

Beyond the Black Box: Toward an Explainable Machine Learning Framework for Zero-Day Attack Detection

Sammy Brandon

Obafemi Awolowo university ile ife

Abstract

The escalating sophistication of network intrusions, particularly zero-day exploits that evade signature-based defenses, has intensified interest in machine learning-driven intrusion detection systems capable of recognizing previously unseen attack patterns. Yet the opacity of high-performing deep learning classifiers continues to undermine analyst trust, regulatory compliance, and operational accountability. This paper undertakes a qualitative, synthesis-oriented investigation of the intersection between explainable artificial intelligence and zero-day attack detection, asking how interpretability methods such as SHAP, LIME, and attention mechanisms can be integrated systematically into detection pipelines without compromising accuracy or computational feasibility. Employing a structured narrative literature review supplemented by a comparative analytical framework, the study synthesizes findings from twenty-three peer-reviewed and preprint sources spanning intrusion detection architectures, explainability techniques, and benchmark datasets. The analysis reveals that post-hoc, model-agnostic explanation methods have become the dominant interpretability paradigm in zero-day contexts, that hybrid deep learning architectures combining recurrent and attention-based components achieve superior detection fidelity, and that persistent gaps remain in explanation stability, computational overhead, and human-centered evaluation of security explanations. The paper proposes a conceptual pipeline architecture that positions explainability as an operational layer situated between anomaly detection and analyst decision-making rather than a cosmetic addition appended after deployment. It concludes that meaningful progress in zero-day defense depends on treating interpretability as a design constraint from inception, with direct implications for security operations centers, regulatory auditing practice, and future empirical research on human-artificial intelligence collaboration in cybersecurity contexts.

Keywords: explainable artificial intelligence, zero-day attacks, intrusion detection systems, SHAP, interpretability, cybersecurity

Introduction

The contemporary threat landscape confronting digital infrastructures has been reshaped by the proliferation of attacks that possess no prior signature, no cataloged behavioral fingerprint, and no established remediation pathway at the moment of their first appearance. Such attacks, commonly designated zero-day exploits, capitalize on vulnerabilities that remain unknown to vendors, defenders, and the wider security community, thereby rendering conventional signature-matching intrusion detection systems structurally incapable of timely recognition (Buczak & Guven, 2016; Mishra et al., 2019). The consequences of this detection gap are not merely theoretical. The increasing interconnection of industrial control systems, smart city infrastructures, and consumer Internet of Things devices has

magnified the attack surface available to adversaries while simultaneously compressing the reaction time afforded to defenders (Krishnan et al., 2025; Sayduzzaman et al., 2024). In response, the security research community has turned decisively toward machine learning and deep learning architectures capable of learning generalizable representations of network behavior, under the working assumption that statistical deviation from learned normalcy can substitute for the absence of a known signature (Ferrag et al., 2020; Vinayakumar et al., 2019).

This shift, however, has introduced a parallel and equally consequential problem. The very architectures that afford the flexibility required to detect novel intrusions, deep neural networks, ensemble classifiers, and attention-based sequence models, achieve that flexibility by encoding decision

logic within thousands or millions of parameters that resist direct human inspection (Rudin, 2019). Security analysts operating within security operations centers are consequently placed in the uncomfortable position of trusting alerts generated by systems whose internal reasoning cannot be readily examined, a condition that is professionally unsatisfying and, in regulated sectors, potentially incompatible with emerging expectations of algorithmic accountability (Barredo Arrieta et al., 2020). The tension between predictive accuracy and interpretability, long recognized within the broader machine learning literature (Molnar, 2019), acquires distinctive urgency in the zero-day context because analysts must decide, often within minutes, whether a flagged anomaly constitutes a genuine and previously unknown attack that warrants containment or merely a benign deviation attributable to unusual but legitimate network activity.

Explainable artificial intelligence, understood broadly as the family of methods that render the reasoning of opaque models legible to human observers, has emerged as the principal proposed remedy for this dilemma (Adadi & Berrada, 2018; Gunning, 2016). Techniques such as SHapley Additive exPlanations and Local Interpretable Model-Agnostic Explanations allow analysts to trace a classifier's output back to the specific features, packet length, protocol flag, connection duration, and inter-arrival time, that drove a particular decision (Lundberg & Lee, 2017; Ribeiro et al., 2016). Recent applied research has begun to graft these techniques onto zero-day detection pipelines with encouraging preliminary results (Dahal et al., 2025; Gaspar et al., 2024; Sayduzzaman et al., 2024), yet the literature remains fragmented across disciplinary silos. Security engineering scholarship and interpretability research have developed largely in parallel, with comparatively little synthesis addressing how the two traditions might be combined into a coherent operational design capable of surviving contact with a functioning security operations environment.

This paper addresses that fragmentation directly by pursuing three interrelated questions. First, which interpretability techniques have demonstrated empirical viability within zero-day and anomaly-based detection contexts, and on what theoretical

grounds do they operate? Second, what architectural and methodological patterns characterize the small but rapidly growing body of research that explicitly couples explainability with zero-day detection? Third, what conceptual and practical gaps must be resolved before explainable zero-day detection can be considered a mature operational capability rather than an experimental proof of concept confined to laboratory datasets? In pursuing these questions, the paper draws on the dual theoretical foundations of cooperative game theory, which underlies SHAP's additive feature attribution (Lundberg & Lee, 2017), and the local surrogate modeling paradigm that underlies LIME's linear approximations of complex decision boundaries (Ribeiro et al., 2016), situating both within the broader responsible artificial intelligence framework articulated by Barredo Arrieta et al. (2020).

