

Deep Learning for Proxy Zero-Day Intrusion Detection: A Systematic Review, Taxonomy, and Evaluation Framework

Maggie Zita

De Montfort University

Mujeeb Ur Rehman

`mu.jeeb.rehman@dmu.ac.uk`

De Montfort University

Muhammad Kazim

De Montfort University

Sohail Khalid

University of Management and Technology

Muhammad Abrar

De Montfort University

Eerke Boiten

De Montfort University

Research Article

Keywords: Zero-day attack, Deep learning, Intrusion detection, Proxy evaluation, Explainability, Systematic review

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11096990/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Deep Learning for Proxy Zero-Day Intrusion Detection: A Systematic Review, Taxonomy, and Evaluation Framework

Margaret Zita¹, Mujeeb Ur Rehman^{1*}, Muhammad Kazim¹,
Sohail Khalid², Muhammad Abrar¹, Eerke Boiten¹

^{1*}Department of Computer Science and Informatics, De Montfort University, Leicester, UK.

²Department of Computer Science, University of Management and Technology, Lahore, Pakistan.

*Corresponding author(s). E-mail(s): mujeeb.rehman@dmu.ac.uk;

Abstract

True zero-day exploits are rarely observable in public intrusion-detection benchmarks, so empirical studies typically evaluate novelty through proxy settings. This systematic review examines how deep learning-based intrusion detection systems operationalize proxy zero-day detection and identifies requirements for defensible comparison. Five corpus-forming databases were searched through 30 April 2026, with supplementary coverage-validation searches in Scopus and Web of Science in July 2026. The corpus-forming search and citation chaining identified 1,101 records; the two validation searches returned 251 raw records and are reported separately because record-level exports were not available for combined deduplication and screening. Corrected venue and eligibility audits retained 37 peer-reviewed empirical primary studies and classified 63 records as background or excluded evidence. We synthesize supervised, unsupervised, self-supervised, hybrid, graph, transformer, federated, few-shot, and reinforcement-learning approaches; characterize datasets and proxy strategies; and assess robustness, explainability, and deployment readiness. Fifteen studies reported an architecture-specific absolute accuracy or F1-score suitable for descriptive quantitative synthesis. Formal meta-analysis was not appropriate because variance estimates, matched comparators, and consistent proxy protocols were generally absent. Reported benchmark performance was often high but was not directly comparable across studies and depended strongly on dataset choice, proxy definition, and reporting completeness. Cross-dataset, temporal, open-set,

adversarial, explainability, and resource evaluation remained limited. We propose a seven-dimension framework linking architecture, dataset context, proxy validity, metrics, robustness, explainability, and deployment readiness, while distinguishing proxy-novelty robustness from true vulnerability-level zero-day discovery.

Keywords: Zero-day attack, Deep learning, Intrusion detection, Proxy evaluation, Explainability, Systematic review

1 Introduction

A zero-day attack exploits hardware or software vulnerabilities that are unknown to the vendor or the public at the time of exploitation (Yin et al., 2017). As a result, vendors have effectively “zero days” to issue a patch once the exploitation begins. Empirical analysis by Bilge and Dumitras (2012) shows that zero-day vulnerabilities remain undisclosed for an average of 312 days, during which exploitation activity can persist undetected, often intensifying after public disclosure.

The lifecycle of a zero-day vulnerability runs from its inadvertent introduction during development, through discovery by adversaries or defenders and active exploitation, to eventual public disclosure and remediation; detection systems must operate throughout this window in the absence of signatures or vendor patches. Because true exploitation of an as-yet-undisclosed vulnerability is what this lifecycle actually describes, and such exploitation is rarely observable in public benchmarks (Section 1.2), empirical evaluation instead relies on the four proxy strategies formally defined in Section 1.2 and summarized in Table 3. Figure 1 situates these four proxies on a single validity ladder, ordered by how closely each approximates true zero-day exploitation, which sits at the top of the ladder as the (largely unattainable) reference point against which every proxy in this review is a stress test rather than a proof.

Expanding on this life cycle perspective, a typical zero-day attack detection pipeline involves traffic monitoring, feature extraction, representation learning, and decision-making components that collectively identify anomalous or previously unseen attack behavior during the pre-disclosure window.

The growing severity of the threat landscape is reflected in reports from both industry threat-intelligence teams and market-analysis firms: Google’s Threat Analysis Group has documented year-on-year increases in publicly catalogued zero-day exploits used in the wild through 2023 (Google Threat Analysis Group, 2025), while separate industry estimates of aggregate cybercrime cost, which include but are not limited to zero-day exploitation, project continued growth over the remainder of the decade (Statista, 2025). These are two conceptually distinct series (actual exploit counts versus a broader economic cost projection) and are not combined into a single figure here to avoid implying a single trend line where none has been independently verified for this review; interested readers should consult the primary sources cited.

Traditional intrusion detection systems (IDSs) rely on signature-based approaches, which can achieve high precision for known threats but cannot detect attacks for

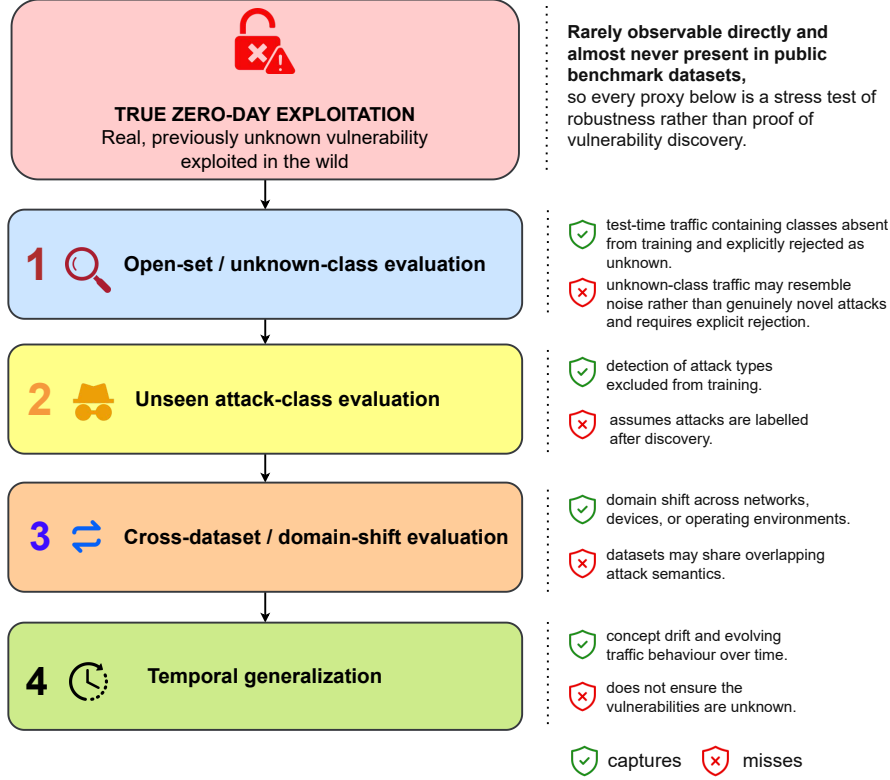


Fig. 1: Proxy zero-day validity ladder: the four proxy evaluation strategies used throughout this review (Section 1.2, Table 3), ordered by closeness to true zero-day exploitation. True zero-day exploitation sits at the top as a largely unattainable reference point; every proxy below it is a stress test of robustness, not proof of vulnerability discovery.

which no matching signature exists (Díaz-Verdejo et al., 2022). Anomaly-based detection systems try to identify deviations from normal behavior and can detect new attacks; however, they often suffer from high false positives and significant computational overhead (Khaleel and Shiltagh, 2023; Narmadha and Balaji, 2025). DL offers a promising alternative by enabling high-capacity feature extraction and generalization from complex, high-dimensional traffic data (Hnamte and Hussain, 2023; Zhang, 2025). DL-based ZDA detectors have demonstrated encouraging performance; however, they remain limited by challenges such as poor interpretability, dataset bias, and susceptibility to adversarial manipulation (Mala et al., 2025; Zhou et al., 2022; ŞAHİN et al., 2025).

Several surveys have reviewed the role of DL in intrusion detection and, to a limited extent, in ZDA detection. Ferrag et al. (2020) provided a broad overview of DL in cybersecurity, but did not quantitatively synthesize the studies. Zhou et al. (2022) examined adversarial DL broadly, with limited focus on ZDA-specific requirements.

Zhang et al. (2022) focused on explainable AI in cybersecurity, but omitted robustness benchmarking and ethical dimensions.

These contributions are valuable, but remain fragmented, emphasizing architectures, datasets, or narrow thematic aspects in isolation. Although recent surveys and studies (2023-2025) have advanced isolated aspects such as federated deployment (Korba et al., 2023), temporal architectures (Lopes et al., 2023), and self-supervised detection (Golchin et al., 2024), no prior work has jointly integrated taxonomy, quantitative synthesis, robustness, and explainability into a unified evaluation framework.

1.1 Comparison with Existing Surveys

To position the present study within the existing literature, Table 1 compares representative surveys on DL-based intrusion detection and ZDA detection on five methodological dimensions: DL-based IDS coverage, explicit ZDA focus, zero-day proxy evaluation strategies, quantitative synthesis, and robustness and explainability considerations.

To ensure transparency and reduce subjective bias, coverage levels were assigned using predefined criteria by the authors of this review, rather than by an independent third party; readers should treat this table as the authors' own comparative judgment rather than a validated bibliometric score. *Dedicated* indicates a dedicated and systematic treatment of a topic, *Partial* indicates that the topic is discussed but not systematically analyzed or incorporated as a primary component of the survey, and *No* indicates that the topic is not explicitly addressed.

Although several surveys have been published in recent years, they typically focus on specific aspects of DL-based intrusion detection, such as model architectures, datasets, or general detection performance. Four recent reviews reinforce this pattern. Ogunseyi et al. (2026) systematically examined explainable DL-based IDS for IoT, while Neto et al. (2025) synthesized DL intrusion detection across emerging technologies. Arnob et al. (2025) surveyed IDS techniques and challenges broadly, and Chinnnasamy et al. (2025) reviewed DL-driven network IDS methods, datasets, and evaluation practices. These studies strengthen coverage of XAI, emerging environments, and general NIDS methodology, but none provides a dedicated synthesis linking explicit proxy zero-day validity, robustness, explainability, quantitative evidence, and deployment readiness. In contrast, explicit zero-day proxy evaluation strategies, quantitative evidence synthesis, robustness assessment, explainability analysis, and unified evaluation practices are generally treated independently. As shown in Table 1, existing surveys provide only partial coverage of these complementary dimensions. Motivated by this gap, the present survey provides a unified evidence-based synthesis that integrates a systematic review of 100 topically screened studies, of which the corrected venue and eligibility audits retained 37 as confirmed, on-topic, peer-reviewed empirical primary studies forming the current primary synthesis corpus (the remaining 63 retained as background evidence; Section 3.4.1), a structured taxonomy, quantitative synthesis, robustness and explainability synthesis, and an evaluation framework within a single study.

Table 1: Comparison of representative surveys on DL-based intrusion detection and proxy zero-day detection.

Reference	Year	DL-IDS	ZDA	Proxy	Quant. synth.	Rob. /XAI
Ferrag et al.	2020	D	N	N	N	N
Zhou et al.	2022	P	N	N	N	D
Zhang et al.	2022	P	N	N	N	P
Alansary et al.	2025	D	D	P	N	P
Guo	2023	D	D	P	N	N
Nair and Mhavan	2023	D	P	N	N	N
Por et al.	2024	D	D	P	D	N
Uttarkar et al.	2024	D	D	P	N	P
Mohamed et al.	2024	D	D	P	N	P
da Silva Ruffo et al.	2024	D	P	N	N	N
Khan et al.	2025	D	P	N	P	D
Manjula and Saravanakumar	2025	D	D	P	N	N
Ogunseyi et al.	2026	D	N	N	P	D
Neto et al.	2025	D	P	P	P	P
Arnob et al.	2025	D	N	N	N	P
Chinnnasamy et al.	2025	D	N	N	P	P
This survey	2026	D	D	D	D	D

Coverage criteria: D indicates dedicated and systematic treatment of the topic, including comparative analysis, methodological evaluation, or a dedicated survey section. P indicates that the topic is discussed but is not systematically analyzed or incorporated as a primary survey component. N indicates that the topic is not explicitly addressed.

Abbreviations: DL-IDS: deep learning-based intrusion detection systems; ZDA: explicit zero-day attack focus; Proxy: unseen-class, temporal-split, or cross-dataset evaluation strategies; Quant. synth.: dedicated quantitative synthesis of reported results, not necessarily a pooled meta-analysis; Rob./XAI: robustness and explainability analysis.

Coding statement: Coding was performed by the authors using the stated criteria and is intended as a positioning aid rather than an independently validated bibliometric ranking.

Building on the comparison in Table 1, this survey integrates complementary aspects of DL-based ZDA detection that have previously been examined in isolation. Specifically, it (i) develops a structured taxonomy derived from the 37-study primary synthesis corpus, (ii) performs a quantitative descriptive synthesis of leading DL architectures across widely used benchmark datasets, (iii) synthesizes current evidence on adversarial robustness and explainability, and (iv) proposes an evaluation framework for the systematic assessment of DL-based IDSs under zero-day proxy conditions.

Figure 2 provides an integrated roadmap of the review. It connects the distinction between true and proxy zero-day detection with the proxy-validity hierarchy, systematic-review workflow, evidence taxonomy, descriptive quantitative synthesis, principal evidence gaps, and the proposed seven-dimension evaluation framework. The workflow summarizes the 1,101-record screened identification set, 100 topically eligible

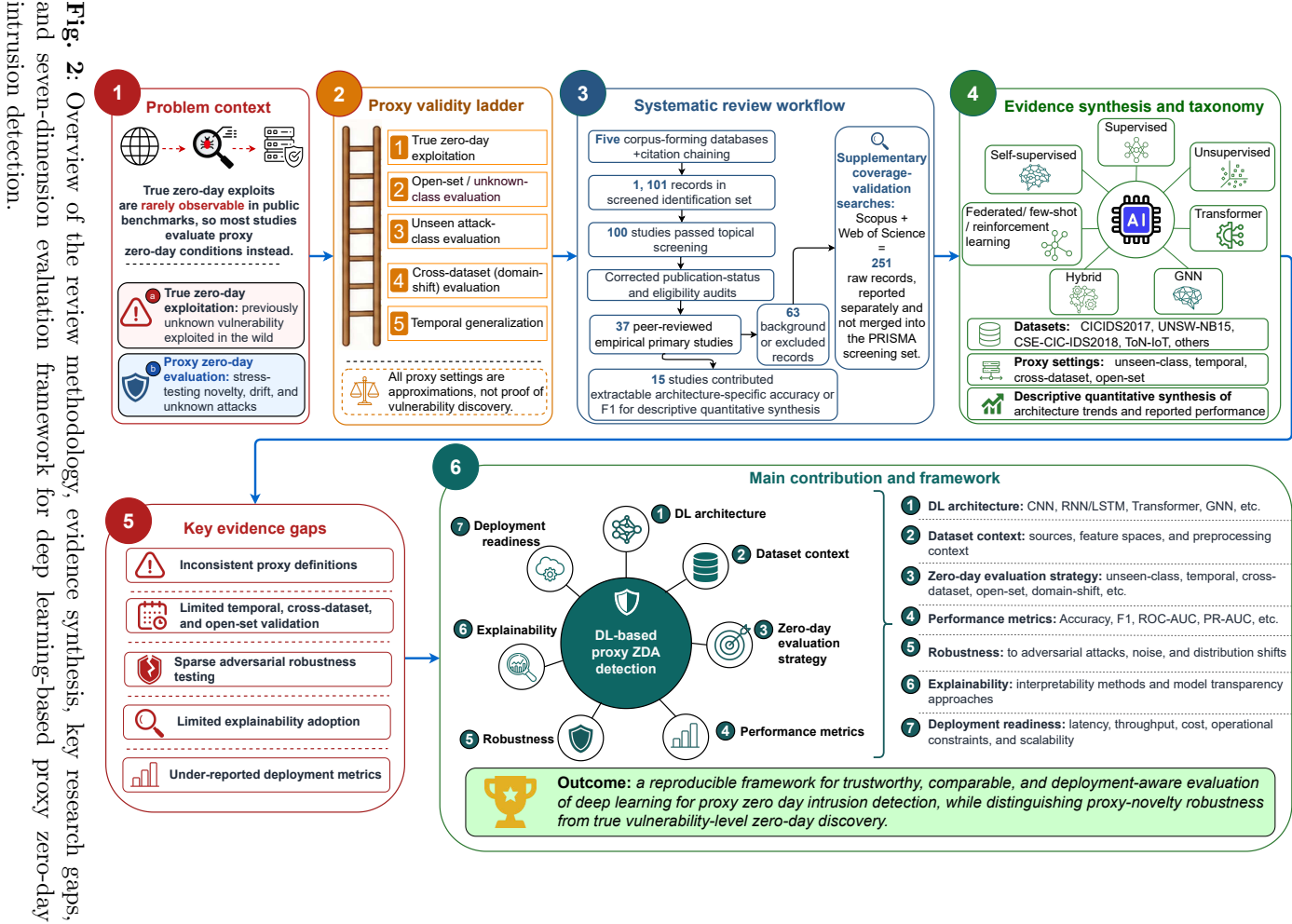
Table 2: Evaluation framework derived from the synthesis of the 37-study primary corpus on DL-based ZDA detection.

Evaluation Dimension	Representative Categories	Assessment Criteria	Research Significance
DL Architecture	CNN, RNN/LSTM, autoencoder, transformer, GNN, and hybrid architectures.	Detection capability, representation learning, and computational complexity.	Compares the suitability of different architectures for ZDA detection.
Dataset Context	CICIDS2017, UNSW-NB15, CSE-CIC-IDS2018, and ToN-IoT.	Dataset diversity, realism, class imbalance, and domain coverage.	Assesses the validity and generalizability of the reported results.
Zero-Day Evaluation	Unseen attacks, cross-dataset testing, temporal splits, and open-set evaluation.	Realism and validity of the adopted zero-day proxy conditions.	Measures the capability of a model to detect previously unseen attacks.
Performance Metrics	Accuracy, precision, recall, F1-score, and AUC.	Detection effectiveness, class-wise performance, and false-alarm behaviour.	Quantifies predictive performance and supports comparison among models.
Robustness	Adversarial testing, robust training, and distribution shifts.	Performance degradation, resilience, and recovery under perturbations.	Evaluates resistance to adaptive attacks and changing data distributions.
Explainability	SHAP, LIME, attention visualization, and TCAV.	Transparency, interpretability, and explanation fidelity.	Supports analyst trust, model auditing, and security decision-making.
Deployment Readiness	Efficiency, scalability, and edge/fog deployment.	Latency, throughput, memory consumption, and energy requirements.	Assesses the feasibility of practical and resource-constrained deployment.

studies, the corrected 37-study primary synthesis corpus, 63 background or excluded records, and the 15-study quantitative subset. The supplementary Scopus and Web of Science searches returned 251 raw records and were used for coverage validation rather than being merged into the completed PRISMA screening set.

The proposed evaluation framework (Table 2) extends beyond a conventional experimental workflow by defining seven complementary evaluation dimensions: DL architecture, data set context, zero-day evaluation strategies, performance metrics, robustness, explainability and readiness for deployment. These dimensions provide a structured basis for comparing heterogeneous studies, improving reproducibility, and identifying methodological gaps in the existing literature.

Beyond predictive performance, the framework explicitly incorporates robustness and interpretability as essential considerations for operational deployment, enabling



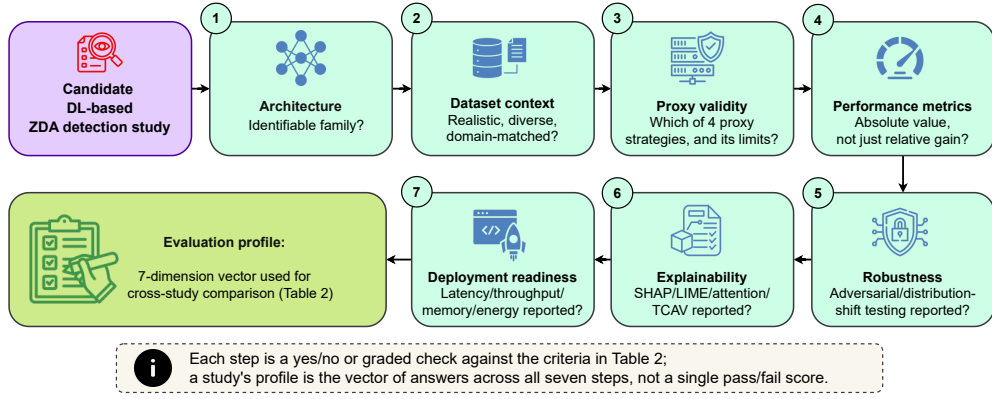


Fig. 3: Seven-dimensional evaluation framework workflow: sequential application of the seven evaluation dimensions in Table 2 to a candidate DL-based proxy-ZDA study, yielding a seven-dimensional evaluation profile.

a more holistic assessment of DL-based IDSs. In addition, this survey examines the relationship between technical performance, explainability, ethical considerations, and deployment suitability, highlighting the importance of transparency, privacy preservation, and responsible AI practices in cybersecurity applications.

