SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Learned Zero-Day Intrusion Detection: A Systematic Review of Methods, Evaluation Practice and Open Problems, 2020–2026

Rohit Hirvey[1], Dr. Hemant Kumar Gupta[2]

[1]Department of Computer Science and Engineering, SAGE University, Indore, India. E-mail: hirvey.rohit@gmail.com
[2]Department of Computer Science and Engineering, SAGE University, Indore, India. E-mail: hemugupta3131@gmail.com

Abstract
Zero-day intrusions are, by construction, attacks for which no signature, patch or prior labelled example exists, so the defensive burden falls on models that generalise from known behaviour to unknown behaviour. The resulting literature has grown quickly and unevenly, and it is now difficult to see which architectural families are actually distinct, which evaluation practices are shared, and which reported gains are attributable to method rather than to protocol. This article systematically reviews forty-seven studies published between 2020 and 2026, drawn from a curated domain corpus of fifty-one records after four duplicate entries were removed. Each study is coded along a common extraction schema covering methodological family, learning paradigm, target deployment environment, evaluation protocol and the limitation its own authors acknowledge. Six families are identified and characterised: classical and ensemble pipelines, deep and hybrid architectures, outlier and semi-supervised formulations, reinforcement-learning agents, representation-learning front-ends, and an emerging group combining large language models, neuro-symbolic reasoning, temporal graphs, moving-target defence and federated learning. The profile of the corpus shows a decisive shift, with the emerging family absent before 2025 and accounting for nine of the seventeen studies dated 2026. The synthesis identifies five recurring weaknesses in evaluation practice: unverified corpus separability, inconsistent leakage control, near-absent calibration and cost reporting, single-run experimentation without significance testing, and the almost complete absence of cross-corpus transfer evidence. These are formulated as an agenda of open problems rather than as a ranking of methods, since the reported figures across the corpus are not commensurable.

Keyword: Anomaly detection, autoencoders, deep learning, evaluation methodology, intrusion detection systems, large language models, network security, reinforcement learning, semi-supervised learning, systematic review, zero-day attack detection.

1. Introduction

A zero-day exploit is an attack for which no signature, patch or prior labelled example exists at the moment it is used. Signature matching therefore fails against it in principle rather than by accident, and the defensive burden shifts onto detectors that must generalise from known threat behaviour to behaviour they have never observed [2], [4], [5]. Over the period covered by this review the response to that problem has produced a large and fast-moving literature, spanning feature-engineered classical pipelines [6], [7], ensemble learners tuned by hyperparameter search [9], convolutional and recurrent networks [11], [12], [17], outlier and semi-supervised formulations that model normality rather than attack [2], [4], [23], reinforcement-learning agents that treat detection as sequential decision-making [3], [8], [19], and, most recently, language-model agents and neuro-symbolic reasoners [26], [27], [32], [39].

That growth has outpaced the consolidation of evaluation practice. Reported figures are frequently near the ceiling of the metric, are drawn from single runs on curated corpora, and are rarely accompanied by calibration, cost or transfer evidence. Two recent reviews [42], [46] survey the space descriptively, but neither audits how the surveyed results were produced. The consequence is that a reader cannot presently separate architectural progress from protocol optimism, and cannot tell whether a headline improvement would survive a change of corpus.

This article addresses that gap. It reviews forty-seven studies published between 2020 and 2026 and codes each along a common extraction schema, so the field is described in terms of what was done and how it was measured rather than in terms of numbers that are not commensurable across papers.

1.1 Review Questions

The synthesis is organised around five questions.

RQ1. What methodological families are present in the literature on learned zero-day detection, and how are the studies distributed across them?

RQ2. How has that distribution shifted over the period 2020–2026?

RQ3. What deployment settings and threat channels do these detectors target?
RQ4. What evaluation practice is shared across the corpus, and what is systematically omitted?
RQ5. Which open problems follow from the answers to RQ1–RQ4, and which of them are already partially addressed?

1.2 Contributions

C1. A reproducible selection and extraction protocol is stated in full, including the criteria of Table 1 and the flow of Fig 1, so that the corpus can be reconstructed and extended.
C2. A six-family taxonomy is derived from the corpus rather than imposed on it, and every included study is assigned to exactly one family (Fig 2, Table 2).
C3. Five study-level tables record, for each work, its approach, target setting and the limitation that bears on zero-day generalisation (Tables 3–VII).
C4. Evaluation practice is audited as a first-class object rather than treated as background (Table 8).
C5. Nine open problems are formulated and mapped to the studies that partially address them (Table 9).
The review deliberately does not rank the surveyed methods by reported accuracy. Section 2.4 explains why such a ranking would not be sound on this corpus.

1.3 Organisation

Section 2 states the review protocol. Section 3 profiles the corpus. Section 4 presents the taxonomy and the thematic synthesis. Section 5 audits evaluation practice. Section 6 states the open problems, and Section 7 concludes.

2. Review Methodology

2.1 Corpus Assembly

The review draws on a curated domain corpus of fifty-one bibliographic records covering learned intrusion and zero-day detection published between 2020 and 2026. The records span archival journals, IEEE conference proceedings and two preprints, and include both primary studies and secondary reviews. Four records were found to be exact duplicates of other entries in the same corpus and were removed, leaving forty-seven unique studies. Fig 1 reports the flow.

2.2 Eligibility Criteria

Table 1 states the inclusion and exclusion criteria applied at full-text assessment. The criteria are deliberately permissive on venue and on threat channel, because narrowing either would have concealed exactly the diversification that Section 3 reports; they are strict only on the requirement that a study contribute a learned detection or evaluation artefact.

Table 1. Inclusion and exclusion criteria applied at full-text assessment.