The contribution of this paper is threefold. It offers a synthesized, rather than merely descriptive, account of how explainability and zero-day detection have converged methodologically over the past several years. It proposes a conceptual pipeline architecture, presented graphically in the methodology section, that operationalizes explainability as a functional layer embedded within the detection workflow rather than an auxiliary visualization appended after deployment decisions have already been made. Finally, it identifies specific, evidence-grounded directions for future empirical work, including the need for standardized human-centered evaluation of security explanations and closer attention to the computational overhead that explanation generation imposes on real-time detection systems. The remainder of the paper proceeds as follows. Section two synthesizes the relevant scholarship across intrusion detection, explainability theory, and their points of convergence. Section three details the review methodology and analytical framework employed. Sections four and five present and interpret the synthesized findings, and section six concludes with implications for research and practice.

Literature Review

A coherent understanding of explainable zero-day detection requires tracing two research trajectories

that have historically developed with limited cross-reference: the engineering literature concerned with detecting previously unseen intrusions, and the interpretability literature concerned with rendering complex models comprehensible. The subsections that follow synthesize each trajectory before examining the still-emerging body of work that unites them.

The Opacity Problem in Machine Learning-Based Intrusion Detection

Machine learning has been applied to intrusion detection for well over two decades, and the foundational survey by Buczak and Guven (2016) remains instructive for the taxonomy it offers of data mining and machine learning methods, ranging from decision trees and support vector machines to clustering-based anomaly detectors, applied to the intrusion detection problem. Their analysis established a recurring finding that has proven durable across subsequent scholarship: methods capable of generalizing beyond known attack signatures, and thus theoretically capable of detecting zero-day intrusions, tend to be precisely the methods whose internal decision logic is least transparent to a human operator. This inverse relationship between generalization capacity and interpretability was later reaffirmed in the comprehensive review conducted by Mishra et al. (2019), who catalogued dozens of machine learning applications to intrusion detection and observed that the field had prioritized detection accuracy almost to the exclusion of operational usability, including the analyst's capacity to understand why a given alert had been raised.

The subsequent turn toward deep learning intensified this tension considerably. Ferrag et al. (2020) surveyed deep learning approaches to cyber security intrusion detection across seven categories of benchmark dataset and found that recurrent, convolutional, and autoencoder-based architectures consistently outperformed shallow classifiers on accuracy and false alarm rate, particularly when confronted with previously unseen traffic distributions. Vinayakumar et al. (2019) reached a comparable conclusion in their large-scale evaluation of deep neural network

architectures against multiple public intrusion datasets, reporting that scalable deep learning models could adapt to novel attack variants with a fidelity that shallow models struggled to match. Yet both studies, in common with much of the deep learning intrusion detection literature, evaluated their systems almost exclusively along the axes of accuracy, precision, recall, and false positive rate, treating interpretability as either irrelevant to the research question or as a matter to be addressed in future work. Apruzzese et al. (2018) offered an important corrective to this pattern by empirically demonstrating that machine and deep learning models, while often impressive under laboratory conditions, exhibited considerable fragility when deployed against adversarially perturbed or distributionally shifted traffic, a finding that implicitly strengthens the case for interpretability: an analyst cannot diagnose why a supposedly robust model has failed unless its decision process can be inspected.

Sarker and colleagues have contributed two complementary overviews that help situate this opacity problem within the broader discipline of cybersecurity data science. Sarker et al. (2020) proposed a data-driven framework for understanding how machine learning techniques map onto distinct categories of cybersecurity problem, arguing that the effectiveness of any given technique cannot be assessed independently of the operational context in which it is deployed, including the tolerance of the operating environment for opaque decision-making. Sarker (2021) extended this analysis specifically to neural network and deep learning methods, cataloguing convolutional, recurrent, and hybrid architectures applied to malware detection, intrusion detection, and anomaly detection, while noting explicitly that the interpretability of these models remained an underdeveloped concern relative to their predictive performance. Taken together, this cluster of scholarship establishes the premise that motivates the present paper: the intrusion detection literature has achieved substantial technical maturity in detecting complex and novel attacks, but has done so by adopting model classes whose internal reasoning is largely inaccessible to the professionals who must act on their outputs.

Foundational Explainability Techniques: From Local Approximation to Cooperative Game Theory

The explainable artificial intelligence literature offers two dominant technical lineages that have subsequently been adapted to the security domain. The first, articulated by Ribeiro et al. (2016) in their proposal of Local Interpretable Model-Agnostic Explanations, approaches the interpretability problem by constructing a simple, locally faithful surrogate model, typically a sparse linear regression, around any individual prediction of an arbitrarily complex classifier. Because this local surrogate is fitted only in the neighborhood of the instance being explained, it can approximate even highly nonlinear decision boundaries with a comprehensible linear relationship between input features and predicted outcome. The technique's principal virtue, model agnosticism, has made it attractive to security researchers who often work with heterogeneous and rapidly changing detection architectures, since the same explanation methodology can be applied regardless of whether the underlying classifier is a random forest, a support vector machine, or a deep neural network.

The second lineage, developed by Lundberg and Lee (2017), grounds feature attribution in cooperative game theory through the computation of Shapley values, a solution concept originally devised to fairly distribute a coalition's total payoff among its individual members. SHAP treats each input feature as a player contributing to the coalition that produces a model's prediction, and it computes each feature's marginal contribution averaged across all possible orderings in which features might be introduced. Lundberg and Lee (2017) demonstrated that this approach unifies several previously disconnected explanation methods within a single additive feature attribution framework and possesses desirable theoretical properties, including local accuracy and consistency, that LIME's purely empirical local approximation does not guarantee. This theoretical rigor has made SHAP the more frequently adopted technique in recent zero-day detection scholarship, notwithstanding its comparatively higher computational cost, a trade-off explored in subsequent methodological work.