Finally, by consolidating more than eight years of empirical research, this survey identifies key limitations, open challenges, and emerging research directions. The evidence synthesis highlights persistent issues related to dataset imbalance, attack diversity, and evaluation consistency, motivating future research on self-supervised representation learning, adaptive hybrid architectures, and interpretable transformer-based intrusion detection systems.

1.2 Operational Definition of ZDA Detection in This Review

In a strict security sense, a ZDA refers to the exploitation of a previously unknown vulnerability for which no patch or signature exists at the time of the attack. Such true zero-day exploits are rare, difficult to observe directly, and are generally unavailable in public benchmark datasets.

Consequently, most empirical research, including the studies reviewed in this paper, cannot directly evaluate true zero-day exploitation. Instead, prior studies have adopted proxy evaluation settings that approximate zero-day conditions under controlled assumptions. In this review, the term ZDA detection is used to denote detection under proxy zero-day conditions rather than verified exploitation of unknown vulnerabilities. Because true vulnerability-level zero-day exploitation is rarely observable in public benchmarks, most empirical studies operationalize “zero-day” through proxy evaluation settings. These proxies emulate novelty or distribution shift (for example, holding out attack types, evaluating on later time windows, cross-dataset transfer, or open-set testing), and should be interpreted as stress tests of robustness under proxy zero-day conditions rather than proof of vulnerability discovery. Specifically,

Table 3: Proxy Evaluation Settings Used to Approximate Zero-Day Conditions

Proxy Type	What It Approximates	What It Misses
True zero-day exploitation	Real exploitation of previously unknown software vulnerabilities.	Rarely observable and generally unavailable in public datasets.
Unseen attack classes	Detection of attack types that are absent from the training data.	Assumes that the attacks have already been identified and labeled after discovery.
Temporal split evaluation	Concept drift and evolving network traffic behaviour over time.	Does not ensure that the underlying vulnerabilities or attack mechanisms are genuinely unknown.
Cross-dataset evaluation	Domain shift across networks, devices, datasets, or operational environments.	Different datasets may still contain overlapping attack semantics or closely related attack families.
Open-set or unknown-class evaluation	Test-time traffic containing classes absent from the training label space, which may be rejected or identified as unknown rather than forced into a known class.	Requires an explicit rejection or unknown-class mechanism and still assumes that unknown-class traffic adequately represents genuinely novel attacks rather than benign noise or ordinary distribution shift.

we distinguish between true zero-day exploitation and the following commonly used proxies:

1. unseen attack-class evaluation, where one or more attack classes are held out from training and used only for testing;
2. temporal generalization, where models are trained on historical traffic and evaluated on future traffic distributions;
3. cross-dataset (domain-shift) evaluation, where training and testing data originate from different environments or data collection pipelines; and
4. open-set or unknown-class evaluation, where test-time traffic may include classes not represented in the training label space and must be flagged as unknown.

These proxy settings approximate different aspects of zero-day behavior, such as novelty, distributional shifts, and lack of prior signatures; however, they do not fully capture vulnerability-level exploitation. Therefore, all claims in this paper are framed in terms of robustness under proxy zero-day conditions rather than absolute zero-day vulnerability discovery. Table 3 summarizes the evaluation proxies commonly used in the literature to approximate zero-day conditions, clarifying what each proxy captures and the limitations that prevent it from representing true zero-day vulnerability exploitation.

This study makes four major contributions.

1. a formal operational distinction and validity ladder separating true zero-day exploitation from four proxy evaluation strategies, with explicit limits on the claims supported by each proxy;
2. an eligibility-audited and traceable systematic evidence base, accompanied by study-level extraction, exclusion, quality/reporting, and quantitative-source records;
3. a descriptive synthesis linking architecture families, datasets, proxy strategies, performance reporting, adversarial robustness, explainability, and deployment evidence without presenting non-comparable benchmark values as pooled effects; and
4. a seven-dimension proxy-aware evaluation framework and research agenda that converts the review findings into minimum reporting and validation requirements for future DL-based intrusion-detection studies.

The remainder of this paper is organized as follows. Section 2 presents the background concepts and related work, covering traditional intrusion detection techniques and foundational DL architectures relevant to zero-day threat scenarios. Section 3 presents a systematic review of DL-based approaches for ZDA detection under proxy evaluation settings. Section 4 presents a descriptive quantitative synthesis of the detection performance across representative benchmark datasets. Section 5 reviews the commonly used datasets for evaluating zero-day-like attack detection methods and discusses their limitations. Section 6 examines the adversarial robustness of DL-based proxies for zero-day condition systems. Section 7 discusses explainability and ethical considerations. Section 8 outlines future research directions and open challenges. Section 9 concludes the study.

2 Background and Related Work

To address the limitations of traditional detection methods in identifying ZDA, this section outlines the background of DL-based ZDA detection and the key concepts underpinning the subsequent analysis.

2.1 Traditional ZDA Detection Techniques

Prior to the broad adoption of DL, IDS primarily used traditional methods, including signature-based and anomaly-based detection processes (Otoum and Nayak, 2021). These two groups essentially reflect distinct approaches to detection: signature-based detection is based on known attack patterns, whereas anomaly-based detection is based on statistical deviations from normal behavior. A comparison of their features, operational limitations, and suitability for ZDA detection is presented in Table 4. A signature-based IDS detects intrusions by matching network traffic or system behavior with a previously defined database of known malicious signatures. Common implementations include Snort and ClamAV, which have been the mainstays of enterprise and open-source security infrastructure for years (Samita, 2024; Ravindran et al., 2025; Khraisat and Alazab, 2021). Their strengths include high accuracy, a low number of false positives, and computational efficiency via optimized pattern-matching capabilities. However, their dependence on current signature repositories renders

them ineffective against new or dynamic threats, most notably, zero-day exploits and polymorphic malware.

In contrast, an anomaly-based IDS model normalizes system or network behavior and labels deviations as potential intrusions (Otoum and Nayak, 2021; Erlacher and Dressler, 2020). Early studies applied statistical profiling, clustering, and rule-based inference to generalize feature distribution deviations, such as packet rate, protocol mix, and connection duration. Recent pre-DL methods have supplemented support vector machines (SVMs), k-nearest neighbors (KNNs), and self-organizing maps (SOMs) to generalize and reduce the need for manual tuning. Despite having more theoretical promise for ZDA detection, anomaly-based systems are often hindered by model drift, false-positive rates, and excessive computational costs owing to continuous behavioral modeling.

As summarized in Table 4, signature-based systems are accurate and fast but rigid against emerging threats, in contrast to anomaly-based methods, which are more comprehensive in terms of attack coverage and potential ZDA detection, although with a trade-off in interpretability, resource consumption, and management complexity.

Table 4: Comparison of signature-based and anomaly-based intrusion detection approaches.

Feature	Signature-Based IDS	Anomaly-Based IDS
Detection method	Pattern matching against known attack signatures (Díaz-Verdejo et al., 2022; Khraisat and Alazab, 2021).	Behavioral modeling of normal traffic patterns (Otoum and Nayak, 2021).
ZDA detection	Cannot detect previously unseen attacks (Díaz-Verdejo et al., 2022).	Has the potential to detect zero-day and novel attacks (Sarhan et al., 2023).
False positives	Provides a low false-positive rate for known attacks (Khraisat and Alazab, 2021).	Typically produces a higher false-positive rate because legitimate deviations may be classified as anomalous.
Maintenance	Requires frequent signature updates (Otoum and Nayak, 2021).	Requires less frequent updates once baseline behavior has been learned, although periodic retraining may still be necessary (Sarhan et al., 2023).
Resource usage	Incurs relatively low computational overhead (Khraisat and Alazab, 2021).	Incurs higher computational costs because of behavioral modeling, training, and inference processes.
Coverage	Provides narrow coverage limited to known attack patterns (Díaz-Verdejo et al., 2022).	Provides broader coverage and can identify unknown and evolving threats (Sarhan et al., 2023).

This dichotomy inspired the development of representation-learning-based detection systems and hybrid deep learning architectures that aim to bridge the flexibility of anomaly-based detection with the precision of signature-based methods (Bahlali et al., 2025; Golchin et al., 2024; Tong and Zhang, 2024; Duraz et al., 2024; Caville et al., 2022; Xu et al., 2024). This paradigm is further motivated in Table 5, which highlights the comparison between manual feature engineering and automatic representation learning in modern detection system design.

Table 5: Feature engineering versus representation learning in IDS.

Aspect	Feature Engineering	Representation Learning
Human involvement	High reliance on domain expertise and manual feature design (Khraisat and Alazab, 2021).	Minimal manual input, with features learned automatically from data (Vinayakumar et al., 2019).
Feature type	Explicit, handcrafted, and often interpretable features, such as statistical flow characteristics (Díaz-Verdejo et al., 2022).	Implicit latent representations or embeddings learned automatically by deep models (Shone et al., 2018).
Scalability	Limited scalability because of the need for manual feature design, selection, and tuning (Khraisat and Alazab, 2021).	Highly scalable when supported by large datasets and GPU-accelerated training (Vinayakumar et al., 2019).
Adaptability to ZDA	Typically exhibits poor generalization to previously unseen attacks (Díaz-Verdejo et al., 2022).	May provide better generalization through hierarchical, transferable, and data-driven representations (Sarhan et al., 2023).
Example in IDS	Manual encoding of packet statistics, traffic-flow characteristics, or protocol-specific rules (Otoum and Nayak, 2021).	CNNs, autoencoders, or transformers learn traffic representations directly from network data (Ullah et al., 2024).

As shown in Table 5, representation learning alleviates the scalability and generalization limitations of manual feature engineering, making it particularly suitable for ZDA detection, wherein prior signatures or handcrafted features are unavailable.

2.2 Taxonomy of DL Techniques for ZDA Detection

This section reviews recent advances in the application of DL techniques for ZDA detection. Furthermore, it provides a structured overview of existing research on key learning paradigms, architectural trends, and operational contexts. A taxonomy of DL-based approaches is presented to enable systematic comparison across methodologies with respect to their data requirements, generalization capabilities, and computational feasibility. The taxonomy in Table 6 categorizes DL-based detection models into five broad groups: supervised, unsupervised, self supervised, hybrid, and reinforcement learning. We also highlight emerging techniques (e.g., few-shot, federated, and continual learning) as extensions that address label scarcity and deployment constraints.

These paradigms are discussed not only in terms of architectural diversity (e.g., CNNs, RNNs, transformers, and autoencoders) but also in terms of their ability to accommodate realistic cybersecurity scenarios with scarce labels, evolving attack surfaces, and resource-constrained environments.

Table 6: Taxonomy derived from thematic coding of the 37-study primary synthesis corpus; references (Yin et al., 2017; Mala et al., 2025; Golchin et al., 2024; Khan et al., 2024) are provided as representative examples.

Learning Type	DL Architecture	Application Context	Notes
Supervised	CNN, RNN, and transformer	Well-labeled datasets and signature-like attack detection.	Provides high detection accuracy but may generalize poorly to previously unseen attacks.
Unsupervised	Autoencoder, variational autoencoder, and GNN	Anomaly detection primarily on normal network traffic.	Useful for ZDA detection but may produce false positives and offer limited interpretability.
Self-supervised	Contrastive models and SSL frameworks	General representation learning from large amounts of unlabeled traffic data.	Effective for pre-training but generally requires task-specific adaptation or fine-tuning.
Hybrid	CNN-LSTM and AE-transformer	Combination of spatial-temporal analysis, anomaly detection, and supervised classification.	Often provides strong performance but introduces greater architectural and computational complexity.
Reinforcement	RL agents with deep learning backbones	Dynamic policy learning for adaptive detection and response in IDS environments.	Less commonly studied but promising for adaptive and sequential security decision-making.

In general, a comparative analysis of DL paradigms for ZDA detection revealed distinct strengths and limitations in different methodological categories. Supervised learning models consistently achieve high accuracy in identifying known attack patterns because they can learn discriminative features from labeled data. However, their reliance on static decision boundaries often leads to poor generalization to previously unseen or evolving threats.

In contrast, unsupervised approaches, such as autoencoders and graph neural networks, have demonstrated strong potential for anomaly detection by modeling the normal behavior of network traffic and flagging deviations as potential intrusions (Shahid et al., 2019; Shone et al., 2018; Villegas-Ch et al., 2025). Despite this advantage, these models often lack interpretability, which limits their operational

trustworthiness and adoption in real-world cybersecurity systems. Self-supervised learning (SSL) has recently emerged as a promising alternative for representation learning in low-label or unlabeled environments, enabling models to capture intrinsic structural patterns within the data (Bahlali et al., 2025; Golchin et al., 2024; Tong and Zhang, 2024; Abdulrazzaq et al., 2024). Nevertheless, SSL-based systems often require careful fine-tuning and downstream adaptation to maintain robustness and high detection fidelity. Hybrid architectures that combine spatial-temporal learning components (e.g., CNN-LSTM) or reconstruction-classification mechanisms (e.g., AE-transformer) have achieved competitive performance and improved resilience against zero-day threats, albeit at the expense of greater architectural complexity and computational demands (Vinayak and Jarin, 2025; Walczyna et al., 2024; Maheshwari et al., 2022; Jakka and Alsmadi, 2022; Basati and Faghih, 2023; Yang et al., 2022).

Reinforcement learning remains a promising but under-explored direction, offering the ability to design adaptive, feedback-driven intrusion response systems capable of dynamic policy learning (Adawadkar and Kulkarni, 2022; Praveena et al., 2022; Benaddi et al., 2022a,b). Emerging paradigms, such as few-shot and federated learning, present new opportunities for decentralized and privacy-preserving ZDA detection, although their practical utility and scalability require further empirical validation (Yang et al., 2022; Lu et al., 2023; Althiyabi et al., 2024; Ohtani et al., 2024; Jin et al., 2023). Collectively, these observations underscore the ongoing evolution of DL-based intrusion detection from static classification to adaptive, interpretable, and resource-efficient zero-day defense mechanisms.

2.3 DL Architectures for ZDA Detection

DL architectures exhibit complementary capabilities for detecting zero-day proxy attacks by capturing spatial, temporal, relational, or representation-based characteristics of network traffic. Table 7 provides a comparative summary of their strengths, limitations, and typical application scenarios.

2.3.1 CNN-Based Intrusion Detection Models

Convolutional neural networks (CNNs) automatically learn hierarchical feature representations from packet- or flow-level network data, making them effective for intrusion detection tasks that rely on identifying local spatial patterns (Fernando et al., 2025; Purwono et al., 2022; Ayeni, 2022). Their ability to perform end-to-end feature extraction has resulted in high detection accuracy and efficient real-time inference across several benchmark datasets (Shebl et al., 2024). However, CNNs do not explicitly model temporal dependencies within network traffic and may exhibit reduced performance for sequential attack behaviors unless combined with temporal or attention-based architectures. Furthermore, their susceptibility to adversarial perturbations and domain shifts motivates the integration of adversarial training and domain adaptation techniques (Gueriani et al., 2024; Bamber et al., 2025; Phalaagae et al., 2025; Farhanath et al., 2024; Gao, 2022).

Table 7: Qualitative comparison of DL architectures for ZDA detection.

Representative Studies	Architecture	Strengths	Weaknesses	Reported Use Context
(Liu and Wu, 2023; Halbouni et al., 2022; Kim et al., 2020; El Alami and Rawat, 2023)	CNN	Fast and efficient packet-level feature extraction suitable for real-time IDS deployment.	Limited temporal awareness and vulnerability to adversarial perturbations.	Packet-level anomaly detection.
(Khaleel and Shiltagh, 2023; Gueriani et al., 2024; Said et al., 2023; Hossain, 2025)	RNN/LSTM	Captures sequential traffic behavior and provides strong temporal modeling of evolving attacks.	Training instability and high computational cost.	Detection of evolving or time-dependent attacks.
(Narmadha and Balaji, 2025; Korba et al., 2023; Vinayak and Jarin, 2025; Meliboev et al., 2022)	Autoencoder	Supports unsupervised anomaly detection and is suitable for unlabeled or imbalanced data.	Limited interpretability and sensitivity to adversarial manipulation.	Detection of novel or previously unknown attacks.
(Xu et al., 2024; Akuthota and Bhargava, 2025; Altaf et al., 2024; Liu et al., 2025)	Graph Neural Network	Models relational structures among hosts and flows and may provide resilience to noisy observations.	High computational overhead and comparatively opaque decision processes.	IoT and distributed network environments.
(Ullah et al., 2024; Wu et al., 2022; Manocchio et al., 2024; Xi et al., 2024)	Transformer	Self-attention captures long-range dependencies and achieves high benchmark detection accuracy in the cited studies.	High memory consumption and computational requirements.	Large-scale enterprise and backbone network traffic.

Note: The comparison was synthesized from the representative studies cited in each row. The reported strengths, weaknesses, and application contexts reflect claims made by the respective authors rather than results from an independent benchmark conducted in this review.

2.3.2 RNN/LSTM-Based Temporal Detection Models

Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are designed to capture sequential dependencies within network traffic, enabling the detection of evolving attack patterns that approximate zero-day conditions (Kasongo, 2023; Yadav and Kumar, 2022; Amutha et al., 2022). LSTMs address the vanishing-gradient limitations of conventional RNNs through gated memory mechanisms, allowing them to preserve long-term contextual information and model complex temporal behaviors

(Awad et al., 2023; Zhou et al., 2023; Aydin et al., 2022). Despite their effectiveness, recurrent architectures remain computationally expensive and difficult to parallelize, limiting scalability for large-scale deployments. Consequently, hybrid architectures combining LSTMs with CNNs, GRUs, or autoencoders have been widely explored to improve efficiency and detection performance (Said et al., 2023; Du et al., 2023; Halbouni et al., 2022; Abdulmajeed and Husien, 2022; Babaey and Faragardi, 2025).

2.3.3 Autoencoder-Based Representation Learning Models

Autoencoders and their variants, including variational and denoising autoencoders, are widely adopted for unsupervised anomaly detection by learning compact representations of normal network traffic and identifying deviations through reconstruction errors (Rehman et al., 2025; Zahoor et al., 2022; DAS and PRAMOD, 2022; Ambekar and Thokchom, 2024). Their ability to exploit unlabeled data makes them particularly attractive for zero-day proxy scenarios where labeled attacks are scarce. Nevertheless, reconstruction-based methods often exhibit limited interpretability, reduced sensitivity to subtle attack variations, and vulnerability to adversarial or out-of-distribution inputs, motivating the development of hybrid and robust representation-learning approaches (Babaey and Faragardi, 2025).

2.3.4 GNN-Based Traffic and Topology-Aware Detection

Graph neural networks (GNNs) extend deep learning to graph-structured cybersecurity data by modeling both node attributes and their relational dependencies (Gupta et al., 2025; Sun et al., 2024; Altaf et al., 2024; Zhou et al., 2020; Wu et al., 2021). This capability makes GNNs particularly suitable for intrusion detection in IoT, SDN, and distributed environments where communication patterns naturally form graphs. Although GNNs effectively capture topological relationships that conventional architectures may overlook, their computational complexity, scalability challenges, and limited interpretability remain significant barriers to practical deployment (Tung et al., 2025).

2.3.5 Transformer-Based Intrusion Detection Architectures

Transformer architectures employ self-attention mechanisms to model long-range dependencies in network traffic and have reported high benchmark performance on several intrusion detection datasets in the cited studies (Xi et al., 2024; Akuthota and Bhargava, 2025; Manocchio et al., 2024; Long et al., 2024), though without a common cross-architecture benchmark comparison this review cannot independently confirm a state-of-the-art ranking. Their ability to capture global contextual information enables effective detection of complex and evolving attack behaviors that may evade conventional CNN- or RNN-based models (Liu and Wu, 2023; Han et al., 2023; Wu et al., 2022; Ullah et al., 2024; Tseng et al., 2024). However, transformers require substantial computational resources, large training datasets, and considerable memory, which limits their applicability in real-time and resource-constrained environments (Zhou et al., 2025).

No single DL architecture consistently provides optimal performance for zero-day attack detection. CNNs efficiently extract local features but lack temporal modeling capabilities, RNN/LSTM models capture sequential dependencies at the expense of scalability, autoencoders enable unsupervised anomaly detection but may struggle with complex attack patterns, GNNs exploit relational structures while incurring high computational costs, and transformers achieve superior contextual modeling but require significant computational resources. Consequently, recent research increasingly favors hybrid, ensemble, and self-supervised approaches that combine the complementary strengths of multiple architectures to improve robustness, generalization, and interpretability under zero-day proxy conditions (Yin et al., 2017; Kim et al., 2020; Vinayakumar et al., 2019; Khan, 2021; Sameera and Shashi, 2020).