Criterion	Decision
Published 2020–2026	Include
Contributes a learned detector, a defence architecture with a learned component, or a review of such work	Include
Targets intrusion, zero-day, novelty, malware or phishing detection	Include
Peer-reviewed venue or archival preprint	Include
Duplicate bibliographic entry	Exclude
Purely cryptographic or purely policy contribution with no learned component	Exclude
No retrievable description of method or evaluation	Exclude

2.3 Data Extraction

Each included study was coded along six fields: methodological family; principal technique; target deployment environment or threat channel; publication year and venue type; the evaluation protocol as far as it is described; and the limitation most relevant to zero-day generalisation. Family assignment is exclusive, so the counts in Table 2 sum to the corpus size. Where a study spans two families, it was assigned to the one carrying its principal claim; the alternative assignment is noted in the corresponding study table.

2.4 Why No Quantitative Meta-Analysis

A pooled comparison of the reported accuracies was considered and rejected. The studies use different corpora, different partitioning rules, different class granularities and different metric definitions, and a substantial fraction report a single run without dispersion. Pooling such figures would produce a ranking whose ordering is driven by corpus difficulty rather than by method quality, and would give that ordering a precision the underlying evidence does not support. The synthesis is therefore qualitative, and the quantitative claims made here concern the corpus itself — how many studies fall in each family, in each year, and report each evaluation dimension — which are facts about the literature rather than about the detectors.

2.5 Threats to the Validity of This Review

Three limitations apply. The corpus is curated rather than assembled by an exhaustive database query, so prevalence figures describe this corpus rather than the whole field. Coding was performed without an independent second coder, so family assignment at the boundaries — particularly between deep hybrids and representation-learning front-ends — carries a subjective element. And extraction depth is limited by what each study reports: an omission in Table 8 means the dimension was not evident as reported, not that the authors did not consider it.

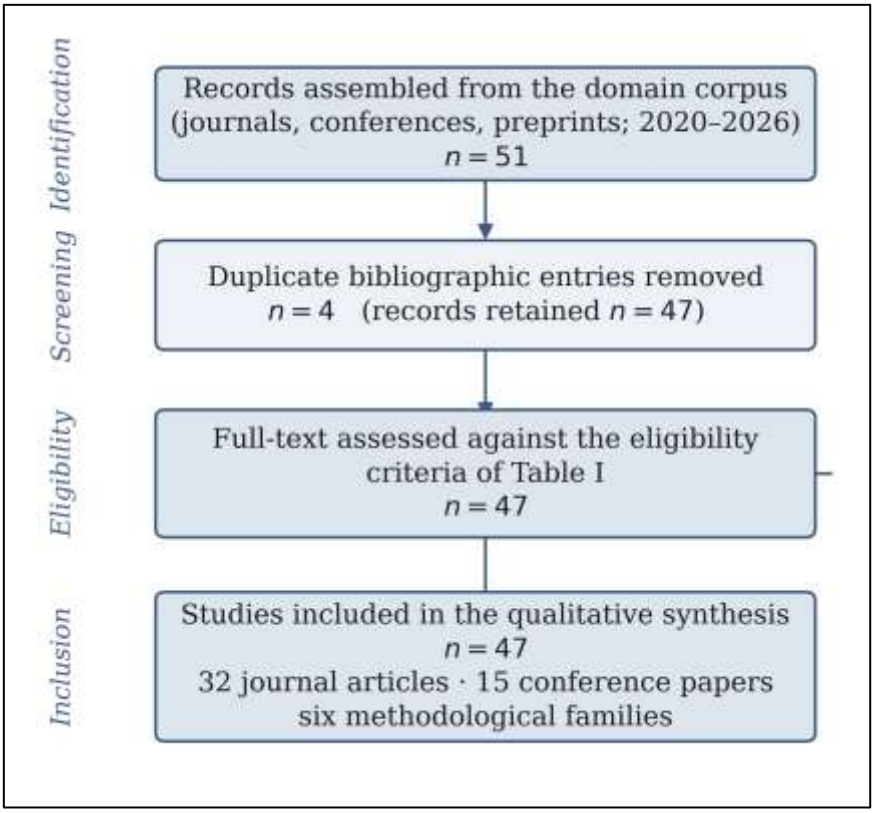

Fig 1. Study selection flow. The corpus is curated rather than exhaustive, so the counts describe the reviewed body of work and not the field as a whole.

3. Profile of the Reviewed Corpus

Fig 3 and Table 2 describe the corpus along its two most informative axes, methodological family and publication year. Three observations follow directly.

First, the field has diversified rather than converged. The two families that dominate the early period, classical or ensemble pipelines and deep or hybrid architectures, account for twenty of the forty-seven studies but for none of the growth after 2025. Second, the emerging family is entirely a phenomenon of the final year of the period: it is absent before 2025 and supplies nine of the seventeen studies dated 2026, covering language-model agents [26], [27], [30], temporal-graph and neuro-symbolic reasoning [28], [32], [39], moving-target defence [38], federated learning under differential privacy [41] and explainable adaptive pipelines [43]. Third, the outlier and semi-supervised family is the only one present in every period of the corpus, which is consistent with its being the formulation most directly matched to the zero-day setting: it does not require a labelled example of the attack it is meant to catch.

The venue split reinforces the reading. Thirty-two studies appeared in journals and fifteen in conference proceedings, but the conference share is concentrated in 2026, where early-stage work in the newest paradigms is being reported at a stage where protocol detail is necessarily compressed.