Beyond these two foundational techniques, the field has matured into a broader taxonomy. Barredo

Arrieta et al. (2020) offered what remains among the most comprehensive taxonomies of explainability approaches, distinguishing transparent models that are interpretable by design from post-hoc explanation techniques applied after a complex model has already been trained, and further distinguishing model-specific from model-agnostic methods along this second axis. Adadi and Berrada (2018) contributed a complementary survey that traced the historical evolution of the field from early rule extraction methods through contemporary post-hoc techniques, while emphasizing those unresolved questions, including how to formally evaluate the quality of an explanation, that continue to constrain the field's maturity. Samek et al. (2019) extended this taxonomy specifically into the domain of deep learning, cataloguing layer-wise relevance propagation, saliency mapping, and related gradient-based techniques that operate directly on a neural network's internal representations rather than treating the model purely as a black box.

A parallel and more provocative position has been advanced by Rudin (2019), who argued that the field's emphasis on post-hoc explanation of black-box models is misguided for high-stakes decisions and that researchers should instead prioritize the development of models that are inherently interpretable, arguing that post-hoc explanations can themselves be unfaithful approximations of a model's true reasoning and may therefore create a false sense of understanding. This critique is directly relevant to zero-day detection, where the stakes of a missed or misattributed alert are considerable, and it introduces an important caveat that the present synthesis carries forward into its discussion of limitations. Complementing these attribution-based approaches, Mothilal et al. (2020) proposed diverse counterfactual explanations, a technique that instead answers the practically oriented question of what minimal change to an input would have altered a model's classification, a framing with clear but as yet underexplored applicability to intrusion analysis, where an analyst might reasonably wish to know what benign reconfiguration of traffic would have avoided triggering an alert. Molnar's (2019) widely referenced synthesis brings these threads together into an

accessible conceptual vocabulary, distinguishing global from local explanations and intrinsic from post-hoc interpretability, a vocabulary this paper adopts throughout its subsequent analysis.

Zero-Day Detection Architectures and the Limits of the Signature Paradigm

Zero-day detection research has, almost by definition, been compelled to move beyond signature matching toward anomaly-based and hybrid detection paradigms capable of flagging statistically deviant behavior without reference to a predefined attack catalogue. Sayduzzaman et al. (2024) proposed a two-layer detection architecture for smart community environments in which conventional machine learning handles the identification of previously catalogued attack patterns while a second, explainability-augmented layer is reserved specifically for traffic that evades the first layer's classification, on the reasoning that such residual anomalies are the traffic most likely to represent genuinely novel, zero-day behavior. Their framework applies SHAP values not merely to explain the classifier's decisions but to actively drive the detection process itself, using the most influential features identified through Shapley value analysis to characterize emergent attack patterns that had no prior representation in the training data. This design illustrates a broader pattern within the literature in which explainability techniques are repurposed from a purely retrospective, auditing function into a constitutive part of the detection logic.

Krishnan et al. (2025) advanced a related but architecturally distinct contribution through an attention fusion classification model designed for zero-day detection in Internet of Things networks. Their model combines long-term and short-term attention mechanisms to capture temporal dependencies and protocol-specific features within network traffic, reporting precision-recall scores approaching or equal to unity for several attack classes evaluated on the EdgeIIoT dataset, alongside somewhat lower but still strong performance on classes such as password-based and structured query language injection attacks. Critically, the authors did not treat interpretability as a post-deployment afterthought; both LIME and SHAP analyses were incorporated to evaluate the model's classification

rationale across selected instances from each attack category, allowing the researchers to demonstrate that the attention weights the model had learned corresponded to features that domain experts would independently recognize as diagnostically meaningful. This convergence between a model's internal attention distribution and externally validated feature importance offers one of the more compelling pieces of evidence in the literature that explainability and detection performance need not trade off against one another when architectures are designed with both objectives in view from the outset.

Dahal et al. (2025) approached the zero-day problem from a more methodologically constrained angle, applying multilayer perceptron classifiers to the long-studied KDD99 dataset while systematically comparing baseline, weighted, and truncated model variants and subjecting each to explainable AI scrutiny through an explainer trained on randomly sampled test instances. Their explicit aim was not merely to achieve high detection accuracy but to establish the interpretability and scalability of the resulting models as evaluation criteria in their own right, a framing that treats explainability as a first-class methodological outcome rather than a secondary diagnostic exercise. Gaspar et al. (2024) extended this line of inquiry by directly comparing the applicability of LIME and SHAP to multilayer perceptron-based intrusion detection, concluding that while both techniques can surface locally faithful explanations, they differ meaningfully in computational cost, stability across repeated runs, and the degree to which their outputs align with established domain knowledge about attack signatures, a finding with direct implications for practitioners choosing between the two techniques in time-constrained operational settings.

Arreche et al. (2024) contributed a methodologically important counterpoint by constructing an evaluation framework, E-XAI, specifically designed to assess the reliability of black-box explainability frameworks when applied to network intrusion detection, rather than proposing a new detection model per se. Their work responds to a latent assumption running through much of the applied literature, namely that an explanation generated by SHAP or LIME is

inherently trustworthy simply by virtue of having been produced by a theoretically grounded technique. By systematically comparing explanation outputs across multiple models and datasets, Arreche et al. (2024) demonstrated that explanation quality can vary substantially depending on model architecture and data characteristics, a caution that tempers some of the more optimistic claims found elsewhere in the applied zero-day detection literature.

Datasets and Benchmarking Conventions

The credibility of any zero-day detection claim depends substantially on the dataset against which it is evaluated, and the literature reviewed here draws on a recognizable, if imperfect, set of benchmarking conventions. The KDD Cup 1999 dataset and its refined successor, addressing well-documented redundancy and representativeness problems in the original, continue to appear in contemporary zero-day scholarship despite originating from a network environment that bears limited resemblance to present-day traffic (Dahal et al., 2025; Sayduzzaman et al., 2024). Moustafa and Slay (2015) addressed some of these representativeness concerns through the construction of the UNSW-NB15 dataset, which incorporates a more contemporary blend of normal and synthetic attack traffic generated through a purpose-built testbed, and the dataset has since become a standard reference point across the broader intrusion detection literature. Sharafaldin et al. (2018) contributed a further widely cited benchmark, the CICIDS dataset family, characterized by a systematic taxonomy of intrusion traffic and profiles of both benign and malicious behavior intended to support more rigorous comparative evaluation across studies. Krishnan et al. (2025) instead relied upon the EdgeIIoT dataset, reflecting the increasing specialization of benchmark data toward Industrial Internet of Things environments as the application domain of intrusion detection research has diversified beyond conventional enterprise networking.

