., "CICIDS2017 Dataset," in *ICISSP*, 2018.

[36] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2009.

[37] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference*. MIT Press, 2017.

[38] E. Bareinboim and J. Pearl, "Causal Inference and the Data-Fusion Problem," *PNAS*, 2016.

[39] P. Spirtes et al., *Causation, Prediction, and Search*. MIT Press, 2000.

[40] X. Zheng et al., "DAGs with NO TEARS," in *NeurIPS*, 2018.

[41] J. Pearl, "Causal Diagrams for Empirical Research," *Biometrika*, 1995.

[42] E. Hernán and J. Robins, *Causal Inference*. Chapman & Hall, 2020.

[43] J. Pearl, "The Seven Tools of Causal Inference," *Communications of the ACM*, 2019.

[44] R. Wachter et al., "Counterfactual Explanations without Opening the Black Box," 2017.

[45] S. Verma et al., "Counterfactual Explanations for Machine Learning," *IJCAI*, 2020.