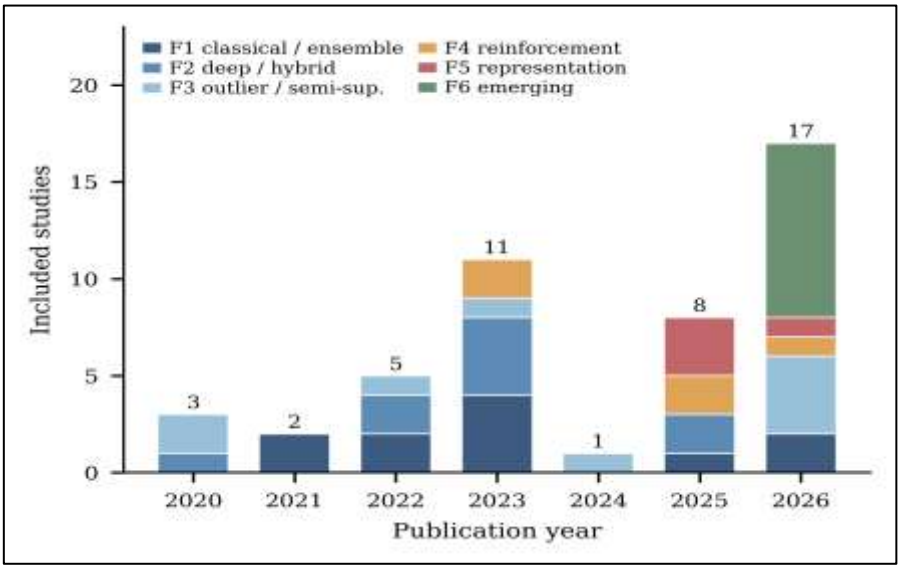

Fig 3. Methodological family of the included studies by publication year. Family assignment is exclusive, so the bars sum to the forty-seven included studies.

Table 2. Distribution of the corpus by methodological family and period.

Family	2020–22	2023–24	2025–26	Total
F1 Classical and ensemble	4	4	3	11
F2 Deep and hybrid	3	4	2	9
F3 Outlier and semi-supervised	3	2	4	9
F4 Reinforcement learning	0	2	3	5
F5 Representation learning	0	0	4	4
F6 Emerging paradigms	0	0	9	9
Total	10	12	25	47

4. Taxonomy and Thematic Synthesis

Fig 2 states the taxonomy. The six families are distinguished by where each approach places its inductive commitment: on engineered features, on architectural depth, on a model of normality, on a decision policy, on the representation itself, or on an external reasoning system. The cross-cutting concerns at the foot of the figure recur in every family and are treated in Section V.

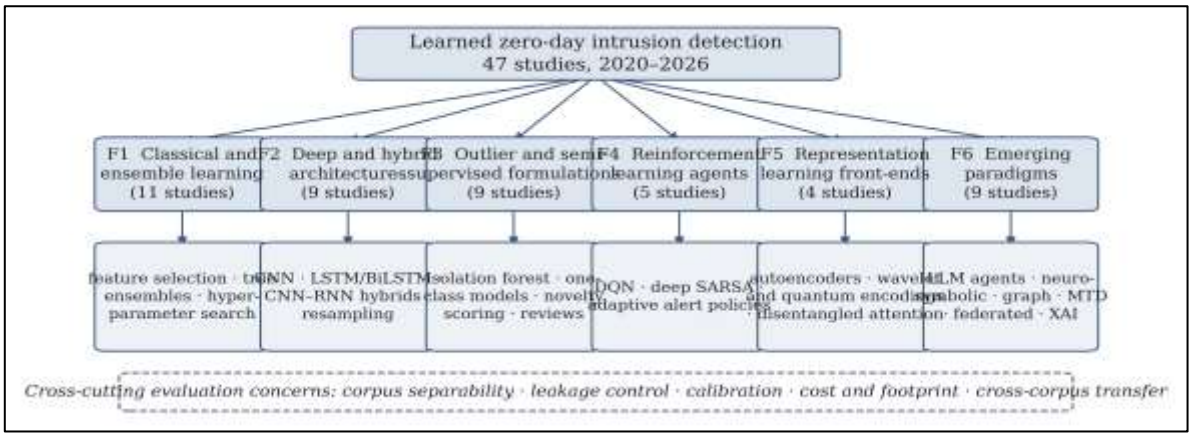

Fig 2. Taxonomy of learned zero-day detection derived from the reviewed corpus. Family assignment is exclusive; the dashed band lists the evaluation concerns that cut across all six families and that Section 5 audits.

4.1 Classical, Ensemble and Feature-Selection Pipelines

Eleven studies build detectors from engineered features and classical learners, and their common thesis is that careful feature construction and selection matters more than model capacity. Feature extraction pipelines for constrained IoT deployments [6], [20], comparative studies across the standard benchmarks [7], and selection-driven designs [10] all report that a well-chosen subset of attributes recovers most of the achievable performance at a fraction of the cost. Ensemble learning with hyperparameter optimisation [9] and the more recent combination of oversampling, LASSO selection and Optuna search [44] extend the same logic to the tuning process itself.

The limitation the family shares is structural, and Table 3 records it study by study: these pipelines are trained and evaluated on closed label sets, so a model that separates a fixed set of known attack classes has no mechanism, beyond the residual behaviour of its decision boundary, for a class it has never seen. Two studies approach the problem from the deployment side instead, tiering the detector for ground control station and UAV links [31] and blending learned detection into a wider cyber-physical defence [33]; both improve operational fit without altering the closed-set assumption. The survey of host-based techniques [15] documents the same pattern in an earlier literature.

Table 3. Classical, ensemble and feature-selection studies (family F1).