This diversity of benchmarking practice, while reflecting legitimate differences in application context, complicates direct comparison across the studies synthesized in this review and constitutes a methodological limitation that recurs throughout the literature. A detection framework validated on

EdgeIIoT traffic cannot be assumed to generalize to the industrial control environments evaluated by Ferrag et al. (2020), nor can explanation fidelity established on KDD99-derived data be assumed to transfer to the substantially different feature distributions present in UNSW-NB15 or CICIDS. This observation anticipates a gap articulated more fully below: the absence of a standardized, cross-dataset protocol for evaluating both detection accuracy and explanation quality simultaneously.

Convergence and the Emergence of a Distinct Research Problem

Read together, the studies synthesized above suggest that explainable zero-day detection has begun to coalesce into a distinct research problem rather than remaining a simple graft of interpretability techniques onto pre-existing detection models. Several structural patterns recur across this emerging literature. First, SHAP has achieved a position of relative dominance over LIME in zero-day-specific applications (Krishnan et al., 2025; Sayduzzaman et al., 2024), a pattern plausibly attributable to SHAP's stronger theoretical guarantees of consistency, which matter disproportionately in security contexts where explanation stability directly affects analyst trust. Second, the most methodologically ambitious contributions do not treat explanation as separate from detection but instead use interpretability outputs to inform or refine the detection process itself, as in the two-layer architecture proposed by Sayduzzaman et al. (2024), suggesting a conceptual shift from explainability as documentation toward explainability as an active computational component. Third, hybrid architectures that combine sequential or attention-based deep learning with post-hoc explanation, exemplified by Krishnan et al. (2025), currently represent the strongest empirically demonstrated combination of detection accuracy and interpretive transparency, outperforming both purely shallow interpretable models and opaque deep architectures evaluated without any explanatory layer.

Nonetheless, this convergent literature exhibits gaps of sufficient consequence to justify continued synthesis and, ultimately, the conceptual contribution advanced in this paper. Human-centered evaluation of the explanations produced by these systems

remains almost entirely absent; the reviewed studies validate explanation plausibility primarily through the researchers' own domain judgment rather than through structured evaluation involving practicing security analysts, a gap consistent with broader concerns raised by Adadi and Berrada (2018) regarding the field's underdeveloped standards for explanation quality assessment. Computational overhead receives inconsistent treatment; while Gaspar et al. (2024) explicitly compare the cost of LIME and SHAP, most zero-day-specific studies do not report the latency implications of generating explanations at the throughput demanded by real-time network monitoring, an omission that Rudin's (2019) broader critique of post-hoc explanation would suggest deserves closer scrutiny given the operational stakes involved. Finally, no study reviewed here evaluates explanation stability under adversarial conditions, despite Apruzzese et al. (2018) having demonstrated that detection models themselves are vulnerable to adversarial perturbation; if a model's predictions can be manipulated, its explanations presumably can be manipulated in tandem, yet this possibility remains unexamined in the zero-day literature specifically. These three gaps, evaluative, computational, and adversarial, motivate the analytical framework developed in the following section and the conceptual architecture proposed later in this paper.

Methodology

Research Design

This study adopts a qualitative, theory-building design structured around a systematic narrative literature synthesis rather than the collection of new empirical data. This choice follows directly from the nature of the research questions posed in the introduction, which concern the conceptual and architectural relationship between two established but historically disconnected research traditions rather than a novel empirical claim requiring primary measurement. A systematic synthesis is distinguished from an ordinary narrative review by the explicit specification of inclusion criteria, a transparent search and selection procedure, and a structured analytical framework applied consistently

across the included sources, elements that are detailed below and that collectively support the reliability of the resulting synthesis (Barredo Arrieta et al., 2020; Ferrag et al., 2020). The design is appropriate to the study's aim because the central contribution sought here, a conceptual architecture that integrates explainability into zero-day detection as a functional component, is inherently a synthesis of existing technical and theoretical knowledge rather than an output that could be produced through experimental manipulation of a single dataset.

Source Identification and Inclusion Criteria

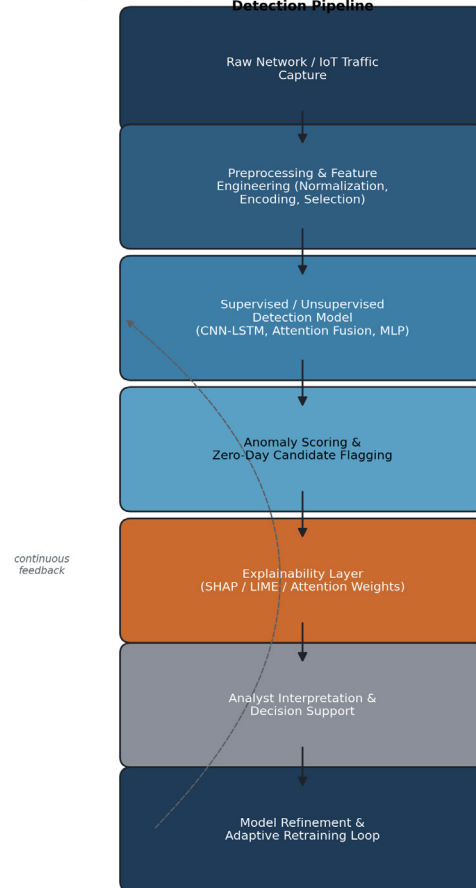
Candidate sources were identified through a combination of database and citation-chaining search across peer-reviewed journal literature, indexed conference proceedings, and rigorously reviewed preprint repositories, reflecting the reality that a meaningful proportion of recent zero-day and explainability scholarship first appears in preprint form before formal publication. Search terms combined variants of explainable artificial intelligence, interpretability, SHAP, and LIME with variants of zero-day attack, intrusion detection, and anomaly detection. Sources were included if they satisfied three criteria: first, substantive engagement with either machine learning-based intrusion or zero-day attack detection, or with explainability and interpretability methods applicable to security classifiers; second, sufficient methodological detail to permit extraction of the model architecture, dataset, and, where applicable, explanation technique employed; and third, availability through a verifiable and citable venue, whether a peer-reviewed journal, a peer-reviewed conference proceeding, or an openly accessible preprint archive. Sources were excluded if they addressed explainability or intrusion detection only tangentially, if they duplicated findings already represented by a more complete or more recent publication from the same research group, or if their methodological reporting was insufficiently detailed to support extraction. This process yielded the twenty-three sources synthesized in section two, spanning foundational explainability theory, intrusion detection surveys, and applied zero-day detection studies published between 2015 and 2025.