2.4 Recent Work on ZDA Detection

Table 7 presents an abstract architectural comparison of DL, and the discussion below (and the more extensive Table 10 in Section 3) summarizes various DL-based approaches to ZDA detection explored in recent studies, considering the architectural choices, data, and learning paradigms applied. Collectively, these studies demonstrate an ongoing shift away from classical supervised models toward adaptive, transferable, and self-supervised paradigms that can be scaled to handle novel or low-frequency attack patterns (Zhao et al., 2019).

Early results from (Khan, 2021) indicated that HCRNNIDS can outperform standard machine learning approaches on large datasets such as CSE-CIC-IDS2018; however, it achieves high accuracy but exhibits limited generalization to new attack modes. Similarly, (Halbouni et al., 2022) reported cross-dataset performance degradation on the CIC-IDS2017, UNSW-NB15, and WSN-DS datasets, while demonstrating the persistent challenge of distributional shift and generalization gaps in purely supervised detection systems.

Follow-up work has involved federated and temporal learning. (Lopes et al., 2023) employed temporal convolutional networks (TCNs) with dilated convolutions to learn long-range relationships in network traffic. In contrast, (Korba et al., 2023) proposed a federated autoencoder architecture that counteracts privacy and labeling constraints by enabling decentralized, communication-efficient learning. These approaches reflect a shift towards architectures to attain scalability, confidentiality, and power.

Other studies have used transfer learning (TL) and generative modeling to improve flexibility. (Ullah et al., 2024) introduced IDS-INT, a transformer-based TL approach that improves minority-class recall across multiple intrusion datasets, demonstrating the utility of pre-trained representations. In parallel, Xu et al. (2023) introduced FSC-GAN, a few-shot conditional GAN that reduces false positives and demonstrates good performance in Internet-of-Vehicles (IoV) environments, demonstrating the utility of generative synthesis for a few-label case.

More recent work has emphasized self-supervised and meta-learning paradigms. Faber et al. (2021) demonstrated that self-supervised Autoencoder frameworks can achieve robust detection across unseen datasets. Ahmad et al. (2025) introduced a stochastic prototypical network for few-shot IoV intrusion detection, demonstrating resilience to labeling scarcity and environmental noise.

In conclusion, the research discussed above and summarized more fully in Table 10 (Section 3) reflects a methodological shift towards increasingly generalizable, data-efficient, and privacy-preserving systems from static supervised learning. The intersection of federated, self-supervised, and transfer learning paradigms appears pivotal for future ZDA detection, in which resilience against emerging threats and adaptability are paramount.

2.5 Emerging Trends and Technologies

The conceptual and chronological development of ZDA detection research traces a shift from conventional rule-based detection systems to distributed, intelligent, and adaptive systems. Most early ZDA detection solutions relied on statistical anomaly- and signature-based detection techniques. These methods laid the groundwork for processes that identify risks by detecting intrusions using predefined attack patterns or behavioral anomalies. Limitations of signature-driven and conventional feature-engineered detection under changing traffic distributions motivated increasing use of representation-learning and adaptive DL approaches that incorporate automated feature extraction and adaptation to different data distributions. Figure 4 complements this qualitative narrative with an empirical view specific to this review’s own audited corpus: the publication-year distribution of its 37 primary-corpus studies, which concentrates heavily in the most recent two years (26 of 37, 70%, published in 2025–2026) rather than spreading evenly across the historical periods just described – a reminder that this corpus’s search window (2018–2026, Table 9) captures only the most recent phase of a much longer conceptual history.

The use of convolutional, recurrent, and transformer-based models has dominated the new phase of proxy research for zero-day conditions following the introduction of DL (Tariq et al., 2024). Compared with traditional methods, these learning algorithms offer more precise and trustworthy temporal and hierarchical representations of complex network traffic. Recent developments represent further steps towards next-generation detection paradigms that emphasize robustness, explainability, and generalization in adversarial and adaptive settings, consistent with the recency of this corpus’s own evidence base shown in Figure 4.

Current research focuses on adversarial robustness, explainability, FL, few-shot adaptation, and edge deployment. Adversarial robustness protects DL-based detectors against poisoning and evasion attacks by leveraging robust training techniques and uncertainty-aware inference. To create model transparency that enables justifiable judgments, attention visualization, feature attribution, interpretable reasoning, explainability, and interpretability are becoming increasingly important. Federated learning allows models to train together across multiple devices or enterprises without sharing raw data by implementing decentralized and privacy-preserving training methods. Few-shot and meta-learning strategies simultaneously address labeled-sample sparsity, thereby enabling rapid adaptation to unseen or ZDA. Finally, in latency-sensitive environments, the availability of edge and fog computing enables lightweight, power-conserving detection near data sources to enable real-time analysis and response.

In conclusion, these developments demonstrate a research path in which autonomous, interpretable, and robust systems replace static rule-based ZDA detection. The evolution of the field from signature- and anomaly-based detection to intelligent, distributed, and adversarially robust systems that can adapt to the ever-changing threat posture of contemporary cyberspace is the subject of the qualitative narrative above; Figure 4 shows, empirically, just how recently this review’s own evidence base has been produced.

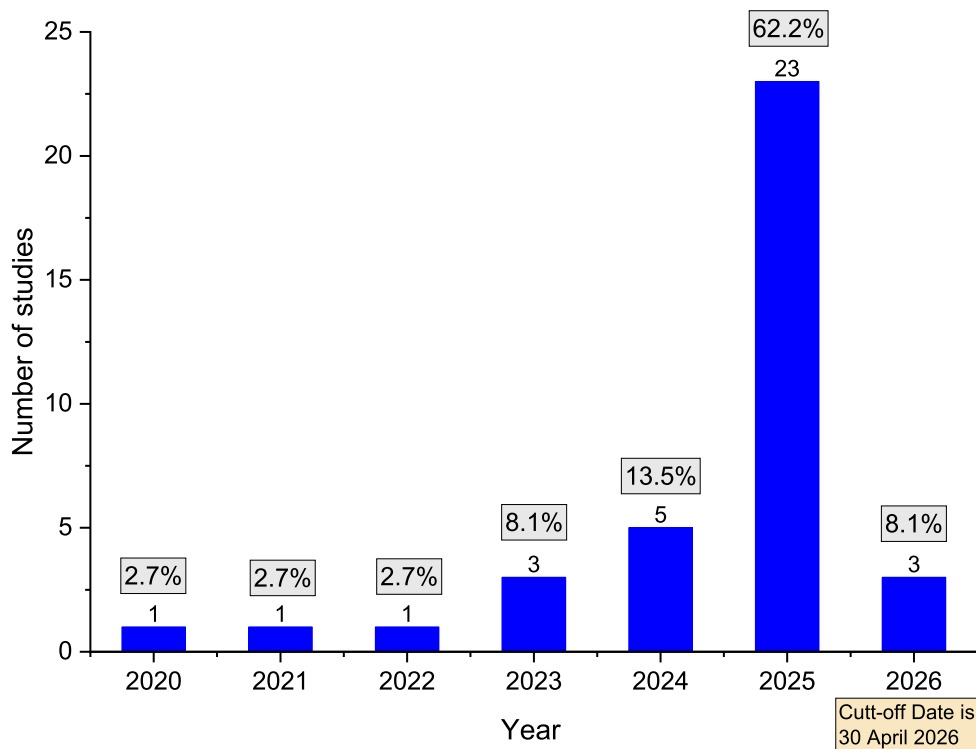


Fig. 4: Publication-year distribution of the 37 studies in the corrected audited primary synthesis corpus (Section 3.4.1) – an empirical bibliometric count of this review’s own evidence base, not a qualitative periodization of the field’s conceptual history (which is instead discussed in prose above).

3 Systematic Review of DL-Based ZDA Detection

To ensure methodological transparency and reproducibility, this systematic review is structured around the research questions (RQs) listed in Table 8, which guide the study selection, data extraction, and synthesis. These RQs provide an empirical foundation for the descriptive quantitative synthesis presented in Section 4.

Table 8: Research questions guiding this systematic review

RQ	Description
RQ1	Which DL architectures are most commonly employed for intrusion detection under zero-day proxy conditions?
RQ2	What datasets, traffic representations, and experimental protocols dominate zero-day proxy evaluation?
RQ3	How are zero-day conditions operationalized and evaluated across the literature?
RQ4	What evidence exists regarding the robustness of DL-based IDSs against adversarial manipulation and distributional shifts?
RQ5	To what extent are explainability techniques integrated into ZDA proxy detection systems, and how do they support operational trust and deployment?

3.1 Search Strategy and Study Identification

The corpus-forming search was conducted across five sources: IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect, SpringerLink, and Wiley Online Library. These sources were selected for their coverage of peer-reviewed research in cybersecurity and machine learning, and the final corpus-forming search was conducted on 30 April 2026. In July 2026, the same core query was also executed in Scopus and the Web of Science Core Collection as a supplementary coverage-validation step. The validation searches returned 105 and 146 raw records, respectively.

Database-specific Boolean strings combined zero-day terminology, deep learning, and IDS/NIDS concepts. Minor syntax adaptations accommodated the indexing rules of each platform while preserving semantic equivalence. The search used the umbrella term “deep learning” rather than individual architecture names to maintain a reproducible cross-database protocol. Because terminology in this field is heterogeneous, backward and forward citation chaining and manual reference screening were used to reduce the risk of missing studies framed as unseen-, unknown-, novel-, open-set-, temporal-, cross-dataset-, or zero-shot detection. The conduct and reporting of the review followed PRISMA 2020 and PRISMA-S guidance (Page et al., 2021; Rethlefsen et al., 2021).

The five corpus-forming database searches yielded 1,083 records; citation chaining added 18 records, producing the completed 1,101-record PRISMA identification set. The Scopus and Web of Science results are reported in Table 9 as supplementary validation counts. Only their aggregate counts were retained, so they were not combined with the original record-level exports and do not change the completed deduplication, screening, or corpus counts.

A literature-currency check conducted in July 2026 identified four recent systematic reviews that were relevant to positioning but did not meet the empirical primary-study criterion: explainable DL-based IDS for IoT (Ogunseyi et al., 2026), DL intrusion detection in emerging technologies (Neto et al., 2025), a broad IDS systematic review (Arnob et al., 2025), and DL-driven network IDS methods (Chinnasamy et al., 2025). These surveys were added to Section 1.1 and Table 1; they were not counted as primary empirical studies.

Table 9: Database-specific search syntax, limits, and raw retrieval counts for 2018–2026.

Database	Database-Specific Syntax and Limits	Records
IEEE Xplore	Common core query applied as a metadata search. Limited to journal and conference papers published during 2018–2026.	68
ACM Digital Library	Common core query applied as a metadata search. Limited to journal and conference papers published during 2018–2026.	87
SpringerLink	Common core query applied to research and review articles in computer science and engineering, published during 2018–2026.	193
Elsevier ScienceDirect	Common core query applied as a metadata search across journal and conference content published during 2018–2026.	676
Wiley Online Library	Common core query applied to publications issued during 2018–2026. The terms <i>IDS</i> and <i>NIDS</i> were omitted where required by the platform syntax.	59
Scopus [†]	TITLE-ABS-KEY field search using the common core query, limited to publications issued during 2018–2026.	105
Web of Science Core Collection [†]	TS= topic search using the common core query, limited to publications issued during 2018–2026.	146
Combined raw retrievals across corpus-forming and validation searches (not a PRISMA denominator)		1,334

Common core query: (“zero-day attack” OR “zero-day intrusion” OR “zero-day detection” OR “zero-day exploit”) AND (“deep learning”) AND (“intrusion detection system” OR “network intrusion detection” OR IDS OR NIDS).

Note: The common core search strategy was applied across all seven databases, with minor syntax adjustments to accommodate database-specific indexing rules while preserving semantic equivalence. The Scopus and Web of Science searches marked with [†] were supplementary coverage-validation searches conducted after completion of the original five-database screening workflow. Their aggregate retrieval counts are included in the raw total; however, record-level exports were not available for combined deduplication and screening. The completed PRISMA workflow therefore remains based on 1,083 records retrieved from the original five databases and 18 records identified through backward and forward citation chaining, giving 1,101 records in the screened identification set.

3.2 Screening and Eligibility

The completed five-source corpus-forming search identified 1,083 records, and backward and forward citation chaining identified 18 additional records, giving the 1,101-record PRISMA identification set. The 251 raw Scopus and Web of Science validation results were not added to this flow because only aggregate counts, rather than record-level exports, were available for the combined deduplication and screening workflow. After removal of 310 duplicate records and exclusion of eight non-English records, 783 records were screened by title and abstract. At this stage, 313 records

were excluded and 470 reports were sought for retrieval. Of these, 120 reports could not be accessed because full text was unavailable or access was restricted, leaving 350 full-text articles for eligibility assessment. Full-text assessment excluded 250 articles that did not evaluate proxy zero-day detection, did not use a DL-based IDS, addressed an unrelated security domain, or lacked empirical validation. One hundred studies met the topical screening criteria.

A corrected publication-status audit retained 59 records in the DOI-verified peer-reviewed venue set and 41 records as venue-stage background evidence (32 journal/proceedings records and nine preprint/repository records; Supplementary Table S4). Publisher, issuing-organization, or repository identification was subsequently completed for bibliographic transparency; this identification alone did not establish peer-review status, domain fit, a qualifying proxy protocol, or empirical eligibility. Reapplication of the prespecified empirical eligibility criteria excluded 22 of the 59 venue-verified records, each documented in Supplementary Table S5. The corrected primary synthesis corpus therefore comprises 37 studies, while 63 records are retained as background or excluded evidence.

3.3 Inclusion and Exclusion Criteria

Studies were included if they proposed or evaluated a DL-based IDS and explicitly assessed its performance under zero-day conditions using recognized proxy evaluation strategies, such as unseen attack classes, temporal holdout splits, open-set evaluation, or cross-dataset testing. Eligible studies were required to employ real, synthetic, or benchmark datasets and report empirical performance metrics derived from experimental evaluations.

Studies were excluded if they were purely theoretical or conceptual in nature, focused exclusively on signature-based intrusion detection, lacked empirical validation, or did not report quantitative performance results. Non-peer-reviewed editorials, opinion pieces, and abstracts without sufficient experimental details were excluded.

3.4 Data Extraction and Synthesis

Data were systematically extracted from each of the 100 studies identified through screening using a standardized extraction form. Online Resource 1 contains Supplementary Tables S1–S5, reporting the corrected study-level record for every screened study, including identifiers, citation keys, venue status, year, datasets, architecture, proxy setting, extractable metrics, appraisal judgments, and exclusion reasons.

Because topical relevance does not establish peer-reviewed publication status, every screened record was checked against its publisher or DOI record and final publication version. The corrected audit identifies 9 preprint/repository records and 32 journal/proceedings records retained outside the verified primary-evidence set under the review’s conservative verification rule, producing 41 venue-stage background records in Supplementary Table S4. Fifty-nine records were verified at established peer-reviewed venues and formed the set assessed against the empirical eligibility criteria.

Because topical relevance does not establish peer-reviewed publication status, every screened record was checked against its publisher or DOI record and final publication

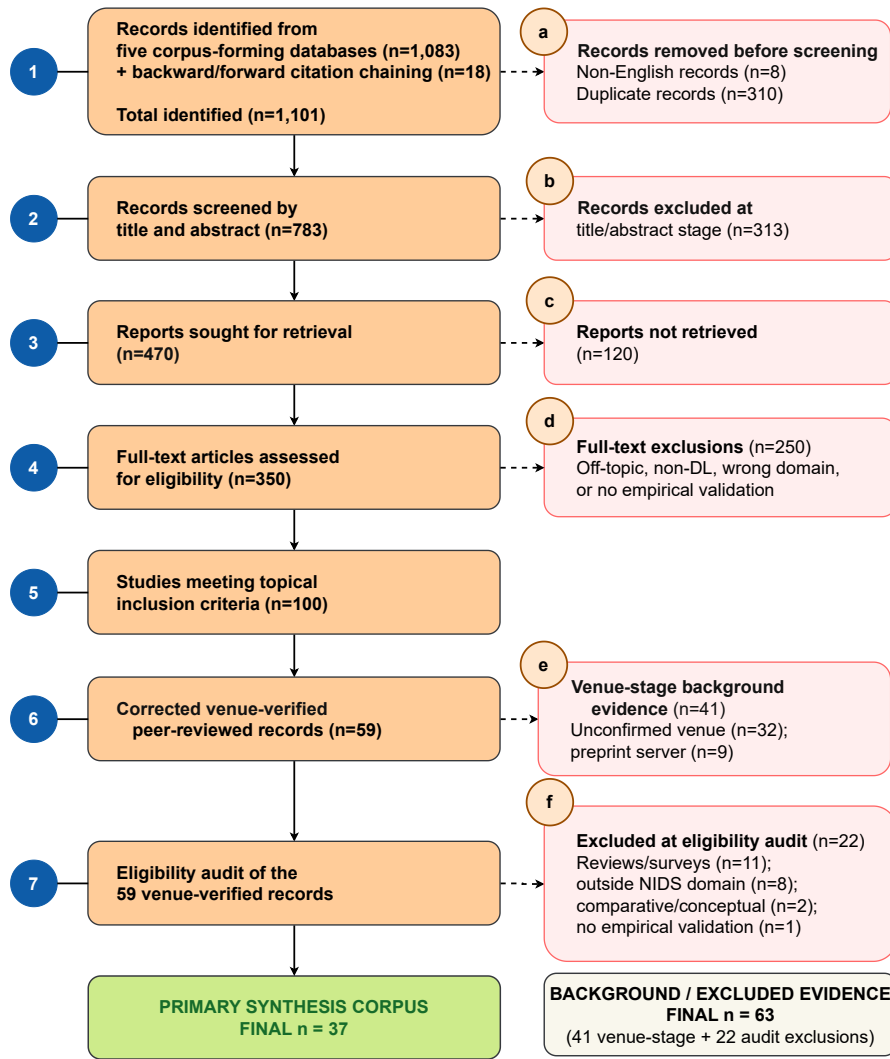


Fig. 5: PRISMA 2020 flow diagram for the completed five-source corpus-forming search, with corrected publication-status and eligibility-audit stages. The supplementary Scopus/Web of Science coverage-validation search (251 raw records) is reported separately in Table 9 and Section 3.1; it is not shown in this flow because record-level exports were not available for combined deduplication and screening. Full-text exclusions are reported as a verified aggregate because a record-level historical exclusion log was not retained; all 22 audit-stage exclusions are individually documented in Supplementary Table S5.

version. The corrected audit identifies 9 preprint/repository records and 32 journal/proceedings records retained outside the verified primary-evidence set under the review’s conservative verification rule, producing 41 venue-stage background records in

Supplementary Table S4. Fifty-nine records were verified at established peer-reviewed venues and formed the set assessed against the empirical eligibility criteria. This verification moved (Soltani et al., 2023) and (Hindy et al., 2020) from venue-stage background evidence into the eligible publication set.

3.4.1 Corrected Corpus Eligibility Audit

Publication-status verification establishes where a record was published but does not by itself establish eligibility as an empirical primary study. The 59 venue-verified records were therefore re-examined against Section 3.3. Twenty-two records were excluded and are listed individually in Supplementary Table S5:

- Eleven literature reviews or surveys were retained only as background or comparative literature (Deldar and Abadi, 2023; Zhang and Maple, 2023; Papalkar and Alvi, 2023; Lal and Ashok, 2026; Siam et al., 2026; Nair and Mhavan, 2023; Tayyab et al., 2022; Kritika, 2024; B et al., 2025; Mote and Arabagatte, 2025; Manjula and Saravanakumar, 2025).
- Eight studies were outside the network-traffic IDS domain (Won et al., 2023; Li and Zhao, 2024; Houda et al., 2024; Mahida, 2026; Zahoor et al., 2022; Okey et al., 2025; Wahid et al., 2025; Elijah et al., 2025). Three of these were identified during a final full-text verification pass: Okey et al. (2025) addresses zero-day *device* identification and fingerprinting rather than unknown-attack detection; Wahid et al. (2025) targets endpoint threat detection rather than network-traffic intrusion detection; and Elijah et al. (2025) detects browser exploits from behavioural and system-log signals rather than network flows.
- Two records were comparative or conceptual and did not establish a qualifying primary empirical DL evaluation (Ali et al., 2022; Anwer et al., 2022).
- One proposed framework did not report demonstrated empirical validation in the available paper record (Naithani et al., 2024).

This correction moves Soltani et al. (2023) and Hindy et al. (2020) from the venue-background table into the primary corpus and moves Zahoor et al. (2022), Anwer et al. (2022), Naithani et al. (2024), Okey et al. (2025), Wahid et al. (2025), and Elijah et al. (2025) into the eligibility-exclusion table. The final corrected corpus contains 37 primary studies and 63 background or excluded records.