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[6] 2023	Feature extraction combined with several supervised learners — IoT network traffic	Attack classes are known at training time; no unseen-family protocol
[7] 2023	Comparative study of classical learners — UNSW-NB15 and NSL-KDD	Benchmark-bound comparison; no transfer between the two corpora
[9] 2023	Ensemble learning with hyperparameter optimisation — General network intrusion	Gains reported on a closed label set; tuning cost not separated
[10] 2023	Feature-selection-driven detector design — General network intrusion	Selection criteria evaluated on the full corpus rather than a training split
[14] 2021	Anomaly-detection framework with a classical learner — CIC-IDS2017	Single-corpus evidence; anomaly threshold not calibrated
[15] 2021	Survey of machine-learning host-based detection techniques — Host telemetry	Descriptive scope; no unified re-evaluation of the surveyed methods
[16] 2022	Classical classification of web intrusion patterns — Web application traffic	Application-layer scope; zero-day generalisation not tested
[20] 2022	Comparison of supervised models for constrained devices — Internet of Things	Cost and footprint discussed informally; no unseen-family evaluation
[31] 2026	Multi-tier real-time detection pipeline — Ground control station and UAV links	Tiering raises latency variance; unseen-attack behaviour not isolated
[33] 2026	Cyber-physical defences combined with learned botnet detection — Consumer and industrial IoT	Blended system; the learned component is not ablated
[44] 2026	Oversampling with LASSO selection and Optuna search — Zero-day attack detection	Optimisation performed on the evaluation partition in the reported design

4.2 Deep and Hybrid Architectures

Nine studies replace engineered features with learned ones. Convolutional models are applied to flow representations in software-defined networks [11], [21], recurrent and bidirectional recurrent layers are stacked on convolutional front-ends to capture order-dependent structure [12], and the two are combined explicitly for unseen-attack detection in networked autonomous systems [17]. Class imbalance, which is severe in intrusion corpora, is addressed by resampling before deep classification [13]. Two later studies extend the family in different directions: lightweight detection of tunnelling behaviour in DNS traffic [29], and explainable deep detection trained on synthetic data for next-generation network infrastructure [34]. Adversarial robustness is treated as a separate axis, with genetic optimisation used to harden deep detectors against white-box and backdoor attacks [18].

Table 4 records the family. Two limitations recur. Depth is bought at an inference cost that is rarely reported alongside the accuracy it purchases, which matters because these detectors are proposed for line-

rate deployment. And the sequential machinery of the recurrent layers is frequently applied to tabular flow records whose feature ordering is arbitrary, so the temporal interpretation of what the model learns does not hold; the architecture may still work, but not for the stated reason.

Table 4. Deep and hybrid architecture studies (family F2).

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[11] 2023	Deep detector for software-defined IoT networks — SDN-enabled IoT	Controller-level assumptions limit portability
[12] 2023	Hybrid convolutional and bidirectional recurrent network — General network intrusion	Depth increases inference cost; calibration not reported
[13] 2023	Data resampling combined with deep classification — Imbalanced intrusion corpora	Resampling applied before partitioning risks optimistic estimates
[17] 2023	Convolutional and recurrent layers for unseen-attack detection — Networked autonomous systems	Short-format study; protocol details insufficient for replication
[18] 2020	Genetic optimisation of deep detectors under adversarial attack — Adversarial and backdoor settings	Robustness is measured against crafted attacks, not novel families
[21] 2022	Convolutional detection in software-defined networks — SDN traffic	Flow representation fixed a priori; no novelty formulation
[22] 2022	Deep classification of malicious activity and connections — Enterprise network security	Supervised throughout; unknown classes fall outside the label set
[29] 2025	Lightweight accurate detection of tunnelling behaviour — DNS tunnels	Protocol-specific; generalisation to other exploit channels untested
[34] 2025	Explainable deep detection trained with synthetic data — NextG network infrastructure	Synthetic augmentation may not preserve the novelty structure of real traffic

4.3 Outlier, Semi-Supervised and Novelty Formulations

Nine studies take the formulation that matches the zero-day problem most directly: model what normal traffic looks like, and treat departures from it as candidate attacks. Outlier-based deep learning [2] and its companion study on deep techniques for zero-day detection [5] establish the pattern; semi-supervised learning from partially labelled traffic [4] relaxes the requirement for exhaustive labels; and a supervised classification filter placed in front of a hybrid anomaly detector [23] shows that the two paradigms can be composed. More recent entries compare isolation forests against autoencoders under a common task [40], apply isolation forests to evil-twin-assisted attacks [45], and assess vulnerability exposure rather than traffic [24]. Two reviews cover the unsupervised sub-literature [42] and AI-powered zero-day models more broadly [46].

The family's difficulty, recorded in Table 5, is the threshold. A novelty score is only as useful as the operating point chosen for it, and that choice is where benign false positives are traded against missed attacks. Few studies in this group characterise the sensitivity of their results to that threshold, and fewer still fit it on a partition held out from the one used to report performance. The assumption of a clean benign pool for training is a second and equally consequential commitment, since operational traffic is not reliably attack-free.

Table 5. Outlier, semi-supervised and novelty-based studies (family F3).