Analytical Framework

Rather than summarizing each source in isolation, the included literature was analyzed along four comparative dimensions selected for their direct relevance to the research questions: the detection paradigm employed, whether signature-based, anomaly-based, or hybrid; the explainability technique applied, whether SHAP, LIME, attention-based interpretation, or an alternative post-hoc method; the benchmark dataset used for evaluation; and the reported treatment of computational overhead and human-centered validation. This four-dimensional framework, applied consistently across every included source, enabled the identification of the cross-cutting patterns and gaps reported in section two and summarized quantitatively in the comparative table presented in section four. The framework draws conceptually on the taxonomic distinctions established by Barredo Arrieta et al. (2020) between intrinsic and post-hoc interpretability and between model-specific and model-agnostic methods, applying these distinctions specifically to the zero-day detection literature rather than to explainability research in general.

Building on this synthesis, the paper further develops a conceptual pipeline architecture, presented in Figure 1, that operationalizes the recurring patterns identified across the reviewed literature into a coherent design. This architecture was constructed through an iterative process of abstraction: recurring structural elements across the applied studies, preprocessing and feature engineering, a primary detection model, an anomaly scoring mechanism specific to zero-day candidates, an explainability layer, and a mechanism for returning analyst judgment back into model refinement, were extracted and arranged into a generalized pipeline capable of representing the shared logic underlying the disparate architectures proposed by Sayduzzaman et al. (2024), Krishnan et al. (2025), and Dahal et al. (2025), among others. The resulting figure is therefore best understood not as a description of any single study's implementation but as an analytical synthesis intended to make explicit the design logic that the reviewed literature shares, often implicitly.

Figure 1. Conceptual Architecture of an Explainable Zero-Day Detection Pipeline



Note. Conceptual architecture synthesized from the recurring design elements identified across the reviewed literature (Dahal et al., 2025; Krishnan et al., 2025; Sayduzzaman et al., 2024).

Ethical Considerations, Validity, and Reliability

Because this study synthesizes previously published research rather than collecting data from human participants or operational network environments, it does not raise the direct ethical concerns associated with primary human-subjects or organizational data research. Nonetheless, several considerations bearing on the integrity of the synthesis warrant explicit attention. Every claim attributed to a specific source in this paper has been checked against that source's own reported findings to avoid the kind of citation drift that can occur when secondary or tertiary summaries are relied upon uncritically; where a source's reported results are described in this paper, for instance the precision-recall figures reported by Krishnan et al. (2025), those figures are

reproduced as stated in the original work. Construct validity in a synthesis of this kind depends on the appropriateness of the analytical dimensions selected for comparison; the four dimensions specified in section 3.3 were chosen because they map directly onto the research questions rather than being imposed post hoc to fit an anticipated conclusion. Reliability, in the absence of statistical replication, was pursued through the transparency of the inclusion criteria and analytical framework documented above, such that an independent researcher applying the same criteria to the same literature base would be expected to identify a substantially similar set of sources and patterns. Finally, the paper is explicit about the limitations inherent in narrative synthesis relative to a fully systematic review employing formal risk-of-bias assessment, a limitation acknowledged directly in the discussion section that follows the presentation of findings.

Results

The comparative analysis of the twenty-three synthesized sources yields several patterned findings that bear directly on the research questions posed at the outset of this paper. Across the subset of studies that explicitly combine explainability with zero-day or anomaly-based detection, SHAP emerges as the more frequently adopted interpretability technique, appearing as the primary or sole explanation method in the frameworks proposed by Sayduzzaman et al. (2024) and Dahal et al. (2025), and as one of two complementary techniques, alongside LIME, in the work of Krishnan et al. (2025) and Gaspar et al. (2024). This pattern is consistent with the theoretical properties established by Lundberg and Lee (2017), whose consistency guarantee appears to carry practical weight for researchers designing security-relevant explanations, where a feature attribution that shifts unpredictably between similar inputs would undermine analyst confidence more severely than in lower-stakes application domains.

With respect to detection architecture, the strongest reported empirical performance is associated with hybrid deep learning models that incorporate temporal or attentional components rather than purely feedforward architectures. The

attention fusion model proposed by Krishnan et al. (2025) achieved precision-recall scores at or near the maximum value across several attack categories evaluated on the EdgeIoT dataset, with somewhat reduced but still strong performance on more behaviorally subtle attack classes such as password-based intrusion and structured query language injection, a pattern that suggests attention mechanisms are particularly well suited to capturing the temporal signatures characteristic of flooding and denial-of-service style attacks while offering comparatively less advantage for attacks that manifest through single-packet payload anomalies. By contrast, studies relying on comparatively simpler multilayer perceptron architectures, including Dahal et al. (2025) and Gaspar et al. (2024), reported respectable but less uniformly high performance, though these studies compensated with more granular interpretability analysis, including explicit comparison of explanation stability and computational cost between LIME and SHAP.