The audit was conducted by the review team through re-examination of titles, abstracts, publication records, and extracted method and dataset descriptions, followed by a final full-text verification pass over the records whose domain fit or proxy protocol had been recorded as uncertain during extraction. That pass confirmed an explicit, auditable proxy protocol for some records: for example, Kara et al. (2025) withholds PortScan and DDoS classes from training and retains them for testing, Rodríguez et al. (2022) trains on four UNSW-NB15 attack families and tests on five families absent from training, and Wang et al. (2025) evaluates temporally separated unseen attack patterns. For a further group of retained studies, the specific proxy construction could not be independently confirmed from accessible full text, most often because the record was paywalled or reported only an aggregate zero-day claim without stating which attacks were withheld. Those studies are retained because they

satisfy the domain, deep-learning, and empirical-evaluation criteria and report a zero-day or unknown-attack evaluation, but they are coded conservatively as providing no explicit proxy evidence in the proxy-strategy synthesis (Figures 7 and 9); this is disclosed as a limitation rather than treated as evidence that those protocols are absent. Screening and the eligibility audit were conducted by the review team rather than by two independently blinded reviewers, so a formal inter-rater agreement statistic is not reported. This is stated explicitly as a limitation of the review process.

The extracted information included publication metadata, dataset characteristics, application domains, deep learning architectures and learning paradigms, zero-day proxy evaluation strategies, and reported performance metrics.

Representative empirical studies are summarized in Table 10 to illustrate the dominant architectural trends, dataset usage patterns, and proxy-based zero-day evaluation strategies. These studies were selected based on methodological diversity and relevance to the defined research questions rather than publication venue or year.

The following information was extracted from each study:

1. Publication year and venue;
2. Dataset(s) used and application domain (enterprise, IoT, SDN, ICS, vehicular);
3. DL architecture and learning paradigm;
4. Zero-day assessment strategy (e.g., unseen attack classes, temporal split, cross-dataset testing); and
5. Reported performance metrics.

3.4.2 Representative Empirical Studies

Table 10 summarizes selected illustrative studies from the broader DL-IDS and proxy-ZDA literature used for contextual comparison. It is not restricted to the audited 37-study primary synthesis corpus and is not used to compute corpus-dependent counts or quantitative summaries in Tables 11–13. This structured narrative summary does not rank methods; instead, it consolidates what the literature most consistently reports and surfaces reporting gaps that limit comparability across studies.

Table 10: Representative studies from the broader DL-based IDS and proxy-ZDA literature, included for contextual comparison rather than corpus-dependent synthesis. The table is not a performance ranking; corpus-level results are derived only from the audited 37-study set documented in Supplementary Tables S1–S3.

Study	Dataset(s)	Model	Reported Headline Metric	Key Findings
(Mala et al., 2025)	CICIDS2018, Edge-IIoTset	DCNN + BiLSTM IDS for IoT detection	Not ZDA reported	Joint spatial and temporal feature learning improves ZDA detection in heterogeneous IoT environments.

Continued on next page

Table 10 – continued from previous page

Study	Dataset(s)	Model	Reported Headline Metric	Key Findings
(Akuthota and Bhargava, 2025)	NSL-KDD, UNSW-NB15	Transformer-based IDS for IoT	$\approx 98\text{--}99\%$	Self-attention captures long-range dependencies and outperforms recurrent IDS models.
(Bahlali et al., 2025)	UNSW-NB15, CSE-CIC-IDS2018	SSL + custom autoencoder IDS	$>99\%$	Self-supervised learning improves encrypted-traffic anomaly detection while reducing dependence on labeled data.
(Gupta et al., 2025)	NSL-KDD	GNN-based IDS	Not reported	Graph representations model relational dependencies among network flows.
(Bamber et al., 2025)	NSL-KDD	Hybrid CNN-LSTM IDS	$\approx 99\%$	Hybrid spatio-temporal learning outperformed individual DL architectures.
(Jiang et al., 2025)	CICIDS2017	Memory Flow Transformer (MFT)	$>99\%$	A memory-enhanced transformer improves detection accuracy and processing efficiency for high-dimensional traffic.
(Ahmad et al., 2025)	CAN-bus, IoV	Few-shot DL IDS	Not reported	Few-shot learning enables detection of previously unseen vehicular attacks using limited labeled samples.
(Bensaoud and Kalita, 2025)	NSL-KDD, UNSW-NB15, CICIoT2023	Hybrid DL IDS for IoT attacks	$\approx 98\text{--}99\%$	Hybrid models improve cross-dataset generalization and robustness against ZDA in IoT networks.
(Sinha et al., 2025)	BoT-IoT	Hybrid LSTM + CNN IoT IDS	$>99\%$	CNN feature extraction combined with LSTM temporal modeling produces high detection accuracy for IoT traffic.

Continued on next page

Table 10 – continued from previous page

Study	Dataset(s)	Model	Reported Headline Metric	Key Findings
(Zhou et al., 2025)	ToN-IoT, Edge-IIoTset	HiViT-IDS, vision-transformer IDS	a Not reported	Vision-transformer models applied to traffic-feature representations demonstrate scalability for large-scale IoT intrusion detection.
(Manocchio et al., 2024)	NSL-KDD, UNSW-NB15, CSE-CIC-IDS2018	FlowTransformer: flow-level transformer NIDS	$\approx 98-99\%$	Flow-level tokenization with self-attention captures long-term dependencies across heterogeneous datasets.
(Golchin et al., 2024)	CICIDS2017, UNSW-NB15, CTU-13, CICDoS, Botnet	SSCL-IDS: self-supervised contrastive IDS	Not reported	Contrastive self-supervised learning improves generalization to unseen attacks while reducing reliance on labeled data.
(Xi et al., 2024)	NSL-KDD, CIC-DDoS2019, UNSW-NB15	Multi-scale transformer IDS	$\approx 98-99\%$	Multi-scale attention improves detection of attacks exhibiting different temporal characteristics.
(Xu et al., 2024)	NF-Bot-IoT, NF-CSE-CIC-IDS2018 (v1/v2)	Self-supervised GNN-based NIDS	Not reported	Self-supervised GNN learning captures relational dependencies among network entities without requiring labeled data.
(Lopes et al., 2023)	CICDDoS2019, CSE-CIC-IDS2018	Temporal CNN-based IDS	$\approx 99\%$	Temporal convolution models capture sequential traffic patterns for high-accuracy intrusion detection.
(Wu et al., 2022)	CICIDS2017, CIC-DDoS2019	RTIDS: robust transformer-based IDS	$>99\%$	Robustness-oriented transformer training improves resistance to adversarial noise and traffic-distribution shifts.
(Faber et al., 2021)	CICIDS2017	Autoencoder-based IDS	$\approx 98-99\%$	Reconstruction-error autoencoders support anomaly detection without explicit attack labels.

Continued on next page

Table 10 – continued from previous page

Study	Dataset(s)	Model	Reported Headline Metric	Key Findings
(Hindy et al., 2020)	CICIDS2017, NSL-KDD	DL-based IDS with zero-day evaluation	$\approx 99\%$	Provides an early systematic evaluation of DL models under zero-day scenarios and emphasizes generalization capability.
(Vinayakumar et al., 2019)	NSL-KDD, UNSW- NB15, Kyoto, WSN-DS, CICIDS2017	Deep DNN-based multi-dataset IDS	$\approx 97\text{--}99\%$	Deep feedforward networks generalize across diverse datasets when supported by appropriate regularization.
(Shang et al., 2018)	NSL-KDD	Deep autoencoder- based anomaly detection	$\approx 98\%$	Deep autoencoders learn normal traffic patterns and identify anomalies using reconstruction errors.

3.5 Reporting-Completeness Screen

Rather than a validated risk-of-bias instrument, methodological and reporting completeness were assessed using an IDS-adapted, PROBAST-*inspired* checklist that this review’s authors constructed for this purpose; it has not itself been validated against an established risk-of-bias framework and should be read as a structured reporting-completeness screen rather than a certified bias assessment. A validated tool such as ROBIS or AMSTAR 2 was not used because those instruments assess the conduct of a review rather than the predictive-model choices of each primary study, which is what this checklist targets. The checklist re-purposes four PROBAST-style domain labels as follows: participants (datasets and traffic sources named), predictors (whether the architecture could be classified into a specific family), outcomes (whether an absolute accuracy or F1 value was extractable), and analysis (DOI-based venue verification, which is Low for every record in this corpus by construction). A per-study scoring rubric is provided in Supplementary Table S3, alongside the DOI-based peer-review verification described in Section 3.4. Because this checklist encodes reporting completeness rather than genuine methodological safeguards against bias (e.g., it does not verify absence of test-set leakage, class-balance handling, or independence between training and proxy-evaluation splits), its domain labels should not be read as equivalent to a PROBAST, ROBIS, or Cochrane risk-of-bias verdict.

Following the corrected eligibility audit (Section 3.4.1), the primary synthesis corpus comprises 37 studies, and Supplementary Table S3 reports the per-study screen for exactly these 37 records. Across the four domains, the distribution is: participants, 33 low / 4 high; predictors, 24 low / 13 some concerns; outcomes, 15 low / 22 some concerns; analysis, 37 low. Combining domains, the overall screen is low for 10 studies, some concerns for 23 studies, and high for 4 studies, with an unreported dataset the sole driver of every high overall rating.

The conduct and reporting of this systematic review followed PRISMA 2020 and PRISMA-S guidance (Page et al., 2021; Rethlefsen et al., 2021), noting that Section 4 reports a descriptive quantitative synthesis rather than a formal meta-analysis (Section 4.1).

3.6 Sensitivity to Publication-Status Verification

Publication-status verification changed the provisional evidence classification of two records: the final journal versions of Soltani et al. (2023) and Hindy et al. (2020) were confirmed as peer-reviewed publications and included in the corrected primary corpus. Six previously included or provisionally included records were excluded on domain or empirical-eligibility grounds. Because these corrections alter architecture and dataset counts, all descriptive tables and figures in Sections 4–5 and Supplementary Tables S1–S5 were recalculated from the corrected 37-study corpus. No pooled effect estimate is reported, so the correction does not imply a recalculated meta-analytic effect.

4 Quantitative Synthesis of DL-Based ZDA Detection Systems

4.1 Overview and Rationale

This section builds on the systematic review presented in Section 3 and provides a quantitative synthesis of empirical evidence on the effectiveness, robustness, and explainability of DL-based IDSs under zero-day proxy conditions. The synthesis draws on the corrected primary corpus of 37 venue-verified, eligibility-audited empirical studies (Section 3; Supplementary Table S1) for descriptive analysis, including architecture prevalence, dataset usage, and explainability trends.

This section reports a *descriptive quantitative synthesis* rather than a pooled meta-analysis, because the comparator means, standard deviations, and sample sizes required to compute and audit a pooled effect size (e.g., Hedges’ g with I^2 heterogeneity and Egger’s regression) are not reliably extractable from the reviewed studies, most of which report only point accuracy/F1 values without variance or a matched non-DL baseline evaluated under identical proxy conditions. Rather than present pooled effect sizes that a reviewer could not recompute, this section reports: point performance values (accuracy, F1-score) as reported by the primary studies themselves, aggregated by architecture family and by dataset, with sample sizes (n) stated explicitly at every level of aggregation. This is a narrower claim than a meta-analysis, but every number in Tables 11 and 12 below traces to a specific study in Supplementary Table S2.

Of the 37 studies in the corrected primary synthesis corpus, 15 reported a numeric accuracy and/or F1-score in a form suitable for direct extraction (i.e., not solely a relative improvement over an unspecified baseline, and specific to the study’s own deep-learning system rather than a classical-ML comparator it reported alongside). The remaining 22 studies either reported metrics only as relative gains, reported non-comparable metrics (e.g., latency, AUC without accuracy), reported a metric whose definition could not be confirmed as accuracy (Section 4.2), or did not report a quantitative headline metric. These 22 studies still contribute to the descriptive and thematic

synthesis, including the corpus-adoption counts in Tables 11–12, but they do not contribute to the reported accuracy/F1 summaries in those tables. Per-architecture and per-dataset sample sizes in this synthesis are therefore small (often $n = 1\text{--}6$) and are reported as medians and ranges rather than means or inferential estimates; no claim of statistical significance is made anywhere in this section.

In the descriptive synthesis of reported values, the main source of variation is not only architecture family but also dataset choice, proxy definition, class imbalance, feature representation, and reporting completeness. Because the reviewed studies rarely report variance estimates, matched baselines, or identical zero-day proxy protocols, the observed ranges should be interpreted as descriptive evidence of heterogeneity rather than as pooled comparative effects.

4.2 Extraction and Aggregation Protocol

Performance values were extracted directly from the reported abstract, results, and findings text of each of the 37 primary-corpus studies (Supplementary Tables S1–S2). To avoid conflating relative improvement claims (e.g., “a 20% improvement in detection accuracy”) with absolute performance values, an extraction rule excluded any candidate value whose surrounding text contained comparative language (e.g., “improve”, “increase”, “reduction”, “up to”, “compared to”); only explicit absolute values (e.g., “achieves 98.7% accuracy”) were retained. A final consistency check compared each retained value against its own source sentence in Supplementary Table S2; this check corrected two entries where the original coding had selected the wrong quantity from a multi-metric sentence (one case of precision entered as accuracy; one case of a classical-machine-learning comparator’s accuracy entered in place of the study’s own deep-learning result, which also changed that study’s architecture-family tag from Ensemble/classical-ML hybrid to Reinforcement) and removed two entries lacking any value meeting this rule (Supplementary Table S2, footnote). Where a study reported results on more than one dataset or attack configuration, the value associated with the study’s primary proxy zero-day evaluation was retained. Each study was assigned one primary architecture family for corpus-adoption counts; composite architectures such as CNN–LSTM were coded once as “Hybrid”, so the architecture counts sum to the 37-study corpus.

4.3 Architecture-Level Descriptive Synthesis

Table 11 reports, for each architecture family, the number of studies in the audited 37-study corpus that adopt it, the number reporting an extractable accuracy or F1 value, and the range of those values. Architecture family was coded directly from each study’s full source text for the 15 studies in the quantitative subset (Supplementary Table S2), and from each study’s title using a documented keyword rule for the remaining 22 studies (Supplementary Table S1, coding note); the two coding methods agree on every study checked under both (Supplementary Table S1, footnote).

The residual “Unspecified/other DL” category is the largest single family (13 of 37), reflecting titles and short-form descriptions that name a broad technique (e.g., “deep neural network”, “AI-driven”) without a more specific architectural commitment; two

Table 11: Architecture-level adoption and reported performance under zero-day proxy conditions in the audited 37-study corpus.

Architecture Family	Corpus <i>n</i>	Metric <i>n</i>	Accuracy (%)	F1 (%)
Unspecified/other DL	13	3	96.15 [96.1–96.2]	64.0 (<i>n</i> = 1)
Hybrid, multi-architecture	7	5	97.5 [97.0–99.1]	98.2 (<i>n</i> = 1)
Federated	5	0	–	–
Reinforcement learning	3	1	94.0 (<i>n</i> = 1)	–
CNN	2	0	–	–
Autoencoder	3	3	98.0 [92.96–99.3]	–
Few-shot/meta-learning	2	1	84.0 (<i>n</i> = 1)	–
Ensemble/classical-ML hybrid	1	1	94.0 (<i>n</i> = 1)	94.0 (<i>n</i> = 1)
RNN/LSTM	1	1	98.7 (<i>n</i> = 1)	97.1 (<i>n</i> = 1)
Corpus total	37	15		

Note: Values are descriptive medians with minimum–maximum ranges; a single value denotes *n* = 1.

“Corpus *n*” records the adoption of each primary architecture family among the 37 studies in the corrected final audited corpus reported in Supplementary Table S1. Each study was assigned exactly one primary architecture family, and the column therefore sums to 37.

“Metric *n*” and the reported accuracy and F1 values are restricted to the 15 studies from which an absolute, architecture-specific, non-comparative, and non-classical-baseline performance value could be extracted, following the protocol in Section 4.2. “Metric *n*” counts studies with at least one extractable accuracy or F1 value; the metric-specific number contributing to the Accuracy and F1 columns may therefore be smaller than Metric *n*. These results are descriptive summaries over small samples and must not be interpreted as pooled effect-size estimates or statistically significant comparisons.

Transformer and GNN families occur in the wider proxy zero-day literature discussed in Section 2, but no study meeting the final eligibility criteria in Section 3.4.1 was coded to either family. These rows are therefore omitted rather than reported as zero. Full per-study values, source sentences, and study identifiers are provided in Supplementary Table S2.

of its thirteen members use a Bayesian or probabilistic-graphical-model formulation (Ranpara et al., 2025; Sunkara et al., 2025) whose classification as “deep learning” under this review’s own taxonomy (Section 2.2) is acknowledged as a boundary case rather than a clear-cut fit, and one is an LLM-agent-based approach (Saheed and Chukwuere, 2025). Architectures combining more than one representation-learning mechanism (coded “Hybrid”) form the next-largest family (7 of 37) and account for the largest share of studies reporting a usable metric (5 of 15) and the highest reported accuracies in this small sample, consistent with the qualitative pattern noted throughout Section 2 that hybrid spatio-temporal or reconstruction–classification designs are frequently proposed specifically to improve zero-day proxy generalization. However, with per-family samples this small, this pattern should be read as descriptive rather than confirmatory; a single study with an unusually high or low reported value can move a family’s median substantially, and no adjustment for study quality, dataset difficulty, or proxy type has been applied. The Hybrid-family value for (Nagarjun and Rajkumar, 2025) (99.1%) comes from a multi-task blockchain/quantum-computing study whose zero-day-proxy-specific evaluation could not be fully isolated from its

Table 12: Dataset usage and reported accuracy in the audited 37-study primary corpus.

Dataset	Corpus n (of 37)	n With Extractable Accuracy	Reported Accuracy (%)
CICIDS2017	12	3	98.0 [94.0–99.3]
UNSW-NB15	5	2	97.1 [96.2–98.0]
NSL-KDD	5	2	94.58 [92.96–96.2]
CIC-IoT2023	3	0	–
CSE-CIC-IDS2018	2	1	97.5 ($n = 1$)
N-BaIoT	1	0	–
KDD99	1	1	99.1 ($n = 1$) [†]
Edge-IIoTset	1	0	–
BoT-IoT	1	0	–
ToN-IoT	1	0	–

Dataset coverage: Custom, proprietary, or other non-standard datasets were used in 17 of the 37 studies. Four of these studies reported no dataset information, whereas the remaining 13 described datasets that were not among the widely shared benchmarks listed in the table.

Note: Values are descriptive medians with minimum–maximum ranges; a single value denotes $n = 1$. Corpus- n values reflect dataset names identified from each study’s dataset description in Supplementary Table S1 using a documented keyword match against standard benchmark names. Studies using more than one named dataset are counted under each applicable dataset; therefore, the corpus- n column does not sum to 37.

Reported accuracy values include only studies from which an absolute accuracy value could be extracted according to the protocol described in Section 4.2. These summaries are descriptive and should not be interpreted as pooled effect-size estimates or statistically validated comparisons.

[†]The KDD99 value is reported by (Nagarjun and Rajkumar, 2025), which is identified in Section 4.3 as a multi-task study; see also Table 11.

other security tasks; it is retained because its architecture and dataset are independently verifiable (Supplementary Table S2), with this scope limitation noted for readers rather than treated as grounds for exclusion.

4.4 Dataset-Level Descriptive Synthesis

Table 12 reports how frequently each major benchmark dataset appears across the final, audited 37-study primary corpus, identified directly from each study’s own dataset description (Supplementary Table S1), and the reported accuracy among the subset of studies using that dataset that also report an extractable accuracy value. **Because n with an extractable metric is 1–3 for every dataset listed, these accuracy values must not be read as a performance ranking between datasets; the corpus-adoption column, by contrast, is a complete, verified count over all 37 studies.**

CICIDS2017 and UNSW-NB15 remain the most frequently used named benchmarks in the corrected final corpus (12 and 5 of 37 studies respectively), consistent with their established role in the wider intrusion-detection literature, but a substantial minority of the corpus (17 of 37 studies, 46%) uses a dataset other than the

widely-shared benchmarks listed in Table 12 – most commonly a custom or proprietary collection that is still described in some detail (13 of the 17), with a further 4 studies providing no dataset information at all (Supplementary Table S3, Participants domain). This dataset fragmentation, rather than non-reporting alone, is the more precise reporting-completeness finding relevant to Section 5.2. Because so few studies per named dataset report a directly comparable accuracy figure, this synthesis does not support a claim that any one dataset is associated with systematically higher or lower reported performance; the differences visible in Table 12 are at least as plausibly explained by differences in architecture, proxy definition, and reporting practice as by the dataset itself, and readers should not treat the accuracy column as a dataset performance ranking.

4.5 Reported Performance, Illustrated

Figure 6 plots the extracted accuracy values by architecture family for the 14 studies (of the 15-study quantitative subset, and of 37 primary-corpus studies) with an extractable accuracy value (Supplementary Table S2). The remaining quantitative-subset study contributes an extractable F1 value only. The figure shows both the spread of accuracy values within each family and the small number of studies behind each point; no pooled effect size is computed in this synthesis (Section 4.1).