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[2] 2020	Outlier-based deep learning for unseen attacks — General network intrusion	Preprint; threshold sensitivity not characterised
[4] 2022	Semi-supervised learning from partially labelled traffic — Enterprise traffic	Assumes a clean benign pool, which is rarely available operationally
[5] 2020	Deep learning applied to zero-day attack detection — General network intrusion	Evaluated on curated corpora with high class separability

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[23] 2023	Supervised classification filter in front of hybrid anomaly detection — Flow-based detection	Filter inherits the blind spots of its supervised stage
[24] 2024	Machine-learning assessment of zero-day vulnerability exposure — Vulnerability triage	Vulnerability-level rather than traffic-level evidence
[40] 2026	Comparison of isolation forest and autoencoder detectors — Proactive zero-day detection	Two-model comparison; no shared tuning budget reported
[42] 2026	Review of unsupervised techniques for unseen attacks — Cross-domain	Narrative review; no re-evaluation under a common protocol
[45] 2026	Isolation-forest detection of evil-twin-assisted attacks — Wireless network traffic	Single attack scenario; benign false-positive cost not quantified
[46] 2026	Comprehensive review of AI-powered zero-day models — Cross-domain	Breadth-oriented; evaluation practice not audited

4.4 Reinforcement-Learning Agents

Five studies formulate detection as sequential decision-making, in which an agent receives traffic observations and is rewarded for correct classification or for a well-timed alert. Deep Q-learning is the dominant instrument, appearing in a self-learning detection system [3] and in a self-adaptive formulation aimed explicitly at unseen attacks [25]; an on-policy alternative is explored with deep SARSA [8]; and two further studies apply the paradigm to general network detection [19] and to zero-day detection with classical baselines [47].

The appeal is adaptivity: a policy can in principle be updated as the threat surface moves, without retraining a classifier from scratch. The recurring objection, recorded in Table 6, is that the environment is usually a static corpus replayed as a sequence, so the Markov structure the agent exploits is an artefact of the replay order rather than a property of the traffic. Where the reward is a fixed positive or negative scalar, the agent also converges to a single operating point and the resulting scores are poorly calibrated, which limits the use of these detectors in any policy that thresholds on confidence.

4.5 Representation-Learning Front-Ends

Four studies place the inductive commitment in the representation rather than the classifier. An adaptive wavelet and projection autoencoder feeding a hybrid detection network [1] is the most developed instance and is the closest antecedent of much of the recent hybrid work. The remaining three explore alternative encodings: genetic evolution for discovering threats without prior knowledge [35], a comparative study of quantum-classical encodings [36], and disentangled attention for phishing detection from URL structure [37].

This family is the newest apart from the emerging group, with all four studies dated 2025 or later, and it is the one where component-level evidence is weakest. Compression, denoising and projection stages are typically presented as a pipeline whose contributions are asserted rather than isolated, so it is not established which stage carries the benefit. Table 6 records both families together.

Table 6. Reinforcement-learning agents and representation-learning front-ends (families F4 and F5).

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[3] 2025	Deep Q-learning detection agent with adaptive policy — Adaptive cybersecurity	Reward shaping fixes the operating point; calibration is poor
[8] 2023	Deep SARSA agent for anomaly-based detection — General network intrusion	On-policy training is sample-inefficient at corpus scale
[19] 2023	Reinforcement-learning formulation of network detection — General network intrusion	Episode construction from static traffic is only nominally sequential
[25] 2025	Self-adaptive deep Q-network for unseen attacks — Zero-day detection	Adaptation evaluated within a single corpus

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[47] 2026	Deep reinforcement learning for zero-day detection — Zero-day detection	Comparison pool limited to classical baselines
[1] 2025	Adaptive wavelet and projection autoencoder with a hybrid detection network — Zero-day exploit detection	Component contributions asserted rather than isolated by ablation
[35] 2025	Genetic evolution for discovering threats without prior knowledge — Unknown cyber threats	Search cost grows with the representation dimension
[36] 2025	Comparative study of quantum-classical encodings — Network intrusion detection	Encoding advantage is reported under simulation, not deployment
[37] 2026	Disentangled attention for content-aware phishing detection — URL and phishing detection	Single threat channel; not a general traffic detector

4.6 Emerging Paradigms

Nine studies, all dated 2026 or in press, move the reasoning outside the detector. Language-model agents are used to reason over device logs [27] and, with memory augmentation and attribution, over edge and industrial IoT telemetry [26]; a further study couples a language model to digital twins for moving-target defence in real-time cyber-physical systems [30]. Symbolic and causal machinery appears in a neuro-symbolic framework for consumer electronics [32] and in a self-evolving verifiable reasoning architecture for zero-trust ecosystems [39]. Contrastive learning over temporal graphs targets advanced persistent threats with explanations attached [28]. The remaining three address orthogonal concerns: moving-target defence with cyber deception against lateral movement [38], federated threat intelligence under differential privacy [41], and explainable adaptive deep learning for high-speed links [43].

Table 7 records the family. These systems are attractive precisely because they can incorporate knowledge absent from the training corpus, which is the resource a zero-day detector most obviously lacks. But their cost profile differs qualitatively from the earlier families: a language-model call is orders of magnitude more expensive than a forward pass through a small convolutional network, and no study in this group budgets that cost against the arrival rate of the traffic it inspects. Explanation quality is likewise claimed more often than measured.

Table 7. Emerging paradigms: language models, neuro-symbolic reasoning, graphs, moving-target defence and federation (family F6).

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[26] 2026	Memory-augmented language-model agent with attribution — Edge IoT and industrial IoT	Agent latency and prompt cost are not budgeted against line-rate traffic
[27] 2026	Language-model reasoning over device logs — IoT threat detection	Log-level reasoning depends on the completeness of instrumentation
[28] —	Contrastive learning over temporal graphs with explanations — Advanced persistent threats	Graph construction cost; explanation fidelity not measured
[30] 2026	Digital twins with language-model moving-target defence — Real-time cyber-physical systems	Twin fidelity dominates the reported gains
[32] 2026	Neuro-symbolic proactive security reasoning — Consumer electronics	Symbolic rules must be authored and maintained per domain
[38] 2026	Moving-target defence combined with cyber deception — Industrial IoT lateral movement	Defence-side evaluation; detection quality is secondary
[39] —	Self-evolving causal and verifiable reasoning architecture — Zero-trust ecosystems	Verifiability claims are architectural, not empirically bounded

Ref. / yr	Approach and target setting	Limitation bearing on zero-day generalisation
[41] 2026	Federated threat intelligence with differential privacy — UNSW-NB15 federation	Privacy budget and detection quality are not jointly reported
[43] 2026	Explainable adaptive deep learning for high-speed data — High-throughput links	Adaptivity and explanation are evaluated separately

5. Evaluation Practice across the Corpus

The most consequential finding of this review concerns not the detectors but the protocols under which they are assessed. Table 8 summarises seven dimensions of evaluation practice, what the corpus commonly reports along each, and what it commonly omits.