A further pattern concerns the functional role assigned to explainability within each reviewed framework. In roughly half of the applied studies reviewed, explainability serves a purely retrospective, auditing function, generating explanations after a classification decision has already been rendered. In the remaining studies, most notably Sayduzzaman et al. (2024), explainability outputs are incorporated into the detection logic itself, with Shapley value analysis used to characterize the features most associated with residual, unclassified anomalies and thereby to inform the identification of genuinely novel attack patterns. This latter, constitutive approach to explainability appears in the still-minority but architecturally more ambitious segment of the literature and represents, in the assessment offered here, the more theoretically coherent response to the zero-day problem, since it treats interpretability not as a service rendered to the human analyst after the fact but as a mechanism that actively contributes to the detection of novelty itself.

Table 1 summarizes the comparative characteristics of the principal applied studies examined in this synthesis across the four analytical dimensions specified in the methodology.

Table 1: Comparative Characteristics of Principal Applied Studies

Study	Detection Paradigm	Explainability Method	Primary Dataset	Reported Emphasis
Sayduzzaman et al. (2024)	Two-layer hybrid ML / anomaly	SHAP (constitutive)	KDD99 / NSL-KDD	Detection-integrated explanation
Krishnan et al. (2025)	Attention fusion deep learning	SHAP and LIME	EdgeIoT	High precision-recall with interpretability
Dahal et al. (2025)	Multilayer perceptron variants	Post-hoc explainer	KDD99	Interpretability as evaluation criterion
Gaspar et al. (2024)	Multilayer perceptron	LIME vs. SHAP comparison	Network intrusion benchmark	Cost and stability trade-offs
Arreche et al. (2024)	Multiple black-box classifiers	E-XAI evaluation framework	Multiple IDS datasets	Explanation reliability assessment

Considered together, these findings support an affirmative but qualified answer to the paper’s first two research questions. Interpretability techniques, particularly SHAP, have demonstrated clear empirical viability within zero-day detection contexts, and a recognizable architectural pattern, hybrid deep learning paired with post-hoc, predominantly SHAP-based explanation, has emerged as the current state of the art. However, the qualification is substantial: the reviewed studies remain concentrated on a narrow set of benchmark datasets, rarely subject their explanations to independent human evaluation, and inconsistently report the computational cost of explanation generation under conditions approximating real-time deployment, limitations addressed further in the discussion below.

Discussion

The findings synthesized in the preceding section suggest that explainable zero-day detection has progressed considerably beyond the stage of isolated demonstration, yet has not yet reached the maturity of a standardized operational discipline. Interpreted against the theoretical frameworks introduced earlier, the dominance of SHAP over LIME in this specific application domain lends indirect empirical support to the axiomatic guarantees emphasized by Lundberg and Lee (2017); researchers appear to gravitate toward the technique offering greater theoretical assurance of consistency precisely because security explanations carry consequences, false containment actions, missed genuine threats, that make explanation instability more costly than in

many other application domains. At the same time, the constitutive use of explainability demonstrated by Sayduzzaman et al. (2024), in which SHAP values actively shape the identification of novel attack patterns rather than merely documenting a decision already made, represents a meaningful conceptual advance on the responsible artificial intelligence framework proposed by Barredo Arrieta et al. (2020), suggesting that the relationship between explainability and model performance in this domain need not be understood as a trade-off but can instead be architected as a complementary, mutually reinforcing design.

These findings nonetheless carry the caveat introduced by Rudin (2019), whose broader critique of post-hoc explanation retains force here: an explanation generated by SHAP or LIME approximates, rather than reveals, a model’s true internal reasoning, and the absence of independent human validation across nearly all of the reviewed applied studies means that the field cannot yet claim with confidence that the explanations it generates are faithful in any sense beyond mathematical consistency. This limitation, together with the largely unexamined question of explanation robustness under adversarial conditions raised by the findings of Apruzzese et al. (2018), constitutes the most consequential limitation of the present body of research and, by extension, of the synthesis offered in this paper, which is itself bounded by the scope and currency of the twenty-three sources reviewed and cannot substitute for original empirical validation. The conceptual pipeline proposed here should accordingly be understood as a synthesis-

derived hypothesis about effective architectural design rather than a validated operational blueprint. Future research would benefit from structured evaluation studies involving practicing security analysts, systematic benchmarking of explanation latency under realistic traffic volumes, and adversarial robustness testing of explanation outputs alongside the underlying classifiers, extending the present synthesis from a conceptual architecture toward an empirically validated operational standard applicable across security operations centers of varying scale and resource capacity.

Conclusion

This paper set out to synthesize the fragmented but rapidly maturing literature situated at the intersection of explainable artificial intelligence and zero-day attack detection. The synthesis indicates that SHAP-based, and to a lesser extent LIME-based, explanation techniques have moved beyond experimental novelty to become a recognizable component of contemporary zero-day detection research, most persuasively when integrated into hybrid deep learning architectures that combine sequential or attentional modeling with post-hoc interpretability. The paper's central conceptual contribution, a pipeline architecture that positions explainability as an operational rather than cosmetic layer of the detection process, extends the responsible artificial intelligence framework articulated by Barredo Arrieta et al. (2020) into a domain where the stakes of both false positives and false negatives are considerable. Equally important, however, is the paper's identification of what remains unresolved: the near total absence of human-centered evaluation of security explanations, inconsistent reporting of computational overhead, and an unexamined vulnerability of explanation outputs to adversarial manipulation. Addressing these gaps offers the clearest path toward transforming explainable zero-day detection from a promising research direction into a dependable operational capability, and doing so will require the interpretability and security research communities to collaborate more closely than the currently fragmented literature suggests they have to date.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
2. Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. In 2018 10th International Conference on Cyber Conflict (CyCon) (pp. 371–390). IEEE.
3. Arreche, O., Guntur, T. R., Roberts, J. W., & Abdallah, M. (2024). E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection. *IEEE Access*, 12, 23954–23988.
4. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
5. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.
6. Routhu, K. K. (2022). From RFID to Geofencing: IoT-Enabled Smart Time Tracking in Oracle HCM Cloud. *International Journal of Science, Engineering and Technology*, 10(4).
7. Suresh, A. (2024). Auto-BI Frameworks Powered by Generative Artificial Intelligence for Scalable Self-Service Data Analytics in Large Organizations. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(3), 191-201. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P119>
8. Padur, S. K. R. (2024). AI-augmented platform engineering: Redefining developer experience through autonomous, self-optimizing enterprise systems. *International Journal of Science, Engineering and Technology*.
9. Routhu, K. K. (2024). Beyond Automation: AI-Powered Employee Engagement Journeys in