Because the underlying counts are small, Figure 6 is presented as a descriptive illustration of reported values rather than as a forest plot of pooled effect sizes, and no formal test for publication bias (e.g., Egger’s regression) is reported, since such a test presumes a pooled effect-size model that this synthesis does not compute.

4.6 Architecture-Level Evaluation-Rigour Coverage

Figure 7 cross-tabulates the primary architecture families against eight evaluation-rigour dimensions: explicit unseen- or unknown-attack testing, temporal or drift evaluation, cross-dataset or domain-shift testing, open-set rejection, adversarial or robustness evaluation, explainability, latency or resource reporting, and evaluation across multiple datasets or external settings. Each cell reports the number of studies satisfying the dimension relative to the number of studies in that architecture family. Coding was conservative: a dimension was counted only when it was explicitly supported by the supplied study-level extraction and audit records.

Figure 7 shows that explicit unseen- or unknown-attack testing is more common than temporal, cross-dataset, or open-set evaluation. Robustness, explainability, and resource reporting are concentrated mainly in the larger Hybrid and Unspecified/other DL categories, whereas several smaller architecture families contain little or no explicit evidence for these dimensions. Because architecture-family sizes range from one to thirteen studies, the matrix describes evidence coverage and should not be interpreted as a ranking of architecture quality or predictive performance.

4.7 Explainability: Reported Adoption

Explainability mechanisms are of particular interest for operational trust in security-critical deployments, where they support alert triage, analyst review of false positives,

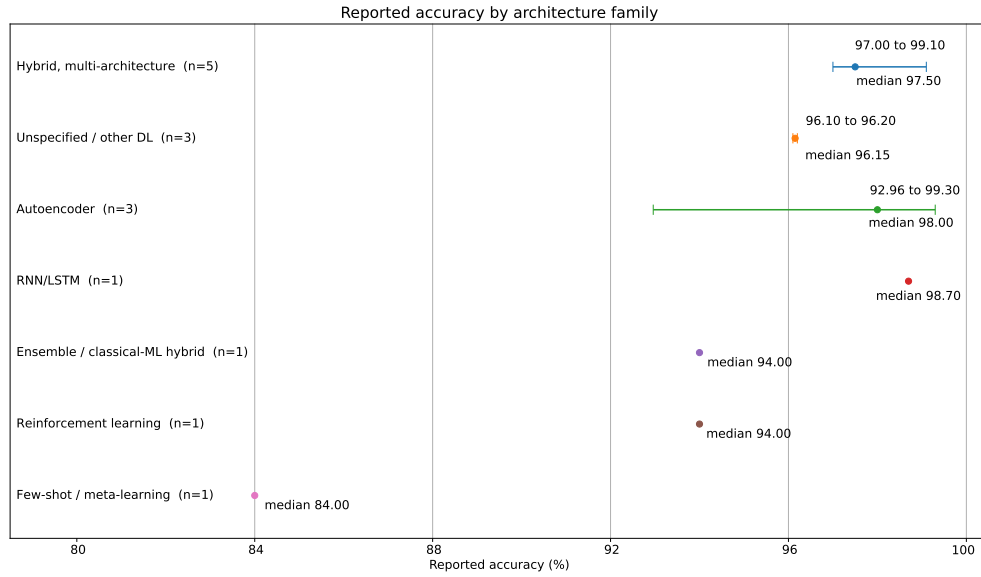


Fig. 6: Reported accuracy by architecture family across the 14 studies (of 37 in the corrected audited primary corpus) with an extractable, architecture-specific accuracy value. Markers show the family median; the grey horizontal line, where present, spans the reported minimum–maximum range (not a confidence interval); a marker with no line denotes $n=1$. The n label beside each point is the number of studies contributing an accuracy value to that family. The 15-study quantitative subset contains one additional study contributing an F1 value only. CNN, Federated, and Transformer families are omitted because no study in those families reported an extractable absolute value in the audited set (the only CNN-tagged study with an extractable value did not meet the review’s final empirical primary-study criteria and was removed at the eligibility audit; Section 3.4.1). Source data: Supplementary Table S2.

and auditability of automated decisions. Table 13 reports how many of the 37 studies in the corrected final audited primary corpus mention each explainability technique, based on keyword searches across the study-level title, methodology-summary, and findings-summary fields used in the review extraction workflow. Supplementary Table S1 provides the corresponding study identifiers and corpus membership; the keyword coding was performed on the underlying study-level extraction records rather than on the displayed columns of Table S1. The explainability keyword search was checked against the corrected 59-record venue-verified set; none of the 22 records excluded at the eligibility audit matched an explainability keyword in the supplied title, methodology, or findings material, so the count remains 8 while the corrected denominator is 37. This is an adoption count, not a moderator analysis: the corpus does not contain the matched with/without-explainability performance comparisons that would be required to estimate a moderating effect on detection performance, so no effect-size column is reported.

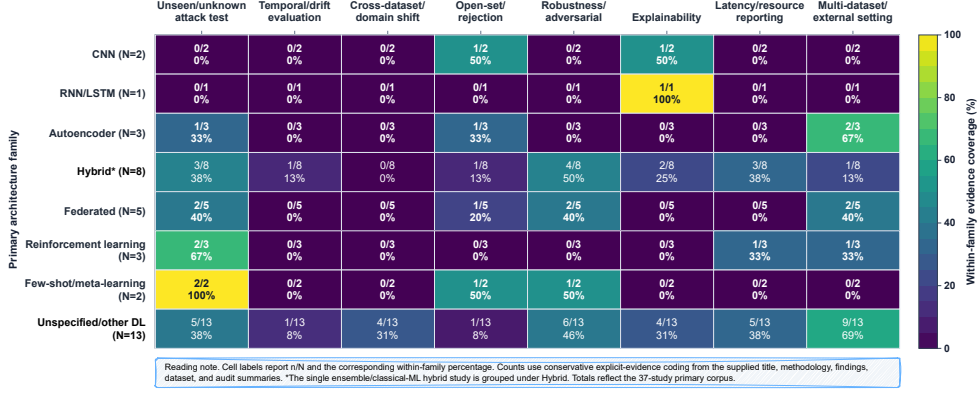


Fig. 7: Architecture-by-evaluation-rigour matrix for the 37-study primary corpus. Cells report study counts and within-family percentages based on explicitly documented evidence. The single ensemble/classical-ML hybrid study is grouped under Hybrid.

Attention visualization is the most frequently mentioned explainability mechanism in this corpus (6 of 37 studies), consistent with its architectural convenience in the increasingly common transformer- and LSTM-based designs (Section 4.3). SHAP is mentioned in 2 studies and LIME is not mentioned at all in this corpus. Overall, only 8 of the 37 studies (22%) mention any explainability technique, which is itself an informative descriptive finding; explainability remains a minority practice in this literature rather than a standard component of the evaluation pipeline.

It has been argued in the wider literature (Section 7) that architecturally integrated interpretability, as opposed to purely post-hoc explanations, provides a more reliable route to improving both trust and performance in DL-based ZDA detection systems; this corpus’s adoption counts are consistent with attention-based (architecturally integrated) mechanisms being more commonly reported than post-hoc SHAP/LIME explainers, though the sample sizes here are too small to treat this as more than a descriptive observation.

4.8 Robustness: Adversarial Attacks and Defenses

Thirteen of the 37 studies in the corrected final audited primary corpus (35%) mention adversarial robustness, evasion, poisoning, or perturbation-related evaluation in their title, methodology, or findings text, coded from the underlying study-level extraction records used for the review coding; the remaining 65% report detection performance without an explicit robustness evaluation. Table 14 and Table 15 summarize the adversarial attack types and defensive mechanisms discussed qualitatively in the representative studies cited in this section and in Section 6, rather than as a pooled quantitative estimate: the corpus does not contain enough studies reporting a matched clean-versus-attacked accuracy pair, with variance, to support a pooled effect-size column. The qualitative pattern reported here (iterative/optimization-based attacks are more damaging than single-step attacks; adversarial training is the most consistently

Table 13: Adoption of explainability techniques in the audited 37-study primary corpus.

Method	Typically Applied To	Studies Mentioning Method	IDS Operational Use
SHAP	CNN, transformer, and ensemble/tree-based hybrid models.	2 of 37	Provides post-hoc feature attribution for SOC analyst review of individual alerts, although its computational cost may limit large-scale use.
LIME	Model-agnostic architectures.	0 of 37	Provides local surrogate explanations for case-by-case alert review. It was not mentioned in the audited corpus but has been reported as unstable under noisy traffic in the wider literature discussed in Section 7.
Attention visualization	Transformer and LSTM architectures.	6 of 37	Highlights the traffic features that influence a classification, supporting rapid analyst triage. It is natively available in attention-based architectures.
Saliency maps	CNN architectures.	0 of 37	Could support visual feature auditing, but it was not explicitly mentioned in the corpus text examined.
TCAV/concept-based explanations	CNN and RNN architectures.	0 of 37	Could support compliance-oriented and concept-level model review, but it was not mentioned in the audited corpus.
Any explainability method mentioned	–	8 of 37 (22%)	The remaining 29 studies reported detection performance without an explicitly identified explainability component.

Note: The values represent adoption counts rather than a moderator or effect-size analysis. Counts were obtained through keyword searches for “shap”, “lime”, “attention”, “saliency”, “tcav”, “explainab*”, and “interpretab*” across each study’s title, methodology summary, and findings summary in the underlying study-level extraction records used for the review coding. A keyword match does not necessarily indicate that the technique was used as a primary evaluation component. Conversely, the absence of a match does not confirm that the technique was absent from the full article, because full text was not available for every study.

Table 14: Adversarial attack types discussed in the representative studies reviewed in Section 6.

Attack Type	Threat-Model Class	Reported Directional Impact	Description
FGSM (white-box)	Single-step, gradient-based	Moderate degradation.	Uses computationally inexpensive, single-step gradient-based perturbations to modify input features (Han et al., 2023).
PGD (iterative)	Multi-step, gradient-based	Larger degradation than FGSM.	Uses iterative projected-gradient updates to generate stronger adversarial perturbations (Wu et al., 2022).
CW (optimization)	Optimization-based, white-box	Largest degradation among the reviewed attack types.	Uses optimization-based perturbations designed to produce misclassification while minimizing input distortion (Wu et al., 2022).
Transfer-based (black-box)	Black-box, surrogate-model based	Moderate degradation without requiring direct access to the target model.	Generates adversarial examples using a surrogate model and transfers them to the target detection model (Li et al., 2022).
Adversarial malware (domain-specific)	Feature-space evasion	Moderate, domain-specific degradation.	Manipulates malware feature representations to evade detection while attempting to preserve malicious functionality (Li et al., 2022).

Note: This table provides a qualitative synthesis of the attack-severity ordering reported in the representative literature cited in this section. It does not present a pooled statistical estimate. The corrected 37-study primary corpus does not contain enough matched clean-versus-attacked performance values, together with corresponding variance estimates, to support a pooled percentage-degradation or effect-size analysis. Thirteen of the 37 studies in the final audited corpus (35%) explicitly mention adversarial robustness evaluation, as discussed in Section 4.8.

reported defense) is drawn from, and referenced to, the specific representative studies cited below; none of these representative studies are among the 22 records removed at the eligibility audit.

In general, iterative and optimization-based adversarial attacks (e.g., PGD and CW) are reported to cause larger degradation in detection performance than single-step attacks such as FGSM, highlighting the vulnerability of many DL-based IDSs

Table 15: Defensive mechanisms discussed in the representative studies reviewed in Section 6.

Defence Mechanism	Reported Robustness Benefit	Notes
Adversarial training	Most consistently reported benefit.	Reported as the most effective defence in the cited literature, although it incurs substantial computational and training costs (Wu et al., 2022; Mhawi et al., 2022).
Input pre-processing	Small reported benefit.	Provides a lightweight and model-agnostic defence that can reduce the effect of input perturbations (Mhawi et al., 2022).
Gradient regularization	Small reported benefit.	Encourages smoother decision boundaries and reduces model sensitivity to small input perturbations (Han et al., 2023).
Robust optimization	Reported benefit comparable to adversarial training.	Has a strong theoretical foundation but is rarely evaluated in the reviewed corpus (Wu et al., 2022).

Note: This table presents a qualitative synthesis of the relative defence effectiveness reported in the representative studies cited in this section. It does not constitute a pooled statistical estimate or a controlled comparison across identical datasets, models, and threat settings.

to adaptive threat models. These findings motivate the inclusion of explicit defensive mechanisms to improve robustness, as summarized in Table 15.

From a broader perspective, attacks that use iterative optimization algorithms, such as projected gradient descent (PGD) and Carlini–Wagner (CW) attacks, consistently result in the largest accuracy degradation. In terms of defense, adversarial training and robust optimization stand out as the most effective methods for improving robustness, offering the greatest improvement. Nevertheless, they require a greater number of computations and have been applied only to a limited extent.

4.9 Heterogeneity, Reporting-Credibility Summary, and Implications

In the descriptive synthesis of reported values, apparent variation across studies reflects differences in the underlying datasets, class-imbalance handling, feature engineering, and evaluation-metric protocols, rather than a formally quantified heterogeneity statistic (Section 4.1). The reporting-completeness and methodological-credibility issues discussed in Section 3.5 are directly relevant here: common limitations across the primary corpus include insufficient validation across multiple datasets, underreporting of variance, and reliance on a single benchmark dataset per study.

Despite these limitations, the descriptive quantitative synthesis suggests a recurring pattern of high reported accuracy for DL techniques under proxy zero-day conditions across the architectures listed in Table 11, albeit over small per-architecture

samples (Section 4.3). Attention-based interpretability and adversarial training are the most frequently reported explainability and robustness practices in this corpus (Tables 13–15), but both remain minority practices relative to studies reporting detection accuracy alone.

To improve replicability and generalizability, future studies should adopt a more uniform reporting procedure that includes variance estimates and confidence intervals. More robust ZDA proxy approaches that incorporate both time-series and cross-domain assessments, as well as robustness checks using adaptive adversary models, should also be adopted.

4.10 SWOT Analysis of DL-Based ZDA Detection

Building on the systematic review and descriptive quantitative synthesis presented in Sections 3 and 4, the audited primary corpus is dominated by the Unspecified/other DL and Hybrid/multi-architecture categories (Table 11), while the broader contextual literature in Table 10 also includes transformer, graph-neural-network, and self-supervised approaches. Reported benchmark accuracies are frequently high in the small quantitative subset, but this corpus’s small per-architecture samples do not support a general claim that any one family consistently outperforms traditional machine learning approaches (Section 4.3). These methods are designed to learn high-dimensional spatiotemporal representations and to generalize to previously unseen attack patterns.

However, the synthesis also revealed persistent limitations. Despite their strong detection performance, many DL-based IDSs remain sensitive to dataset bias, evaluation protocol design, and the specific operationalization of zero-day proxies. As reflected by the heterogeneity and robustness analyses reported in Tables 14 and 15, the reported performance varies substantially across studies, indicating inconsistent experimental practices and limited cross-dataset validation. Although explainability and robustness mechanisms have been explored (Table 13), they have not yet been systematically integrated into evaluation pipelines, thus constraining operational trust and deployment readiness.

Taken together, these observations motivate a balanced deployment-aware perspective on ZDA detection. Rather than focusing solely on architectural novelty or peak accuracy, the literature increasingly points to the need for robust, interpretable, and generalizable detection systems. The following SWOT analysis synthesizes these insights by mapping the strengths and limitations of current DL approaches against emerging opportunities and external threats.

1. **Strengths:** DL-based IDSs frequently report strong benchmark performance under proxy zero-day conditions. In the audited corpus, Hybrid/multi-architecture systems account for the largest share of studies contributing an extractable performance value (5 of 15; Table 11). The literature also demonstrates methodological breadth across supervised, unsupervised, self-supervised, federated, few-shot, and reinforcement-learning approaches. These findings indicate flexibility across heterogeneous security settings, while remaining descriptive because per-family samples are small and proxy protocols differ.

2. **Weaknesses:** Despite their performance advantages, current approaches inherit several limitations. As shown in Table 12 and Figure 9, evaluation remains concentrated on a small set of benchmark datasets and explicit temporal, cross-dataset, and open-set validation is limited; Table 13 and Figure 7 further show sparse explainability and deployment-oriented reporting. Reliance on simplified proxy definitions of zero-day behavior can inflate performance estimates and limit real-world generalizability. High computational overhead, particularly for transformer-based and ensemble models, further constrains deployment in resource-limited environments. In addition, explainability mechanisms are often treated as post-hoc additions rather than as integral components of system design, reducing transparency and analyst trust.
3. **Opportunities:** The synthesis reveals multiple opportunities for advancing ZDA detection research. Table 12 highlights the growing effectiveness of self-supervised and representation-learning approaches, suggesting a strong potential for reducing the reliance on labeled data. In parallel, the corpus does not contain matched with/without-explainability comparisons sufficient to assess whether explainability is associated with any gain or loss in detection performance; the defensible corpus-level finding is limited to adoption, with 8 of 37 studies mentioning an explainability technique (Table 13). Emerging research directions, including continual learning, federated deployment, and robustness-aware evaluation, remain important opportunities for improving adaptability, privacy, and operational resilience.
4. **Threats:** Several external and methodological threats may hinder progress. As summarized in Tables 14 and 15, high between-study heterogeneity and inconsistent reporting practices limit reproducibility and cross-study comparability. The limited and uneven evaluation of adversarial robustness and distributional shifts exposes deployed models to evasion, poisoning, and concept drift. Moreover, the tendency to equate proxy-based results with true ZDA detection capability poses a risk of overclaiming effectiveness, potentially undermining trust in operational deployments.

Figure 8 summarizes the SWOT synthesis derived from the descriptive corpus evidence in Tables 11–15 and the thematic analyses in Sections 4–7. The same 13-study “Unspecified/other DL” and 7-study Hybrid families that dominate corpus adoption (Strengths) are also, by construction, the families in which most of the corpus’s few robustness- and explainability-reporting studies fall (Weaknesses), since these two families jointly account for the majority of the 37-study corpus.

4.10.1 Implications

This SWOT-based analysis contextualizes the quantitative findings of the descriptive synthesis by linking empirical performance trends to broader methodological and deployment considerations. The results emphasize the need for standardized zero-day proxy evaluation protocols, greater emphasis on robustness and explainability, and more realistic cross-dataset and longitudinal testing methods. These insights directly inform the open challenges and future research directions discussed in the following section, particularly in the pursuit of scalable, interpretable, and resilient DL-based IDSs that can operate under realistic zero-day conditions.



Fig. 8: SWOT analysis of deep learning-based intrusion detection under proxy zero-day conditions, derived from the reviewed evidence.

5 Datasets for ZDA Detection

Selecting a dataset for use is a critical element in determining both the performance and variability in DL-based ZDA detector studies. As indicated in Table 12, dataset realism, class imbalance handling, and ZDA proxy definition are all plausible contributors to the variability in reported accuracy across studies, though with the small per-dataset samples in this corpus these associations are descriptive rather than statistically tested.

Of the 37 corrected audited studies in the primary corpus, many relied on benchmark datasets to approximate zero-day conditions through attack-class hold-out, temporal evaluation, or cross-dataset testing. Despite their use to mimic novelty, these studies introduce their respective biases, which directly influence the descriptive synthesis reported in Section 4.

5.1 Dataset Characteristics and Reported Performance

As shown in Table 12, CSE-CIC-IDS2018, which uses a more recent, temporally organized traffic collection design, is associated with a reported accuracy of 97.5% among corpus studies reporting a value, though this rests on a single study ($n = 1$) and cannot be generalized.

UNSW-NB15 appears in 5 of the 37 primary-corpus studies. Among the two studies reporting an extractable accuracy value, the reported values range from 96.2–98.0%, with a median of 97.1%. These values remain descriptive and should not be interpreted as a dataset-level performance estimate.

CICIDS2017 is the most frequently used named benchmark in the final corpus, appearing in 12 of 37 studies. Among the three studies with an extractable accuracy

value, the reported values range from 94.0–99.3%, with a median of 98.0%. Differences among these values may reflect architecture, proxy design, data processing, and reporting practices rather than an intrinsic property of the dataset.

CIC-IoT2023, CSE-CIC-IDS2018, KDD99, Edge-IIoTset, N-BaIoT, BoT-IoT, and ToN-IoT together account for a further 10 named-dataset uses across the corpus (Table 12). CIC-IoT2023 and ToN-IoT contribute no value passing the extraction filter in Section 4.2, while the single KDD99-based study reports an extractable accuracy value (99.1%, flagged in Section 4.3 as a multi-task study).

Figure 9 extends the dataset-frequency analysis by cross-tabulating each named benchmark against four proxy zero-day evaluation strategies: unseen- or unknown-attack testing, temporal or drift evaluation, cross-dataset or domain-shift testing, and open-set rejection. This representation distinguishes frequent benchmark use from diversity in evaluation design. Coding was conservative and was based only on explicitly documented evidence in the supplied extraction and audit records.