Benchmark choice is narrow and stable. Where a corpus is named it is usually one of a small set of established captures: UNSW-NB15 and NSL-KDD are compared directly in [7], UNSW-NB15 recurs in the federated setting of [41], and CIC-IDS2017 anchors [14]. Several are old relative to the threat behaviour they represent, and a detector that separates their classes cleanly has demonstrated nothing about a genuinely novel family. The complementary risk is saturation: on a highly separable corpus many models reach the metric ceiling and the benchmark loses the ability to rank them at all.

Leakage control is the dimension most often left implicit. A leakage-controlled pipeline requires partitioning to precede every fitted transform, with scaling statistics, selection criteria and calibration parameters estimated on the training partition alone. Selecting features on the full corpus [10], or resampling before partitioning [13], departs from that requirement in ways that inflate the reported estimate. Neither study is unusual; the point is that the practice is rarely stated either way, so a reader cannot distinguish a controlled pipeline from an uncontrolled one.

Calibration and cost are the two dimensions most often absent. Discrimination metrics dominate reporting, yet a detector deployed behind a confidence-thresholded alerting policy is only usable if its probabilities mean what they claim, and a detector deployed in line is only usable if its latency fits the traffic rate. Where cost is reported at all it tends to be a single training time rather than an inference budget. Explanation is reported by a growing minority [26], [28], [34], [43], but almost always as a qualitative artefact rather than as a measured fidelity.

Finally, statistical discipline is thin. Single-run results are the norm, seed dispersion is rarely given and paired significance testing is essentially absent, so differences of a few tenths of a point are reported as improvements when they lie within the range seed variation alone would produce.

Table 8. Audit of evaluation practice across the forty-seven included studies.

Dimension	Commonly reported	Commonly omitted
Benchmark provenance	Named public captures	Capture conditions; whether traffic is real or synthetic
Partitioning	Train/test split ratio	Whether transforms were fitted after the split
Unseen-family protocol	Claimed in the framing	An explicit leave-one-attack-out or held-out-family design
Discrimination	Accuracy, F1, ROC-AUC	Chance-corrected agreement on imbalanced label sets
Calibration	—	Log loss, Brier score, reliability behaviour
Computational cost	Training time	Per-sample inference latency; serialised model size
Statistical treatment	Single-run point estimates	Seed dispersion; paired significance testing
Cross-corpus transfer	—	Any evidence that a detector generalises beyond its corpus

6. Open Problems and Research Directions

Table 9 states nine open problems that follow from the synthesis, each paired with the direction it implies and with the studies in this corpus that already address it in part. Four deserve comment.

The first is benchmark adequacy. The field needs corpora whose difficulty is verified before they are used to rank detectors, and a saturation check — does a weak baseline such as Gaussian naive Bayes already solve it? — is a cheap and decisive test that almost no study reports. A benchmark on which many models are indistinguishable at the ceiling cannot support any ranking claim, however carefully the models are compared.

The second is protocol standardisation for the unseen-family claim. The claim that a detector catches zero-day attacks is testable: withhold an entire attack family from training and measure detection rate and benign false-positive rate on it. That design appears in the framing of many studies in this corpus and in the execution of few, and until it is routine the central claim of the field rests largely on assertion.
The third is the cost of the emerging paradigms. Language-model and neuro-symbolic detectors bring exactly the external knowledge the zero-day setting lacks, and dismissing them on cost grounds alone would be a mistake; but a defensible position requires their latency and monetary cost to be reported against the arrival rate of the traffic, and compared against the cheap baselines they are implicitly claimed to displace. The fourth is the reporting of negative results. Component-level evidence that a proposed module does not help is almost entirely absent across the corpus, which is not plausible as a description of what researchers observe. An ablation reporting a negative contribution is more informative than one confirming every design choice.

Table 9. Open problems, implied research directions and partial coverage in the reviewed corpus.

Open problem	Implied direction	Partially addressed by
P1 Benchmark difficulty is unverified	Publish a saturation check with every benchmark result	—
P2 The unseen-family claim is rarely tested	Adopt leave-one-attack-out as a standard protocol	[2], [4], [5], [25]
P3 Leakage control is left implicit	State the fit order of every transform relative to the split	[23]
P4 Novelty thresholds are fitted on the test partition	Fit operating points on a separate calibration split	[40], [45]
P5 Probabilities are uncalibrated	Report log loss, Brier score and reliability with accuracy	—
P6 Inference cost is unbudgeted	Report per-sample latency and model size against traffic rate	[29], [31], [43]
P7 Reasoning-based detectors are unpriced	Budget language-model cost against arrival rate and cheap baselines	[26], [27]
P8 Results are single-run	Repeat over seeds and apply paired significance testing	—
P9 Cross-corpus transfer is untested	Evaluate a common feature schema across at least two corpora	[36], [41]