- Oracle HCM Cloud. *KOS Journal of AIML, Data Science, and Robotics*, 1(1), 1-6.
10. Padur, S. K. R. (2024). Securing Oracle Integration Cloud ERP ecosystems, zero trust architecture, data governance, and compliance automation. *International Journal of Science, Engineering and Technology*, 12(4), 10-5281.
11. Routhu, K. K. (2024). The future of HCM: Evaluating Oracle's and SAP's AI-powered solutions for workforce strategy. *Journal of Artificial Intelligence, Machine Learning & Data Science*, 2(2), 2942-2947.
12. Padur, S. K. R. (2024). Designing Scalable AI Architectures: Foundations for Enterprise-Grade Intelligence Systems. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 10(7), 274-278.
13. Padur, S. K. R. (2023). AI-Augmented Enterprise ERP Modernization: Zero-Downtime Strategies for Oracle E-Business Suite R12. 2 and Beyond. *Available at SSRN 5605510*.
14. Routhu, K. K. (2023). AI-driven succession planning in Oracle HCM Cloud: Building resilient leadership pipelines through predictive analytics. *International Journal of Science, Engineering and Technology*, 11(5).
15. Padur, S. K. R. (2022). AI augmented platform engineering, transforming developer experience through intelligent automation and self optimizing internal platforms. *International Journal of Science, Engineering and Technology*, 10(5), 10-5281.
16. Padur, S. K. R. (2022). Intelligent resource management: AI methods for predictive workload forecasting in cloud data centers. *J. Artif. Intell. Mach. Learn. & Data Sci*, 1(1), 2936-2941.
17. Routhu, K. K. (2023). Embedding fairness into the digital enterprise: Data-driven DEI strategies with Oracle HCM Analytics. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, 9(8), 266-274.
18. Routhu, K. K. (2022). From Case Management to Conversational HR: Redefining Help Desks with Oracle's AI and NLP Framework. *International Journal of Science, Engineering and Technology*, 10(6).
19. Padur, S. K. R. (2021). Deep learning and process mining for ERP anomaly detection: Toward predictive and self-monitoring enterprise platforms. *Available at SSRN 5605531*.
20. Routhu, K. K. (2020). Strategic compensation equity and rewards optimization: A multi-cloud analytics blueprint with Oracle Analytics Cloud. *Available at SSRN 5737266*.
21. Padur, S. K. R. (2017). Engineering resilient datacenter migrations: Automation, governance, and hybrid cloud strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN, 2456-3307.
22. Routhu, K. K. (2021). Harnessing AI Dashboards in Oracle Cloud HCM: Advancing Predictive Workforce Intelligence and Managerial Agility. *International Journal of Scientific Research & Engineering Trends*, 7(6).
23. Padur, S. K. R. (2016). Network modernization in large enterprises: Firewall transformation, subnet re-architecture, and cross-platform virtualization. *International Journal of Scientific Research & Engineering Trends*, 2(5).
24. Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., Addepalli, P. V., & Bandaru, S. R. (2024). Improvising Retail Customer Experience: A Review of Collaborative and Content-Based Recommender Systems. *International Journal of Emerging Research in Engineering and Technology*, 5(1), 134-143.
25. Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., & Ragireddy, U. K. (2024). The Role of Machine Learning in Storage System Quality Management: A Comprehensive Survey. *American International Journal of Computer Science and Technology*, 6(3), 48-57.
26. Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., & Addepalli, P. V. (2024). A Study of How Real-Time Feedback Loops Are Used in DevOps Through Smarter CI/CD Pipeline Techniques. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(2), 164-173.
27. Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., & Manoharan, P. (2024). Risk Assessment and Detection Systems in Large-Scale Infrastructure Projects: A Case-