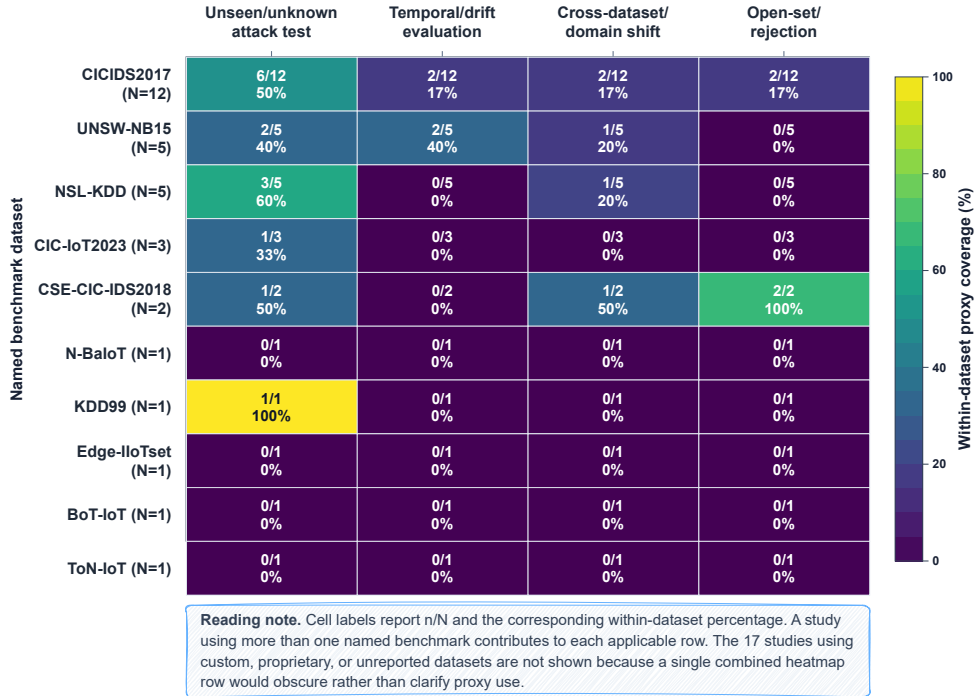


Fig. 9: Dataset-by-proxy-strategy heatmap for named benchmarks in the 37-study primary corpus. Cells report study counts and within-dataset percentages; studies using multiple datasets contribute to each applicable row.

Figure 9 shows that CICIDS2017 has the broadest proxy coverage among the named benchmarks, although its temporal, cross-dataset, and open-set coverage

remains limited. UNSW-NB15 and NSL-KDD are used mainly for unseen- or unknown-attack evaluation, with little explicit evidence of open-set rejection. Several datasets occur in only one study and therefore cannot support comparative conclusions about proxy suitability. Zero-valued cells indicate absence of explicitly documented evidence in the coded corpus, not that a dataset is inherently unsuitable for that evaluation. The 17 studies using custom, proprietary, heterogeneous, or unreported datasets are omitted because combining them into one heatmap row would obscure their methodological differences.

5.2 Dataset Evolution and Zero-Day Validity

Older datasets, such as KDD Cup 99 and NSL-KDD, are widely reported in the wider literature to be less representative of current traffic patterns and attack categories, and their continued use is a recognized source of optimistic bias in intrusion-detection benchmarking generally. NSL-KDD still appears in 5 of the 37 audited primary-corpus studies (Table 12), illustrating that this concern applies directly to the present corpus and not only to the field in the abstract.

Mid-generation datasets, such as UNSW-NB15, provide greater feature variability than KDD-era datasets but remain synthetically generated and have known limitations in transferring to real-world environments. More recent datasets, such as CICIDS2017 and CSE-CIC-IDS2018, provide richer flow features and better-documented temporal dynamics, which is consistent with (though not proof of) the comparatively high reported accuracies for these two datasets in Table 12. Class imbalance and rigid attack definitions remain a limitation across all of the datasets in the corpus, regardless of generation.

5.3 Implications for Generalisation and Quantitative Synthesis

The descriptive synthesis highlights that differences in datasets are as crucial as model architectures in reported performance. This indicates that these improvements result from a combination of dataset properties, model architecture, and evaluation setup.

In cases in which studies are split along the temporal dimension or tested on other datasets, the magnitude of the reported effects is somewhat lower but more generalizable overall. This indicates the need for greater rigour in testing and dataset creation.

5.4 Future Dataset Directions

To overcome the limitations of ZDA proxy evaluations, next-generation datasets must incorporate properties such as encrypted traffic, concept drift, adaptive attackers, and class distributions. Collaboration with security operations centres and cyber ranges will play a vital role in next-generation datasets.

Such datasets would be beneficial for conducting more rigorous quantitative comparisons and for reducing confounds arising from the datasets used across studies.

6 Adversarial Robustness of DL-Based ZDA Detection

Although DL-driven intrusion detectors often perform well on clean test datasets, they are susceptible to adversarial manipulation. A model can be tricked into making the wrong decision by maliciously constructing perturbations that preserve malicious traffic and payloads. This is a serious risk, particularly in cybersecurity, where adversaries can easily repurpose or modify known exploits to evade defenses without altering the attack methodology.

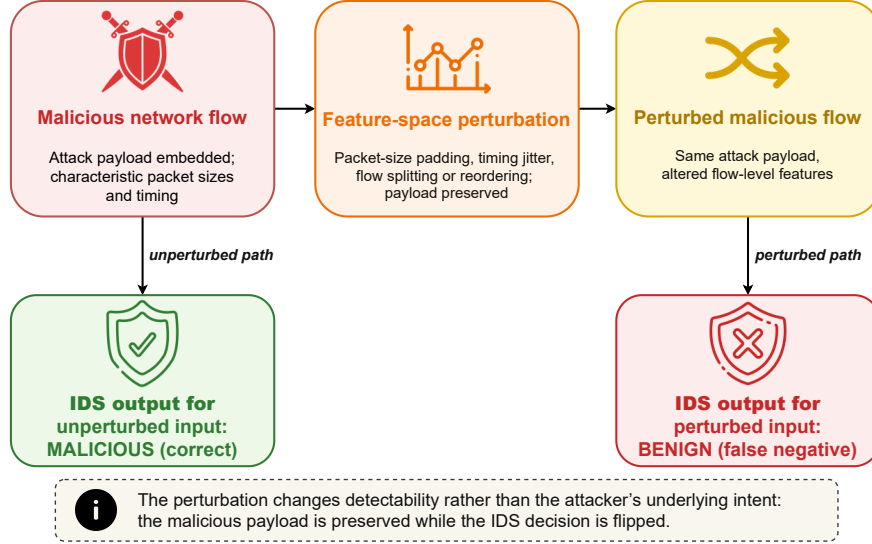


Fig. 10: Schematic example of an adversarial evasion attack on a network-flow-based IDS. The attacker starts from genuinely malicious traffic and perturbs flow-level features (packet-size padding, timing jitter, flow splitting/reordering) while preserving the attack’s functional payload. An IDS that correctly classifies the unperturbed malicious flow can be induced to misclassify the perturbed version as benign (a false negative), illustrating evasion rather than a change in the attacker’s underlying intent.

These flaws are particularly concerning in settings where safety is critical, such as industrial control systems, hospital networks, automotive systems, and large enterprise networks. Figure 10 illustrates a network-flow-based instance of this problem: an attacker perturbs flow-level features, such as packet-size padding, timing jitter, or flow splitting, while preserving the underlying malicious payload, causing an otherwise accurate IDS classifier to output “benign”. Such undetected evasions can have serious operational, financial, or safety repercussions. In this section, adversarial threat models, attack types, robustness in real-world scenarios, and defensive tactics that support DL-based ZDA detection are systematically reviewed.

6.1 Threat Model and Adversarial Attack Taxonomy

When discussing IDS attacks, the focus is typically on the information that the attacker has access to in the target model. For instance, in white-box attacks, an attacker has full knowledge of the model architecture, parameters, and learning methods, enabling optimized attacks that leverage gradient information. In contrast, in black-box attacks, limited or no information is available in terms of access, with attacks based on query feedback or the ability of adversarial samples to be transferred from other models.

According to the literature on DL-based IDS systems, attack techniques against DL-based IDS systems mainly fall under four categories: (i) Gradient-based attack techniques, (ii) optimisation-based attack techniques, (iii) Transfer-based Black-box attack techniques, and (iv) Adversarial Malware/Traffic. These strategies differ in terms of the computational complexity involved and the distance an attack must travel to remain undetected by an IDS system. Technical details are provided in Table 14.

6.2 Quantitative Impact of Adversarial Attacks

Across the representative studies discussed in this section, adversarial attacks are consistently reported to degrade IDS performance. One-shot gradient attacks, such as the fast gradient sign method, reduce accuracy by approximately 7%–9% on average, as shown in Table 14, whereas deeper and more aggressive projected gradient descent attacks or optimization-based Carlini–Wagner attacks may result in a reduction of more than 10%.

Black-box attacks based on transfer learning introduce an additional layer of complexity, particularly in practical attack scenarios in which the attacker lacks access to the model parameters. These attacks exploit the fact that adversarially transferred attacks exist, particularly when scalability is achieved in the architecture or data distribution. For applications in IoT and malware classification, domain adversarial attacks and manipulations in the feature space confirm evasion success without undermining malicious characteristics.

These qualitative patterns are consistent with the descriptive performance synthesis in Section 4 and support the conclusion that adversarial robustness remains a significant bottleneck in the practical deployment of DL-based ZDA detection systems; no pooled effect size is claimed or computed for either synthesis.

6.3 Defensive Mechanisms and Their Limitations

Several defensive strategies have been proposed to mitigate adversarial vulnerabilities. These strategies for countering adversarial attacks include adversarial training, input preprocessing, gradient regularization, and robust optimization. As shown in Table 15, adversarial training is one method that has demonstrated the highest average robustness gain; however, it is associated with a high computational overhead.

In light of these mechanisms, although robustness to first-order attacks is improved, adaptive, iterative, and transfer-based attacks remain problematic for DL models. Generative models, such as GANs and VAEs, are highly vulnerable to adversarially designed attack samples that exploit uncertainties in latent representations. Moreover,

defenses designed to specifically counter a certain attack norm or threat may not generalize well to other attack types.

Such deficiencies indicate the need to implement defensive tactics to address the variability in attack types, mitigate dataset bias, and assess defensiveness and robustness.

6.4 Formal Definitions of Common Adversarial Attacks

For completeness and reproducibility, the main text presents concise canonical mathematical definitions of commonly used adversarial attacks, including FGSM, PGD, Carlini–Wagner, and transfer-based black-box attacks, which are sufficient to support the robustness analysis.

Here, $J(\cdot)$ denotes the training loss, $L(\cdot)$ is the surrogate loss, $\Pi_{B_p(x,\epsilon)}(\cdot)$ is the projection operator onto an ℓ_p -bounded perturbation set, and $f(\cdot)$ is a margin-based misclassification objective.

6.4.1 Fast Gradient Sign Method.

The FGSM is a single-step, gradient-based attack that perturbs the input in the direction of the loss gradient:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)),$$

where ϵ controls the perturbation magnitude under an ℓ_∞ constraint.

6.4.2 Projected Gradient Descent.

In PGD, FGSM is generalized by applying iterative updates followed by projection onto a norm-bounded perturbation set:

$$x^{(t+1)} = \Pi_{B_p(x,\epsilon)} \left(x^{(t)} + \alpha \cdot \text{sign} \left(\nabla_{x^{(t)}} J(\theta, x^{(t)}, y) \right) \right)$$

6.4.3 Carlini–Wagner Attack.

The CW attack formulates adversarial example generation as a constrained optimization problem that minimizes the perturbation magnitude while inducing misclassification as follows:

$$\min_{\delta} \|\delta\|_2 + c \cdot f(x + \delta) \quad \text{s.t.} \quad x + \delta \in [0, 1]^n.$$

6.4.4 Transfer-Based Black-Box Attacks.

Transfer-based attacks exploit the transferability of adversarial examples generated by surrogate model M_s .

$$\delta^* = \arg \max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{M_s}(x + \delta, y),$$

where the attack success is defined as $M_t(x_{\text{adv}}) \neq y$ when applied to a target model M_t .

6.5 Summary and Security Implications

Adversarial attacks pose a clear threat to DL systems for detecting ZDA, as attackers can craft extremely small and carefully designed perturbations to cause models to misclassify or exploit their blind spots.

Although current defense mechanisms can increase resilience against certain attack methods, they are ineffective against adaptive or transfer attackers. Previous results, which indicate how threat scenarios, dataset biases, and assessment methods interact from the perspective of an adversarial attack, are consistent with the descriptive quantitative synthesis introduced in Section 4.

In general, robustness-aware benchmarking, open reporting of adversarial settings, and dataset designs that accurately reflect real-world threat scenarios are required to improve the practical feasibility of ZDA detection systems. Performance enhancements under benign evaluation assumptions may overstate real-world security guarantees in the absence of such advancements.

7 Explainability and Ethics

In cybersecurity, explainability facilitates debugging, improves the performance of detection systems, fosters confidence among security professionals, and supports well-informed decision-making. Figure 11 illustrates a typical DL-based IDS explainability pipeline and its role in improving threat detection. To verify predictions and take appropriate action, analysts require more than classification outputs; they also need insights into the reasoning behind each conclusion (Hassija et al., 2024). In security-critical applications, DL models are unlikely to be adopted at scale without explainability. Furthermore, accountability and transparency are increasingly required by legal frameworks, such as the GDPR and the EU AI Act, particularly when AI systems process sensitive data or control critical infrastructure (Vinayakumar et al., 2019).

7.1 Techniques for Interpreting DL Models

Several methods have been proposed to provide insights into how DL-based IDS and ZDA detectors make decisions. Table 16 compares the strengths and limitations of these methods qualitatively and reports, for each, the exact number of primary-corpus studies mentioning it (Table 13). These observed patterns align with the descriptive explainability-adoption synthesis outlined in Section 4, which found that only a minority of corpus studies (8 of 37, 22%; Section 4.7) report an explainability technique alongside detection performance.

7.1.1 SHapley Additive exPlanations

SHAP works by attributing feature-level contributions to estimate the extent to which a given input feature pushes a prediction. In the context of IDS, SHAP has been used

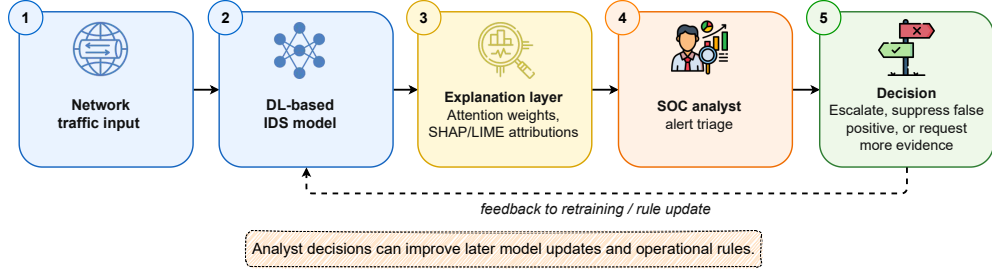


Fig. 11: DL-based IDS explainability pipeline: raw traffic is scored by the detection model; an explanation layer (e.g., attention weights, SHAP/LIME attributions) attaches a rationale to each alert; a SOC analyst uses that rationale during alert triage; the resulting decision (escalate, suppress as false positive, or request more evidence) can feed back into model retraining or rule updates.

Table 16: Comparison of explainability methods applied to DL-based ZDA detection.

Method	Primary mentions (	Strengths	Limitations
SHAP (Annapareddy et al., 2025)	2/37	Provides global and local feature attribution with strong theoretical foundations.	Can be computationally expensive when applied to large or complex models.
LIME (Wu et al., 2022)	0/37	Provides simple, model-agnostic, and case-specific explanations with relatively low computational requirements.	May produce unstable explanations when applied to noisy or perturbed network traffic.
Attention (Vaswani et al., 2017)	6/37	Is naturally integrated into attention-based architectures and highlights salient input features or temporal relationships.	Attention weights may not accurately represent the causal reasoning underlying the model prediction.
Saliency (Vinayakumar et al., 2019)	0/37	Provides intuitive visual indications of influential input features through saliency scores or heatmaps.	Offers limited temporal and contextual information and may be sensitive to small input changes.
TCAV (Manocchio et al., 2024)	Rare	Links human-defined concepts with internal neural activations, supporting concept-level interpretation.	Requires carefully curated concept datasets and additional validation of the selected concepts.

to identify traffic attributes that have the most significant influence on malicious classifications. This has proven useful during post-hoc analysis and alert validation (Hindy

et al., 2020). However, this approach incurs a trade-off in terms of computational overhead, which may limit its use in real-time or resource-constrained settings.

7.1.2 Local Interpretable Model-Agnostic Explanations

LIME generates a local explanation by approximating a complex model around its prediction using a simpler surrogate model. Due to its small footprint and model-agnostic characteristics, it is suitable for quick, case-by-case interpretation; however, explanations can be unstable under noisy or adversarial conditions (Wu et al., 2022).

7.1.3 Attention Visualisation

Attention-based explainability methods highlight the parts of the input sequence that the model emphasizes during prediction. Because transformer-based IDS models naturally expose attention weights, they provide an interpretable understanding of which packets or flow-level features drive such detection decisions. Empirical studies suggest that attention mechanisms strike a strong balance between interpretability and performance for modern IDS designs (Manocchio et al., 2024; Han et al., 2023).

7.1.4 Saliency Maps.

All saliency-based techniques compute the gradients of the output with respect to the input features to identify the most influential parts of the input. Because these generated visualizations can only help analysts identify salient traffic characteristics, they fail to capture long-range temporal dependencies in sequential network data (Vinayakumar et al., 2019).

7.1.5 Concept-Based Interpretability

Concept-driven approaches aim to link the model’s choice of actions with overall human-interpretable patterns of activity, rather than direct features. Although these methods have the potential to support relevance validation, they are not widely used in the IDS research community because they require high-quality concept-based datasets.

In summary, explainability is an important component of trustworthy DL-based proxy-ZDA detection. In the audited 37-study primary corpus, 8 studies mention an explainability technique (Section 4.7). The defensible corpus-level conclusion is therefore that explainability remains sparsely reported; the available evidence does not support a causal conclusion about its effect on detection performance, although explanation methods can support analyst interpretation, alert triage, and auditability.

7.2 Ethical and Governance Considerations

Although ZDA detection is primarily a technical security problem, the deployment of DL-based IDS introduces governance and ethical risks that can affect trust, accountability, and safety. First, traffic telemetry can contain personal or commercially sensitive information; therefore, data minimization, purpose limitation, retention policies, and privacy-preserving learning should be treated as design requirements rather than optional add-ons. Second, dataset bias and deployment shift can lead to uneven

error profiles across networks, devices, and user populations, which is operationally consequential when false positives trigger automated responses. Third, transparency and auditability are needed for incident response and compliance: model updates, training data provenance, and evaluation under explicit proxy settings should be logged to support post-incident review and external auditing. Finally, A secure deployment requires attention to model security itself (e.g., adversarial robustness, data poisoning, and update integrity) because compromised IDS models can amplify risk rather than reduce it.

8 Future Directions and Open Challenges

DL holds great promise for the large-scale detection of ZDA in enterprises, IoT, SDN, and industrial networks. However, the 37 studies in the corrected audited primary corpus analyzed here revealed numerous research gaps regarding the robustness, interpretability, generalizability, and practicality of applying this solution in real-world settings.

8.1 Research Gaps and Future Work

Adversarial robustness is a major challenge for existing DL-based IDSs. Reported performance can degrade substantially under gradient-based adversarial attacks associated with optimization tasks, even for the strongest models, such as CNN-LSTM combined models and transformers (Wu et al., 2022; Li et al., 2022; Han et al., 2023). Although adversarial training can be effective to some extent, the substantial computational cost and poor generalization capabilities pose obstacles to its adoption in IDS (Wu et al., 2022; Mhawi et al., 2022). Future work should prioritize adaptive adversarial defenses that remain effective under evolving attack strategies.

A methodological limitation of this review is that the database search relied on the umbrella keyword “deep learning” rather than individual architecture-specific terms. Although this approach ensured a consistent and reproducible search strategy, some studies indexed primarily under alternative descriptors such as “self-supervised learning”, “representation learning”, or “transformer-based intrusion detection” may not have been identified directly and were instead captured through citation chaining where possible.

The lack of model transparency is another factor that hinders the application of DL-based ZDA detectors. Several models that provide high precision remain black boxes and threaten trust (Manocchio et al., 2024; Han et al., 2023). Explaining models by applying trust-enhancing techniques, such as SHAP, LIME, attention maps, and concept-based models, is a promising area of research (Walczyna et al., 2024; Shebl et al., 2024). However, explainability is rarely evaluated quantitatively, underscoring the need for standardized interpretability metrics that assess model faithfulness and stability.