7. Conclusion

This review examined forty-seven studies on learned zero-day intrusion detection published between 2020 and 2026, assigned each to one of six methodological families, and audited the evaluation practice of the corpus as a whole. The field is diversifying rather than converging: classical and deep pipelines still supply the largest share of the literature but none of its recent growth, while an emerging group built on language models, neuro-symbolic reasoning, temporal graphs, moving-target defence and federated learning accounts for nine of the seventeen studies dated 2026 and did not exist in the corpus before 2025.
The more durable finding is methodological. Across all six families the same omissions recur: benchmark difficulty is not verified, the unseen-family claim is asserted more often than it is tested, leakage control is left implicit, calibration and inference cost are seldom reported, results are single-run, and cross-corpus transfer is almost never attempted. These are not defects of any individual study but properties of the shared protocol, and they are the reason a quantitative ranking of the surveyed methods would not be sound. Progress in this area now depends less on further architectural variety than on an evaluation practice capable of telling one architecture from another.

References

[1] A. A. Mohamed, A. Al-Saleh, S. K. Sharma, and G. G. Tejani, "Zero-day exploits detection with adaptive WavePCA-Autoencoder (AWPA) adaptive hybrid exploit detection network (AHEDNet)," Scientific Reports, vol. 15, no. 1, p. 4036, 2025.
[2] H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Towards an effective zero-day attack detection using outlier-based deep learning techniques," arXiv preprint arXiv:2006.15344, 2020.
[3] M. A. Hossain, "Deep Q-learning intrusion detection system (DQ-IDS): A novel reinforcement learning approach for adaptive and self-learning cybersecurity," ICT Express, 2025.

[4] I. Mbona and J. H. P. Eloff, "Detecting zero-day intrusion attacks using semi-supervised machine learning approaches," IEEE Access, vol. 10, pp. 69822–69838, 2022.
[5] H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Utilising deep learning techniques for effective zero-day attack detection," Electronics, vol. 9, no. 10, p. 1684, 2020.
[6] D. Musleh, M. Alotaibi, F. Alhaidari, A. Rahman, and R. M. Mohammad, "Intrusion detection system using feature extraction with machine learning algorithms in IoT," Journal of Sensor and Actuator Networks, vol. 12, no. 2, p. 29, Mar. 2023.
[7] F. Türk, "Analysis of intrusion detection systems in UNSW-NB15 and NSL-KDD datasets with machine learning algorithms," Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, vol. 12, no. 2, pp. 465–477, 2023.
[8] S. Mohamed and R. Ejbali, "Deep SARSA-based reinforcement learning approach for anomaly network intrusion detection system," International Journal of Information Security, vol. 22, no. 1, pp. 235–247, 2023.
[9] A. Sarkar, H. S. Sharma, and M. M. Singh, "A supervised machine learning-based solution for efficient network intrusion detection using ensemble learning based on hyperparameter optimisation," International Journal of Information Technology, vol. 15, no. 1, pp. 423–434, 2023.
[10] S. Venkatesan, "Design an intrusion detection system based on feature selection using ML algorithms," Mathematical Statistician and Engineering Applications, vol. 72, no. 1, pp. 702–710, 2023.
[11] R. Chaganti, W. Suliman, V. Ravi, and A. Dua, "Deep learning approach for SDN-enabled intrusion detection system in IoT networks," Information, vol. 14, no. 1, p. 41, 2023.
[12] V. Hnamte and J. Hussain, "DCNNBiLSTM: An efficient hybrid deep learning-based intrusion detection system," Telematics and Informatics Reports, vol. 10, p. 100053, 2023.
[13] A. Abdelkhalek and M. Mashaly, "Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning," The Journal of Supercomputing, pp. 1–34, 2023.
[14] R. Singh and G. Srivastav, "Novel framework for anomaly detection using machine learning technique on CIC-IDS2017 dataset," in Proc. Int. Conf. Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 632–636, doi: 10.1109/ICTAI53825.2021.9673238.
[15] K. Soliman, M. A. Sobh, and A. M. Bahaa-Eldin, "Survey of machine learning HIDS techniques," in Proc. 16th Int. Conf. Computer Engineering and Systems (ICCES), Cairo, Egypt, 2021, pp. 1–5, doi: 10.1109/ICCES54031.2021.9686138.
[16] M. Bhatnagar, G. Rozinaj, and P. K. Yadav, "Web intrusion classification system using machine learning approaches," in Proc. Int. Symp. ELMAR, Zadar, Croatia, 2022, pp. 57–60, doi: 10.1109/ELMAR55880.2022.9899790.
[17] H. E. Alami and D. B. Rawat, "Detection of zero-day attacks using CNN and LSTM in networked autonomous systems," in Proc. IEEE Conf. Communications and Network Security (CNS), Orlando, FL, USA, 2023, pp. 1–2, doi: 10.1109/CNS59707.2023.10288680.
[18] K. Alrawashdeh and S. Goldsmith, "Optimizing deep learning based intrusion detection systems defense against white-box and backdoor adversarial attacks through a genetic algorithm," in Proc. IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 2020, pp. 1–8, doi: 10.1109/AIPR50011.2020.9425293.
[19] M. Malik and K. S. Saini, "Network intrusion detection system using reinforcement learning techniques," in Proc. Int. Conf. Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 1642–1649, doi: 10.1109/ICCPCT58313.2023.10245608.
[20] G. Guo, "An intrusion detection system for the Internet of Things using machine learning models," in Proc. 3rd Int. Conf. Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Xi'an, China, 2022, pp. 332–335, doi: 10.1109/ICBAIE56435.2022.9985800.
[21] S. Kumar, J. Kaswa, L. Meena, and M. Porwal, "Convolution neural network-based detection in software defined networks," in Proc. 2nd Int. Conf. Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1–6, doi: 10.1109/CONIT55038.2022.9847748.
[22] M. D. Rokade and S. S. Khatal, "Detection of malicious activities and connections for network security using deep learning," in Proc. IEEE Pune Section Int. Conf. (PuneCon), Pune, India, 2022, pp. 1–6, doi: 10.1109/PuneCon55413.2022.10014736.
[23] D. Macko, P. Goldschmidt, P. Pištek, and D. Chudá, "Assessing the impact of a supervised classification filter on flow-based hybrid network anomaly detection," arXiv preprint arXiv:2310.06656, 2023.
[24] S. SakthiMurugan, S. Kumaar, V. Vignesh, and P. Santhi, "Assessment of zero-day vulnerability using machine learning approach," EAI Endorsed Transactions on Internet of Things, vol. 10, 2024.