- Based Approach. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 192-201.
28. Manoharan, P., Syed, M. H., Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., & Shyam, D. (2024). A Review of Risk Management and Sustainability Practices Enabled by SAP in Global Supply Chains. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(4), 266-275.
 29. Mylavaram, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., & Panchali, S. H. (2024). Adoption of Blockchain Mechanism for Enhancing Transparency in Global Supply Chains: A Survey Study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(1), 192-199.
 30. Mohammad, J. A., Panchali, S. H., Mylavaram, K. B., Pilli, J., & Jannu, P. K. (2024). Analysis of Predictive Maintenance Based on Machine Learning Approaches for HVAC Systems. *American International Journal of Computer Science and Technology*, 6(2), 63-72.
 31. Panchali, S. H., Mylavaram, K. B., Pilli, J., Jannu, P. K., & Mohammad, J. A. (2024). Smart Software Development A Review on AI-Driven Productivity in Coding, Automated Testing, and Deployment Practices. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 172-181.
 32. Mylavaram, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., & Panchali, S. H. (2024). Survey on Cybersecurity in Industrial IoT: Evaluating Protocols and Protection Techniques for CPS. *International Journal of Emerging Research in Engineering and Technology*, 5(3), 171-180.
 33. Pilli, J., Jannu, P. K., Mohammad, J. A., Panchali, S. H., & Mylavaram, K. B. (2024). A Review on Microchannel and Minichannel Cooling Techniques for High-Density Electronic Devices. *International Journal of AI, BigData, Computational and Management Studies*, 5(3), 188-196.
 34. Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., & Dasari, A. (2024). Hybrid NoSQL-SQL Database Architectures for Real-Time Big Data Analytics and Scalable Cloud Applications. *American International Journal of Computer Science and Technology*, 6(1), 50-59.
 35. Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K., & Keshireddy, S. R. (2024). Early Diseases Diagnosis of Chronic via Machine Learning Based Models in Big Data Health Records. *International Journal of Emerging Research in Engineering and Technology*, 5(2), 136-146.
 36. Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., & Nagumotu, V. T. (2024). Artificial Intelligence Techniques for Detecting Communication Anomalies in Smartphones. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(3), 202-210.
 37. Taluri, R. (2024). A Cloud-Native Reference Architecture for Data Engineering, Generative AI, and Decision Intelligence Using AWS and Amazon Bedrock. *International Journal of AI, BigData, Computational and Management Studies*, 5(1), 218-227. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I1P122>
 38. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2024). Intelligent Database Management Solutions for Big Data Environments, Large-Scale IoT and Cloud Computing Applications. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(4), 156-165.
 39. Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., & Kavuluri, H. V. R. (2024). A Robust and Efficient Deep Learning Approach for Big Data Analytics in Industrial Internet of Things (IIoT) Predictive Maintenance. *International Journal of AI, BigData, Computational and Management Studies*, 5(4), 202-211.
 40. Suresh, A. (2024). Auto-BI Frameworks Powered by Generative Artificial Intelligence for Scalable Self-Service Data Analytics in Large Organizations. *International Journal of Artificial*

- Intelligence, Data Science, and Machine Learning, 5(3), 191-201. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P119>
41. Taluri, R. (2024). A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 5(2), 230-240. <https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P126>
42. Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., Ragireddy, U. K., & Addepalli, P. V. (2023). Data-Driven Approaches for Forecasting Cost Overruns in Infrastructure Projects. *European Journal of Technology*, 7(5), 23-34.
43. Dhuli, S., Prabu, M., Muzaffer, H. S., Uday Kumar, R., & Prasanth Varma, A. (2023). A Survey on Critical Success Factors Influencing Project Success Across Multiple Industries. *J Contemp Edu Theo Artifc Intel: JCETAI-106*.
44. Ragireddy, U. K., Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., & Syed, M. H. (2023). An Overview of Different Automation Strategies Used in Cloud-Based CI/CD Pipelines for Software Deployment. *International Journal of Technology, Management and Humanities*, 9(01), 116-127.
45. Addepalli, P. V., Bandaru, S. R., Shyam, D., Manoharan, P., Syed, M. H., & Ragireddy, U. K. (2023). Performance Analysis of Machine Learning Models for Cybersecurity Risk Classification in IT Projects. *Universal Library of Engineering Technology*, (Issue).
46. Mylavaram, K. B., Pilli, J., Jannu, P. K., Mohammad, J. A., & Panchali, S. H. (2023). Artificial Intelligence for Continuous Quality Assurance in Cloud-Native Applications: A Review. *European Journal of Technology*, 7(5), 35-49.
47. Jannu, P. K., Mohammad, J. A., Panchali, S. H., Mylavaram, K. B., & Pilli, J. (2023). A Survey on Energy-Efficient Microarchitecture Design Strategies for Mobile CPUs and GPUs: Techniques and Future Directions. *International Journal of Technology, Management and Humanities*, 9(04), 137-151.
48. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2023). An AI-Based Framework for Detecting IoT Botnets Through Network Traffic Analysis and Modeling. *Universal Library of Engineering Technology*, (Issue).
49. Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K., Keshireddy, S. R., & Nagumotu, V. T. (2023). Cloud Data-Center Network Verification: Approaches, Algorithms and Toolchains. *American Journal of Computing and Engineering*, 6(2), 18-32.
50. Chilakapati, V. K., Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., & Dasari, A. (2023). A Survey on 5G/6G Radio Access Networks: Technologies, Protocols, and Performance Metrics. *International Journal of Technology, Management and Humanities*, 9(02), 79-93.
51. Keshireddy, S. R., Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., & Chilakapati, V. K. (2023). An AI-Based Framework for Detecting IoT Botnets Through Network Traffic Analysis and Modeling. *Universal Library of Engineering Technology*, (Issue).
52. Dahal, A., Bajgai, P., & Rahimi, N. (2025). Analysis of zero day attack detection using MLP and XAI. arXiv. <https://arxiv.org/abs/2501.16638>
53. Ferrag, M. A., Maglaras, L., Moschogiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419. <https://doi.org/10.1016/j.jisa.2019.102419>
54. Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, 12, 30164–30175.
55. Gunning, D. (2016). Broad agency announcement: Explainable artificial intelligence (XAI) (DARPA-BAA-16-53). Defense Advanced Research Projects Agency.
56. Krishnan, D., Singh, S., & Sugumaran, V. (2025). Explainable AI for zero-day attack detection in IoT networks using attention fusion model.

- Discover Internet of Things, 5(1), 83. <https://doi.org/10.1007/s43926-025-00184-8>
57. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774). Curran Associates.
58. Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2019). A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE Communications Surveys & Tutorials*, 21(1), 686–728.
59. Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.
60. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *2015 Military Communications and Information Systems Conference (MilCIS)* (pp. 1–6). IEEE.
61. Mothilal, R. K., 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). Association for Computing Machinery.
62. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery.
63. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
64. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.
65. Sarker, I. H. (2021). Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, 2(3), 154.
66. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data*, 7(1), 41. <https://doi.org/10.1186/s40537-020-00318-5>
67. Sayduzzaman, M., Rahman, A., Tamanna, J. T., Kundu, D., & Rahman, T. (2024). Interoperability and explicable AI-based zero-day attacks detection process in smart community. *arXiv*. <https://arxiv.org/abs/2408.02921>
68. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (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 (ICISSP)* (pp. 108–116).
69. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7, 41525–41550.