One concern is data imbalance and the limited availability of labeled attack data. Popular benchmark datasets, such as UNSW-NB15, CICIDS2017, and CSE-CIC-IDS2018, exhibit data imbalance, which can lead to overfitting and poor generalization performance (Meliboev et al., 2022; da Silva Ruffo et al., 2024). The contributions of

few-shot, semi-supervised, and self-supervised approaches are highlighted, as they can use data without supervised learning and transfer data from different sources (Sarhan et al., 2023; Golchin et al., 2024). Cross-dataset and temporal analyses require further attention for the evaluation of real-world zero-day data.

A major barrier to real-time deployment of DL-based IDSs is computational cost. Models that involve transformers or ensemble models exhibit high-latency inference and high memory requirements, making their deployment on edge platforms or in resource-scarce settings challenging (Manocchio et al., 2024; Xi et al., 2024). Advances in lightweight models, model compression, pruning, quantization, and federated learning can play a pivotal role in addressing the trade-off between the performance achieved and deployment constraints (Korba et al., 2023; Khan et al., 2024). The use of deployment-oriented benchmarks for latency, memory, and energy must become the norm.

Finally, the potential for further exploration of hybrid and ensemble methods remains. Based on evidence, the potential to improve robustness, interpretability, and generalization by combining anomaly detection (using autoencoders) with time-series analysis (using transformers or LSTMs), as well as other concepts such as explainable AI and privacy-preserving learning, exists (Caville et al., 2022; Vinayak and Jarin, 2025; Dardouri and Almuhan, 2025). This requires further exploration by implementing these systems on diverse datasets, such as UNSW-NB15 and CSE-CIC-IDS2018 (Akshaya et al., 2025; Saurabh et al., 2024; Lilhore et al., 2025). Conventional evaluation metrics, such as accuracy and the F1-score, are widely used but do not reflect performance in the context of the ZDA (Sarhan et al., 2023). A proxy for the zero-day condition rate as a standardized metric for ZDA and proper adversarial robustness benchmarks must be considered (Wu et al., 2022).

As summarized in Table 17, these challenges and the corresponding research directions highlight concrete pathways for advancing DL-based ZDA detection in both academic and operational contexts.

8.2 Implications for the Research and Practitioner Community

The findings of this survey have several implications for both researchers and practitioners developing DL-based intrusion detection systems under zero-day proxy conditions. The proposed taxonomy and evaluation framework provide a structured reference for organizing future studies and encourage consistent reporting of architectural choices, dataset characteristics, robustness evaluation, explainability methods, and deployment considerations. This unified perspective improves the reproducibility and comparability of empirical research while facilitating the identification of methodological gaps.

For researchers, the systematic review and descriptive quantitative synthesis highlight the importance of standardized proxy evaluation protocols, cross-dataset validation, and integrated robustness and explainability assessments. These findings provide evidence-based guidance for selecting benchmark datasets, evaluation metrics, and baseline architectures, thereby supporting the development of more reliable and generalizable IDS models.

Table 17: Summary of key research gaps, proposed technical solutions, and methodological directions for ZDA detection.

Research Gap	Technical Solutions	Research Methodology
Adversarial robustness	Adversarial training and certified defence mechanisms.	Cross-dataset benchmarking under white-box and black-box attack settings (Wu et al., 2022; Li et al., 2022; Han et al., 2023).
Explainability and trust	XAI methods, including SHAP, LIME, attention visualization, and counterfactual explanations.	Quantitative interpretability assessment using faithfulness, sensitivity, and completeness metrics, together with analyst-in-the-loop evaluation (Walczyna et al., 2024; Shebl et al., 2024).
Data imbalance and generalization	Few-shot learning, semi-supervised learning, and contrastive pre-training.	Multi-dataset evaluation using temporal splits and cross-domain generalization testing (Sarhan et al., 2023; Golchin et al., 2024).
Computational efficiency	Lightweight models, pruning, quantization, and federated learning.	Benchmarking of inference latency, throughput, memory footprint, and energy consumption on edge and cloud platforms (Khan et al., 2024; Korba et al., 2023).
Evaluation methodology	Proxy zero-day detection rates and adversarial-robustness benchmarks.	Reproducible cross-laboratory experiments using open datasets and standardized attack scripts (Sarhan et al., 2023; Dai et al., 2024).
Hybrid architectures	Autoencoder-transformer ensembles with integrated explainability mechanisms.	Component-level ablation studies and multi-domain evaluation using enterprise, IoT, and SDN traffic (Vinayak and Jarin, 2025; Tong and Zhang, 2024).

For practitioners, the proposed evaluation framework extends model selection beyond predictive accuracy by incorporating robustness, interpretability, computational efficiency, and deployment readiness. This multidimensional assessment supports the adoption of DL-based IDSs that are not only effective at detecting previously unseen attacks but are also transparent, trustworthy, and suitable for deployment in enterprise, IoT, SDN, and industrial cybersecurity environments.

Collectively, these implications promote the development of reproducible research practices and operationally reliable DL-based IDSs, bridging the gap between academic advances and real-world zero-day attack detection requirements.

9 Conclusion

True zero-day exploitation remains largely absent from public benchmarks; therefore, the current evidence base evaluates generalization under proxy conditions rather than discovery of previously unknown vulnerabilities. This review audited 100 topically screened records and, after final eligibility and venue verification, synthesized 37 peer-reviewed empirical studies. The evidence shows that deep learning architectures often report high accuracy on benchmark datasets, but cross-study comparison is constrained by inconsistent proxy definitions, small metric-reporting subsets, limited variance reporting, and sparse temporal, cross-dataset, open-set, adversarial, explainability, and deployment evaluation. Accordingly, the review does not identify a universally superior architecture and does not treat benchmark accuracy as evidence of true zero-day capability.

The paper contributes four linked outputs: an operational distinction between true and proxy zero-day detection; an auditable taxonomy and evidence synthesis; a map of dataset, architecture, robustness, and explainability gaps; and a seven-dimension evaluation framework covering architecture, dataset context, proxy validity, performance, robustness, explainability, and deployment readiness. The framework shifts evaluation from peak accuracy toward evidence that is reproducible, comparable, and operationally meaningful.

Future studies should predefine the proxy threat model, separate model selection from unknown-attack testing, report uncertainty and false-alarm costs, evaluate temporal and cross-domain generalization, test adaptive adversaries, quantify explanation quality, and disclose latency, memory, and energy. The central conclusion is that progress depends less on reporting another near-perfect benchmark score than on standardizing and stress-testing the conditions under which claims of zero-day generalization are made.

Declarations

- **Funding:** No external funding was received for the conduct of this study.
- **Competing interests:** The authors declare that they have no competing interests.
- **Ethics approval:** Not applicable. This literature review involved no new human participants, animals, or primary data collection.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Protocol registration:** No prospective protocol registration was completed. The search strategy, eligibility criteria, extraction fields, and post-screening audit trail are reported in the manuscript and online resource; the absence of prospective registration is acknowledged as a limitation.
- **Availability of data and materials:** Search strings, verified screening counts, study-level extraction, performance extraction, reporting and methodological credibility judgments, venue-stage background records, and eligibility-audit reasons are supplied as Online Resource 1.

- **Code availability:** The consolidated BibTeX databases and consistency-check materials used to verify corpus-dependent tables are included in the submission source package.
- **Authors' contributions:** M.Z. and M.U.R. conceived and designed the review. M.Z., M.K., and S.K. conducted the literature search, screening, and data extraction, review and editing. M.A. and E.B. contributed to the quantitative synthesis and quality appraisal. M.U.R. supervised the study. All authors drafted or critically revised the manuscript, approved the submitted version, and accept responsibility for the work.

References

- Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access* **5**, 21954–21961 (2017)
- Bilge, L., Dumitras, T.: Before we knew it: An empirical study of zero-day attacks in the real world. In: *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pp. 833–844 (2012). <https://doi.org/10.1145/2382196.2382284>. ACM
- Google Threat Analysis Group: Year in Review: Zero-Day Exploits in 2023. <https://blog.google/threat-analysis-group/year-in-review-2023-zero-day-exploits/>. Accessed: Aug. 2025 (2025)
- Statista: Number of zero-day vulnerabilities worldwide from 2014 to 2023. <https://www.statista.com/statistics/1251649/zero-day-vulnerabilities-worldwide/>. Accessed: Aug. 2025 (2025)
- Díaz-Verdejo, J., Muñoz-Calle, J., Estepa Alonso, A., Estepa Alonso, R., Madinabeitia, G.: On the detection capabilities of signature-based intrusion detection systems in the context of web attacks. *Applied Sciences* **12**(2), 852 (2022)
- Khaleel, T.J., Shiltagh, N.A.: Ddos cyber-attacks detection-based hybrid cnn-lstm. In: *International Conference on Computing and Communication Networks*, pp. 523–537 (2023). Springer
- Narmadha, S., Balaji, N.: Improved network anomaly detection system using optimized autoencoder- lstm. *Expert Systems with Applications* **273**, 126854 (2025)
- Hnamte, V., Hussain, J.: Dcnnbilstm: An efficient hybrid deep learning-based intrusion detection system. *Telematics and Informatics Reports* **10**, 100053 (2023)
- Zhang, H.: Development of an intelligent intrusion detection system for iot networks using deep learning. *Discover Internet of Things* **5**(1), 74 (2025)
- Mala, K., Renuka, S., Lavanya, K., YM, M.: Hybrid deep learning approach for efficient intrusion detection using dcnn and bilstm. In: *2025 3rd International Conference on*

- Data Science and Network Security (ICDSNS), pp. 1–6 (2025). IEEE
- Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., Yu, P.S.: Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys* **55**(8), 1–39 (2022)
- ŞAHİN, E., Arslan, N.N., Özdemir, D.: Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications* **37**(2), 859–965 (2025)
- Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H.: Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications* **50**, 102419 (2020) <https://doi.org/10.1016/j.jisa.2019.102419>
- Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C.Y., Taher, F.: Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access* **10**, 93104–93139 (2022)
- Korba, A.A., Boualouache, A., Brik, B., Rahal, R., Ghamri-Doudane, Y., Senouci, S.M.: Federated learning for zero-day attack detection in 5g and beyond v2x networks. In: *ICC 2023-IEEE International Conference on Communications*, pp. 1137–1142 (2023). IEEE
- Lopes, I.O., Zou, D., Abdulqadder, I.H., Akbar, S., Li, Z., Ruambo, F., Pereira, W.: Network intrusion detection based on the temporal convolutional model. *Computers & Security* **135**, 103465 (2023)
- Golchin, P., Rafiee, N., Hajizadeh, M., Khalil, A., Kundel, R., Steinmetz, R.: Sslids: Enhancing generalization of intrusion detection with self-supervised contrastive learning. In: *2024 IFIP Networking Conference (IFIP Networking)*, pp. 404–412 (2024). IEEE
- Ogunseyi, T.B., *et al.*: Performance analysis of explainable deep learning-based intrusion detection systems for IoT networks: A systematic review. *Sensors* **26**(2), 363 (2026) <https://doi.org/10.3390/s26020363>
- Neto, E.C.P., Iqbal, S., Buffett, S., Sultana, M., Taylor, A.: Deep learning for intrusion detection in emerging technologies: A comprehensive survey and new perspectives. *Artificial Intelligence Review* **58**, 340 (2025) <https://doi.org/10.1007/s10462-025-11346-z>
- Arnob, R.I., Islam, M.R., *et al.*: A comprehensive systematic review of intrusion detection systems: Techniques, challenges, and future directions. *Journal of Edge Computing* **4**(1), 73–104 (2025) <https://doi.org/10.55056/jec.885>
- Chinnasamy, P., *et al.*: Deep learning-driven methods for network-based intrusion

- detection systems: A systematic review. *ICT Express* **11**(1), 181–215 (2025) <https://doi.org/10.1016/j.ict.2025.01.005>
- Alansary, S.A., Ayyad, S.M., Talaat, F.M., Saafan, M.M.: Emerging ai threats in cybercrime: a review of zero-day attacks via machine, deep, and federated learning. *Knowledge and Information Systems* **67**(11), 10951–10987 (2025)
- Guo, Y.: A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer communications* **198**, 175–185 (2023) <https://doi.org/10.1016/j.comcom.2022.11.001>
- Nair, D., Mhavan, N.: Augmenting cybersecurity: a survey of intrusion detection systems in combating zero-day vulnerabilities. In: *Smart Analytics, Artificial Intelligence and Sustainable Performance Management in a Global Digitalised Economy*, pp. 129–153. Emerald Publishing Limited, ??? (2023)
- Por, L.Y., Dai, Z., Leem, S.J., Chen, Y., Yang, J., Binbeshr, F., Phan, K.Y., Ku, C.S.: A systematic literature review on the methods and challenges in detecting zero-day attacks: Insights from the recent crowdstrike incident. *IEEE Access* (2024)
- Uttarkar, A.R., Vanjale, S., Rajpoot, P.: A comprehensive review and novel approach for detection of zero day vulnerabilities. In: *2024 IEEE Pune Section International Conference (PuneCon)*, pp. 1–6 (2024). IEEE
- Mohamed, N., Taherdoost, H., Madanchian, M.: Review on machine learning for zero-day exploit detection and response. In: *International Conference on Smart Technology*, pp. 163–176 (2024). Springer
- Silva Ruffo, V.G., Lent, D.M.B., Komarchesqui, M., Schiavon, V.F., Assis, M.V.O., Carvalho, L.F., Proença Jr, M.L.: Anomaly and intrusion detection using deep learning for software-defined networks: A survey. *Expert Systems with Applications* **256**, 124982 (2024)
- Khan, N., Ahmad, K., Al Tamimi, A., Alani, M.M., Bermak, A., Khalil, I.: Explainable ai-based intrusion detection systems for industry 5.0 and adversarial xai: A systematic review. *Information* **16**(12), 1036 (2025)
- Manjula, L., Saravanakumar, E.: Zero-day attack detection in cloud security: A survey of machine learning and deep learning approaches. In: *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 2055–2062 (2025). IEEE
- Otoun, Y., Nayak, A.: As-ids: Anomaly and signature based ids for the internet of things. *Journal of Network and Systems Management* **29**(3), 23 (2021)
- Samita: A review on intrusion detection system for iot based systems. *SN Computer Science* **5**(4), 380 (2024)

- Ravindran, V.K., Ojha, S.S., Kamboj, A.: A comparative analysis of signature-based and anomaly-based intrusion detection systems. *International Journal of Latest Technology in Engineering, Management & Applied Science* **14**(5), 209–214 (2025)
- Khraisat, A., Alazab, A.: A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. *Cybersecurity* **4**(1), 18 (2021)
- Erlacher, F., Dressler, F.: On high-speed flow-based intrusion detection using snort-compatible signatures. *IEEE Transactions on Dependable and Secure Computing* **19**(1), 495–506 (2020)
- Sarhan, M., Layeghy, S., Gallagher, M., Portmann, M.: From zero-shot machine learning to zero-day attack detection. *International Journal of Information Security* **22**(4), 947–959 (2023) <https://doi.org/10.1007/s10207-023-00676-0>
- Bahlali, A.R., Bachir, A., Labed, A.: Self-supervised learning meets custom autoencoder classifier: A semi-supervised approach for encrypted traffic anomaly detection. *IEEE Access* (2025)
- Tong, J., Zhang, Y.: A real-time label-free self-supervised deep learning intrusion detection for handling new type and few-shot attacks in iot networks. *IEEE Internet of Things Journal* **11**(19), 30769–30786 (2024)
- Duraz, R., Espes, D., Francq, J., Vaton, S.: Secl: A zero-day attack detector and classifier based on contrastive learning and strong regularization. In: *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pp. 1–12 (2024)
- Caville, E., Lo, W.W., Layeghy, S., Portmann, M.: Anomal-e: A self-supervised network intrusion detection system based on graph neural networks. *Knowledge-based systems* **258**, 110030 (2022)
- Xu, R., Wu, G., Wang, W., Gao, X., He, A., Zhang, Z.: Applying self-supervised learning to network intrusion detection for network flows with graph neural network. *Computer Networks* **248**, 110495 (2024)
- Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., Venkatraman, S.: Deep learning approach for intelligent intrusion detection system. *IEEE access* **7**, 41525–41550 (2019)
- Shone, N., Ngoc, T.N., Phai, V.D., Shi, Q.: A deep learning approach to network intrusion detection. *IEEE transactions on emerging topics in computational intelligence* **2**(1), 41–50 (2018)
- Ullah, F., Ullah, S., Srivastava, G., Lin, J.C.-W.: Ids-int: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. *Digital*

- Khan, I.A., Razzak, I., Pi, D., Khan, N., Hussain, Y., Li, B., Kousar, T.: Fed-inforce-fusion: A federated reinforcement-based fusion model for security and privacy protection of iomt networks against cyber-attacks. *Information Fusion* **101**, 102002 (2024)
- Shahid, M.R., Blanc, G., Zhang, Z., Debar, H.: Anomalous communications detection in iot networks using sparse autoencoders. In: 2019 IEEE 18th International Symposium on Network Computing and Applications (NCA), pp. 1–5 (2019). IEEE
- Villegas-Ch, W., Govea, J., Navarro, A.M., Játiva, P.P.: Intrusion detection in iot networks using dynamic graph modeling and graph-based neural networks. *IEEE Access* (2025)
- Abdulrazzaq, M.M., Ramaha, N.T., Hameed, A.A., Salman, M., Yon, D.K., Fitriyani, N.L., Syafrudin, M., Lee, S.W.: Consequential advancements of self-supervised learning (ssl) in deep learning contexts. *Mathematics* **12**(5), 758 (2024)
- Vinayak, H., Jarin, T.: A hybrid model for detecting intrusions using stacked autoencoders and extreme gradient boosting. *Computers & Security* **150**, 104212 (2025)
- Walczyna, T., Jankowski, D., Piotrowski, Z.: Enhancing anomaly detection through latent space manipulation in autoencoders: A comparative analysis. *Applied Sciences* **15**(1), 286 (2024)
- Maheshwari, A., Mehraj, B., Khan, M.S., Idrisi, M.S.: An optimized weighted voting based ensemble model for ddos attack detection and mitigation in sdn environment. *Microprocessors and Microsystems* **89**, 104412 (2022)
- Jakka, G., Alsmadi, I.M.: Ensemble models for intrusion detection systemclassification. *International Journal of Smart Sensor and Adhoc Network* **3**(2), 8 (2022)
- Basati, A., Faghih, M.M.: Apae: an iot intrusion detection system using asymmetric parallel auto-encoder. *Neural Computing and Applications* **35**(7), 4813–4833 (2023)
- Yang, L., Song, Y., Gao, S., Hu, A., Xiao, B.: Griffin: Real-time network intrusion detection system via ensemble of autoencoder in sdn. *IEEE Transactions on Network and Service Management* **19**(3), 2269–2281 (2022)
- Adawadkar, A.M.K., Kulkarni, N.: Cyber-security and reinforcement learning—a brief survey. *Engineering Applications of Artificial Intelligence* **114**, 105116 (2022)
- Praveena, V., Vijayaraj, A., Chinnasamy, P., Ali, I., Alroobaea, R., Alyahyan, S.Y., Raza, M.A.: Optimal deep reinforcement learning for intrusion detection in uavs. *Computers, Materials & Continua* **70**(2) (2022)

- Benaddi, H., Ibrahimi, K., Benslimane, A., Jouhari, M., Qadir, J.: Robust enhancement of intrusion detection systems using deep reinforcement learning and stochastic game. *IEEE Transactions on Vehicular Technology* **71**(10), 11089–11102 (2022)
- Benaddi, H., Jouhari, M., Ibrahimi, K., Ben Othman, J., Amhoud, E.M.: Anomaly detection in industrial iot using distributional reinforcement learning and generative adversarial networks. *Sensors* **22**(21), 8085 (2022)
- Yang, J., Li, H., Shao, S., Zou, F., Wu, Y.: Fs-ids: A framework for intrusion detection based on few-shot learning. *Computers & Security* **122**, 102899 (2022)
- Lu, C., Wang, X., Yang, A., Liu, Y., Dong, Z.: A few-shot-based model-agnostic meta-learning for intrusion detection in security of internet of things. *IEEE Internet of Things Journal* **10**(24), 21309–21321 (2023)
- Althiyabi, T., Ahmad, I., Alassafi, M.O.: Enhancing iot security: A few-shot learning approach for intrusion detection. *Mathematics* **12**(7), 1055 (2024)
- Ohtani, T., Yamamoto, R., Ohzahata, S.: Detecting zero-day attack with federated learning using autonomously extracted anomalies in iot. In: 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), pp. 356–359 (2024). IEEE
- Jin, D., Chen, S., He, H., Jiang, X., Cheng, S., Yang, J.: Federated incremental learning based evolvable intrusion detection system for zero-day attacks. *IEEE Network* **37**(1), 125–132 (2023)
- Liu, Y., Wu, L.: Intrusion detection model based on improved transformer. *Applied Sciences* **13**(10), 6251 (2023)
- Halbouni, A., Gunawan, T.S., Habaebi, M.H., Halbouni, M., Kartiwi, M., Ahmad, R.: Cnn-lstm: hybrid deep neural network for network intrusion detection system. *IEEE Access* **10**, 99837–99849 (2022)
- Kim, J., Kim, J., Kim, H., Shim, M., Choi, E.: Cnn-based network intrusion detection against denial-of-service attacks. *Electronics* **9**(6), 916 (2020)
- El Alami, H., Rawat, D.B.: Detection of zero-day attacks using cnn and lstm in networked autonomous systems ‘ieee cns 23 poster’. In: 2023 IEEE Conference on Communications and Network Security (CNS), pp. 1–2 (2023). IEEE
- Gueriani, A., Kheddar, H., Mazari, A.C.: Enhancing iot security with cnn and lstm-based intrusion detection systems. In: 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), pp. 1–7 (2024). IEEE
- Said, R.B., Sabir, Z., Askerzade, I.: Cnn-bilstm: A hybrid deep learning approach for network intrusion detection system in software-defined networking with hybrid

- feature selection. *IEEE Access* **11**, 138732–138747 (2023)
- Hossain, M.A.: Deep learning-based intrusion detection for iot networks: a scalable and efficient approach. *EURASIP Journal on Information Security* **2025**(1), 28 (2025)
- Meliboev, A., Alikhanov, J., Kim, W.: Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. *Electronics* **11**(4), 515 (2022)
- Akuthota, U.C., Bhargava, L.: Transformer based intrusion detection for iot networks. *IEEE Internet of Things Journal* (2025)
- Altaf, T., Wang, X., Ni, W., Yu, G., Liu, R.P., Braun, R.: Gnn-based network traffic analysis for the detection of sequential attacks in iot. *Electronics* **13**(12), 2274 (2024)
- Liu, W., Li, S., Gao, C., Sang, Y., Ge, H.: An adaptive graph neural network-based intrusion detection system for airborne network. *Engineering Applications of Artificial Intelligence* **152**, 110851 (2025)
- Wu, Z., Zhang, H., Wang, P., Sun, Z.: Rtids: A robust transformer-based approach for intrusion detection system. *IEEE Access* **10**, 64375–64387 (2022)
- Manocchio, L.D., Layeghy, S., Lo, W.W., Kulatilleke, G.K., Sarhan, M., Portmann, M.: Flowtransformer: A transformer framework for flow-based network intrusion detection systems. *Expert Systems with Applications* **241**, 122564 (2024)
- Xi, C., Wang, H., Wang, X.: A novel multi-scale network intrusion detection model with transformer. *Scientific Reports* **14**(1), 23239 (2024)
- Fernando, O.A., Spring, J., Xiao, H.: Cinet: A comparative study of a cnn-based intrusion detection system with tensorflow and pytorch for 5g and beyond. *Network*, 1–28 (2025)
- Purwono, P., Ma’arif, A., Rahmaniar, W., Fathurrahman, H.I.K., Frisky, A.Z.K., Haq, Q.M.: Understanding of convolutional neural network (cnn): A review. *International Journal of Robotics and Control Systems* **2**(4), 739–748 (2022)
- Ayeni, J.: Convolutional neural network (cnn): the architecture and applications. *Appl. J. Phys. Sci* **4**(4), 42–50 (2022)
- Shebl, A., Elsedimy, E., Ismail, A., Salama, A., Herajy, M.: Dcnn: a novel binary and multi-class network intrusion detection model via deep convolutional neural network. *EURASIP Journal on Information Security* **2024**(1), 36 (2024)
- Bamber, S.S., Katkuri, A.V.R., Sharma, S., Angurala, M.: A hybrid cnn-lstm approach for intelligent cyber intrusion detection system. *Computers & Security* **148**, 104146 (2025)

- Phalaagae, P., Zungeru, A.M., Yahya, A., Sigweni, B., Rajalakshmi, S.: A hybrid cnn-lstm model with attention mechanism for improved intrusion detection in wireless iot sensor networks. *IEEE Access* (2025)
- Farhanath, K., Senthilkumar, R., *et al.*: Intrusion detection for cyber-physical systems and iot devices using hybrid deep learning algorithm. In: 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 137–142 (2024). IEEE
- Gao, J.: Network intrusion detection method combining cnn and bilstm in cloud computing environment. *Computational intelligence and neuroscience* **2022**(1), 7272479 (2022)
- Kasongo, S.M.: A deep learning technique for intrusion detection system using a recurrent neural networks based framework. *Computer Communications* **199**, 113–125 (2023)
- Yadav, A., Kumar, A.: Intrusion detection and prevention using rnn in wsn. In: *Inventive Computation and Information Technologies: Proceedings of ICICIT 2021*, pp. 531–539. Springer, ??? (2022)
- Amutha, S., Kavitha, R., Srinivasan, R., Kavitha, M.: Secure network intrusion detection system using nid-rnn based deep learning. In: 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pp. 1–5 (2022). IEEE
- Awad, A.A., Ali, A.F., Gaber, T.: An improved long short term memory network for intrusion detection. *Plos one* **18**(8), 0284795 (2023)
- Zhou, H., Kang, L., Pan, H., Wei, G., Feng, Y.: An intrusion detection approach based on incremental long short-term memory. *International Journal of Information Security* **22**(2), 433–446 (2023)
- Aydn, H., Orman, Z., Aydn, M.A.: A long short-term memory (lstm)-based distributed denial of service (ddos) detection and defense system design in public cloud network environment. *Computers & Security* **118**, 102725 (2022)
- Du, J., Yang, K., Hu, Y., Jiang, L.: Nids-cnnlstm: Network intrusion detection classification model based on deep learning. *IEEE Access* **11**, 24808–24821 (2023)
- Abdulmajeed, I.A., Husien, I.M.: Mlids22-ids design by applying hybrid cnn-lstm model on mixed-datasets. *Informatica* **46**(8) (2022)
- Babaey, V., Faragardi, H.R.: Detecting zero-day web attacks with an ensemble of lstm, gru, and stacked autoencoders. *Computers* **14**(6), 205 (2025)
- Rehman, M.U., Zita, M., Abrar, M., Kazim, M., Khalid, S.: Zero-day attack detection

- system using autoencoders and isolation forest: An unsupervised machine learning approach. In: International Conference on Neural Computing for Advanced Applications, pp. 245–258 (2025). Springer
- Zahoora, U., Rajarajan, M., Pan, Z., Khan, A.: Zero-day ransomware attack detection using deep contractive autoencoder and voting based ensemble classifier. *Applied Intelligence* **52**(12), 13941–13960 (2022)
- DAS, A., PRAMOD, S.: An enhanced optimization model with ensemble autoencoder for zero-day attack detection. *Journal of Theoretical and Applied Information Technology* **100**(22) (2022)
- Ambekar, N.G., Thokchom, S.: Ul-vae: An unsupervised learning approach for zero-day malware detection using variational autoencoder. In: 2024 International Conference on Computational Intelligence and Network Systems (CINS), pp. 1–7 (2024). IEEE
- Gupta, K., Chatterjee, N., *et al.*: A gnn-based novel approach to detect malicious traffic in intrusion detection system. *Procedia Computer Science* **258**, 3184–3193 (2025)
- Sun, Z., Teixeira, A.M., Toor, S.: Gnn-ids: Graph neural network based intrusion detection system. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–12 (2024)
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. *AI Open* **1**, 57–81 (2020) <https://doi.org/10.1016/j.aiopen.2021.01.001>
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* **32**(1), 4–24 (2021) <https://doi.org/10.1109/TNNLS.2020.2978386>
- Tung, N.X., Son, B.D., Geun-Jeong, S., Van Chien, T., Hanzo, L., Hwang, W.J., *et al.*: Graph neural networks for next-generation-iot: Recent advances and open challenges. *IEEE Communications Surveys & Tutorials* (2025)
- Long, Z., Yan, H., Shen, G., Zhang, X., He, H., Cheng, L.: A transformer-based network intrusion detection approach for cloud security. *Journal of Cloud Computing* **13**(1), 5 (2024)
- Han, X., Cui, S., Liu, S., Zhang, C., Jiang, B., Lu, Z.: Network intrusion detection based on n-gram frequency and time-aware transformer. *Computers & Security* **128**, 103171 (2023)
- Tseng, S.-M., Wang, Y.-Q., Wang, Y.-C.: Multi-class intrusion detection based on transformer for iot networks using cic-iot-2023 dataset. *Future Internet* **16**(8), 284 (2024)

- Zhou, H., Zou, H., Li, W., Li, D., Kuang, Y.: Hivit-ids: An efficient network intrusion detection method based on vision transformer. *Sensors* **25**(6), 1752 (2025)
- Khan, M.A.: Hcrnnids: Hybrid convolutional recurrent neural network-based network intrusion detection system. *Processes* **9**(5), 834 (2021)
- Sameera, N., Shashi, M.: Deep transductive transfer learning framework for zero-day attack detection. *ICT Express* **6**(4), 361–367 (2020)
- Zhao, J., Shetty, S., Pan, J.W., Kamhoua, C., Kwiat, K.: Transfer learning for detecting unknown network attacks. *EURASIP Journal on Information Security* **2019**(1), 1–13 (2019)
- Xu, B., Wang, B., Chen, X., Zhao, J., He, G.: A cgan-based few-shot method for zero-day attack detection in the internet of vehicles. In: 2023 Eleventh International Conference on Advanced Cloud and Big Data (CBD), pp. 98–103 (2023). IEEE
- Faber, K., Faber, L., Sniezynski, B.: Autoencoder-based ids for cloud and mobile devices. In: 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pp. 728–736 (2021). IEEE
- Ahmad, J., Latif, S., Djenouri, D., Ullah, F., Khan, M.S., Saad, M.M., Jhaveri, R.H., Verma, P.: A stochastic prototypical network for few-shot intrusion detection in can-based iov network. *IEEE Open Journal of the Communications Society* (2025)
- Tariq, A.H.I.E., Tariq, M.B.I.E., Lu, S.: Hybrid ai-driven techniques for enhancing zeroday exploit detection in intrusion detection system (ids). In: 2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC), pp. 156–160 (2024). IEEE
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., *et al.*: The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**, 71 (2021) <https://doi.org/10.1136/bmj.n71>
- Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S., *et al.*: PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews* **10**(1), 39 (2021) <https://doi.org/10.1186/s13643-020-01542-z>
- Soltani, M., Ousat, B., Jafari Siavoshani, M., Jahangir, A.H.: An adaptable deep learning-based intrusion detection system to zero-day attacks. *Journal of Information Security and Applications* **76**, 103516 (2023) <https://doi.org/10.1016/j.jisa.2023.103516>
- Hindy, H., Atkinson, R., Tachtatzis, C., Colin, J.-N., Bayne, E., Bellekens, X.: Utilising deep learning techniques for effective zero-day attack detection. *Electronics* **9**(10) (2020) <https://doi.org/10.3390/electronics9101684>

- Deldar, F., Abadi, M.: Deep learning for zero-day malware detection and classification: A survey. *ACM Computing Surveys* **56**(2), 1–37 (2023) <https://doi.org/10.1145/3617894>
- Zhang, H., Maple, C.: Deep reinforcement learning-based intrusion detection in iot system: a review. *Iet Conference Proceedings* (2023) <https://doi.org/10.1049/icp.2023.2577>
- Papalkar, R.R., Alvi, A.S.: Review of unknown attack detection with deep learning techniques. In: *Artificial Intelligence, Blockchain, Computing and Security Volume 1*. CRC Press, ??? (2023). <https://doi.org/10.1201/9781003393580-148>
- Lal, P.S., Ashok, A.: Deep learning-based network intrusion detection systems: A comprehensive survey of methods, evolution, and research perspectives. In: *2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*. IEEE, ??? (2026). <https://doi.org/10.1109/icmcsi67283.2026.11412845>
- Siam, A.A., Faruqui, N., Azad, A., Moni, M.A.: Securing the unseen: A comprehensive exploration review of ai -powered models for zero-day attack detection. *Expert Systems* (2026) <https://doi.org/10.1111/exsy.70217>
- Tayyab, U.-E.-H., Khan, F.B., Durad, M.H., Khan, A., Lee, Y.S.: A survey of the recent trends in deep learning based malware detection. *Journal of Cybersecurity and Privacy* (2022) <https://doi.org/10.3390/jcp2040041>
- Kritika, E.: A comprehensive literature review on phishing url detection using deep learning techniques. *Journal of Cyber Security Technology* (2024) <https://doi.org/10.1080/23742917.2024.2378552>
- B, S.C., G, P.K., K, A.B.M., Aswathy, A.A., Gracia, C.D.: Machine learning in cyber security: A comprehensive review. In: *2025 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, ??? (2025). <https://doi.org/10.1109/accthpa65749.2025.11168602>
- Mote, M.S.A., Arabagatte, D.S.: Survey on detection and mitigation strategies for zero-day attacks: An extensive analysis of current developments and case studies. In: *First International Conference on Computer, Computation and Communication (IC3C-2025)*. River Publishers, ??? (2025). <https://doi.org/10.13052/rp-9788743808268a037>
- Won, D.-O., Jang, Y.-N., Lee, S.-W.: Plausmal-gan: Plausible malware training based on generative adversarial networks for analogous zero-day malware detection. *Ieee Transactions on Emerging Topics in Computing* (2023) <https://doi.org/10.1109/tetc.2022.3170544>
- Li, Z., Zhao, D.: Zerod-fender: A resource-aware iot malware detection engine via fine-grained side-channel analysis. *Acm Transactions on Design Automation of*

- Electronic Systems (2024) <https://doi.org/10.1145/3687482>
- Houda, Z.A.E., Moudoud, H., Brik, B.: Federated deep reinforcement learning for efficient jamming attack mitigation in o-ran. *Ieee Transactions on Vehicular Technology* (2024) <https://doi.org/10.1109/tvt.2024.3359998>
- Mahida, A.: An intellectual zero trust security framework using deep reinforcement learning for predictive threat mitigation in ai-based fraud detection systems. *Ieee Access* (2026) <https://doi.org/10.1109/access.2026.3664389>
- Zahoor, U., Khan, A., Rajarajan, M., Khan, S.H., Asam, M., Jamal, T.: Ransomware detection using deep learning based unsupervised feature extraction and a cost sensitive pareto ensemble classifier. *Scientific Reports* (2022) <https://doi.org/10.1038/s41598-022-19443-7>
- Okey, O.D., Dadkhah, S., Rodríguez, D.Z., Kleinschmidt, J.H.: Zero-day device identification for iot security through optimized incremental learning and drift adaptation. In: *Proceedings of the 2025 IEEE SIoT Conference*. IEEE, ??? (2025). <https://doi.org/10.1109/siot68426.2025.11368834>
- Wahid, A.R., Anggreani, D., Hayat, M.A.M., Rahman, A.A., Faisal, M.: Stacking architecture-endpoint detection: a hybrid multi layered architecture for endpoint threat detection. *International Journal of Advances in Applied Sciences* (2025) <https://doi.org/10.11591/ijaas.v14.i4.pp1263-1280>
- Elijah, T.D., Walee, N.A., Shalan, A.M.: Ai-based detection of zero-day exploits: A framework. In: *Proceedings of the 2025 IEEE ICICIS Conference*. IEEE, ??? (2025). <https://doi.org/10.1109/icicis66182.2025.11313144>
- Ali, S., Rehman, S.U., Imran, A., Adeem, G., Iqbal, Z., Kim, K.-I.: Comparative evaluation of ai-based techniques for zero-day attacks detection. *Electronics* **11**(23), 3934 (2022)
- Anwer, M., Ahmed, G., Akhunzada, A., Hussain, S., Khan, M.: Comparative analysis of soft computing approaches of zero-day-attack detection. In: *2022 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, ??? (2022). <https://doi.org/10.1109/icetst55735.2022.9922937>
- Naithani, A., Singh, S.N., Sharma, P.K.: Zero-day vigilance: Proposing a machine learning framework for ddos detection in cloud computing. In: *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*. IEEE, ??? (2024). <https://doi.org/10.1109/globalaisummit62156.2024.10947871>
- Kara, A.O., Kara, M., Boyaci, A.: A comparative analysis of machine learning and deep reinforcement learning approaches for adaptive intrusion detection. *Ieee Access* (2025) <https://doi.org/10.1109/access.2025.3627098>

- Rodríguez, E., Valls, P., Otero, B., Costa, J.J., Verdú, J., Pajuelo, M.A., Canal, R.: Transfer-learning-based intrusion detection framework in iot networks. *Sensors* (2022) <https://doi.org/10.3390/s22155621>
- Wang, J., Huang, N., Zhang, H., Liu, L., Fu, Q., Cao, K., Guo, X., Jung, H.: Self-learning model fusion for network anomaly detection: A hybrid cnn-lstm-transformer framework. *Plos One* (2025) <https://doi.org/10.1371/journal.pone.0332502>
- Jiang, X., Xu, L., Yu, L., Fang, X.: Mft: A novel memory flow transformer efficient intrusion detection method. *Computers & Security* **148**, 104174 (2025)
- Bensaoud, A., Kalita, J.: Optimized detection of cyber-attacks on iot networks via hybrid deep learning models. *Ad Hoc Networks* **170**, 103770 (2025)
- Sinha, P., Sahu, D., Prakash, S., Yang, T., Rathore, R.S., Pandey, V.K.: A high performance hybrid lstm cnn secure architecture for iot environments using deep learning. *Scientific Reports* **15**(1), 9684 (2025)
- Shang, F., Mao, L., Gong, W.: Service-aware adaptive link load balancing mechanism for software-defined networking. *Future Generation Computer Systems* **81**, 452–464 (2018)
- Ranpara, R., Hirpara, J., Doshi, M., Shah, D.: A novel bayesian-inspired framework for proactive detection and mitigation of zero-day attacks in distributed network architectures. *Iet Conference Proceedings* (2025) <https://doi.org/10.1049/icp.2025.1567>
- Sunkara, K.C., Mittal, T., Sripada, N.K., K, B.H., Tlajiya, F., Thakur, V.S.: Probabilistic deep learning algorithmic model for robust anomaly detection in decentralized networked cyber-physical systems. In: 2025 3rd International Conference on Data Science and Information System (ICDSIS). IEEE, ??? (2025). <https://doi.org/10.1109/icdsis65355.2025.11070339>
- Saheed, Y.K., Chukwuere, J.E.: Autonomous llm agent: A memory-augmented, edge-optimized shap explanations with zero-day attack resilience in iot/industrial iot networks. *Ieee Internet of Things Journal* (2025) <https://doi.org/10.1109/jiot.2025.3648649>
- Nagarjun, A.V., Rajkumar, S.: Quantum deep learning-enhanced ethereum blockchain for cloud security: intrusion detection, fraud prevention, and secure data migration. *Scientific Reports* (2025) <https://doi.org/10.1038/s41598-025-22408-1>
- Li, K., Ma, W., Duan, H., Xie, H., Zhu, J.: Few-shot iot attack detection based on rfp-cnn and adversarial unsupervised domain-adaptive regularization. *Computers & Security* **121**, 102856 (2022)

- Mhawi, D.N., Aldallal, A., Hassan, S.: Advanced feature-selection-based hybrid ensemble learning algorithms for network intrusion detection systems. *Symmetry* **14**(7), 1461 (2022)
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45–74 (2024)
- Annapareddy, V.N., Challa, K., Komaragiri, V.B., Sriram, H.K., Kalisetty, S., Goma, T.: Explaining ai techniques such as shap lime and rise in limited sample size neuroimaging studies. In: 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0, pp. 1–6 (2025). IEEE
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al.*: Attention is all you need [j]. *Advances in neural information processing systems* **30**(1), 261–272 (2017)
- Khan, I.A., Pi, D., Kamal, S., Alsuhaibani, M., Alshammari, B.M.: Federated-boosting: a distributed and dynamic boosting-powered cyber-attack detection scheme for security and privacy of consumer iot. *IEEE Transactions on Consumer Electronics* (2024)
- Dai, Z., Por, L.Y., Chen, Y.-L., Yang, J., Ku, C.S., Alizadehsani, R., Pławiak, P.: An intrusion detection model to detect zero-day attacks in unseen data using machine learning. *PloS one* **19**(9), 0308469 (2024)
- Dardouri, S., Almuhanha, R.: A deep learning/machine learning approach for anomaly based network intrusion detection. *Frontiers in Artificial Intelligence* **8**, 1625891 (2025)
- Akshaya, S., *et al.*: Resnet50-based deep convolutional neural network for zero-day attack prediction and detection. *International Journal of Advanced Technology and Engineering Exploration* **12**(124), 507 (2025)
- Saurabh, K., Singh, U., Mishra, A., Vyas, R., Vyas, O.: Zero-shot based hybrid neural network system for enhancing zero-day attack detection. In: 2024 IEEE 21st India Council International Conference (INDICON), pp. 1–6 (2024). IEEE
- Lilhore, U.K., Simaiya, S., Alroobaea, R., Baqasah, A.M., Alsafyani, M., Alhazmi, A., Khan, M.M.: Smarttrust: a hybrid deep learning framework for real-time threat detection in cloud environments using zero-trust architecture. *Journal of Cloud Computing* **14**(1), 35 (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MAGIReviewpaperSupplementary.pdf](#)