[25] M. Alkasassbeh, E. H. Omoush, M. Almseidin, and A. Aldweesh, "A self-adaptive intrusion detection system for zero-day attacks using deep Q-networks," IEEE Access, vol. 13, pp. 174280–174296, 2025, doi: 10.1109/ACCESS.2025.3617792.
[26] Y. K. Saheed and J. E. Chukwuere, "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, vol. 13, no. 7, pp. 14213–14228, Apr. 2026, doi: 10.1109/JIOT.2025.3648649.
[27] J. Yang et al., "LLMEdgeSec: LLM-enabled log reasoning for zero-day IoT threat detection," IEEE Internet of Things Journal, vol. 13, no. 15, pp. 32546–32555, Aug. 2026, doi: 10.1109/JIOT.2026.3686053.
[28] H. B. Ahmad, H. Gao, N. Latif, and P. Wang, "CTGAD-X: Explainable contrastive learning on temporal graphs for advanced threat detection," IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2026.3703460.
[29] N. Sharma, M. Swarnkar, and Divyanshu, "DLAZE: Detecting DNS tunnels using lightweight and accurate method for zero-day exploits," IEEE Transactions on Network and Service Management, vol. 22, no. 3, pp. 2343–2353, Jun. 2025, doi: 10.1109/TNSM.2025.3541234.
[30] R. Mukherjee, M. Azab, and T. Chantem, "Digital twins reimagined: Zero-day LLM-powered moving target defense in-depth for real-time CPS," IEEE Internet of Things Journal, vol. 13, no. 12, pp. 27001–27012, Jun. 2026, doi: 10.1109/JIOT.2026.3679937.
[31] H. J. Hadi, M. K. Khan, and N. Ahmad, "A real-time multi-tier machine learning intrusion detection framework for securing ground control station–UAV communications," IEEE Open Journal of the Communications Society, vol. 7, pp. 3881–3900, 2026, doi: 10.1109/OJCOMS.2026.3683883.
[32] D. Das, "SYNAPS: A neuro-symbolic framework for proactive security in consumer electronics," IEEE Transactions on Consumer Electronics, vol. 72, no. 1, pp. 2300–2308, Feb. 2026, doi: 10.1109/TCE.2025.3642082.
[33] R. Q. Shawl, M. Singh, and M. M. Hassan, "Leveraging cyber-physical security solutions blended with machine learning for advanced IoT botnet detection," IEEE Communications Standards Magazine, doi: 10.1109/MCOMSTD.2025.3636269.
[34] M. J. Hossain, K. Alam, M. F. Monir, M. M. Hoque, and T. Ahmed, "Explainable AI meets synthetic data: A deep learning framework for detecting network intrusion in NextG network infrastructure," IEEE Access, vol. 13, pp. 114979–115001, 2025, doi: 10.1109/ACCESS.2025.3585783.
[35] W. Fang et al., "Unknown cyber threat discovery empowered by genetic evolution without prior knowledge," IEEE Transactions on Information Forensics and Security, vol. 20, pp. 9552–9567, 2025, doi: 10.1109/TIFS.2025.3594569.
[36] A. Kadi, A. Selamnia, Z. A. E. Houda, H. Moudoud, B. Brik, and L. Khoukhi, "An in-depth comparative study of quantum-classical encoding methods for network intrusion detection," IEEE Open Journal of the Communications Society, vol. 6, pp. 1129–1148, 2025, doi: 10.1109/OJCOMS.2025.3537957.
[37] S. Jalil, M. Usman, B. Fong, and H. Kim, "Leveraging disentangled attention: A system-level design for URL-aware phishing detection," IEEE Communications Magazine, doi: 10.1109/MCOM.001.2600012.
[38] X. Zang, F. Hou, X. Liu, M. K. Khan, and D. Liu, "Mitigating the lateral movement of APT in IIoT with an efficient moving target defense and cyber deception approach," IEEE Transactions on Reliability, vol. 75, pp. 1753–1766, 2026, doi: 10.1109/TR.2026.3685261.
[39] Y. C. Krishna and B. Naresh Kumar Reddy, "A neurosymbolic architecture for self-evolving, causal and verifiable reasoning in zero-trust consumer electronic ecosystems," IEEE Transactions on Consumer Electronics, doi: 10.1109/TCE.2026.3713043.
[40] S. Anand and S. Zehra, "A comparative study of isolation forest and autoencoder models for proactive zero-day attack detection," in Proc. IEEE Int. Conf. Consumer Electronics (ICCE), Dubai, UAE, 2026, pp. 1–5, doi: 10.1109/ICCE67443.2026.11449661.
[41] M. U, A. K. B. R, and S. A, "Privacy-preserving federated threat intelligence using differential privacy for zero-day attack detection on UNSW-NB15," in Proc. IEEE Madhya Pradesh Section Conf. (MPCON), Gwalior, India, 2026, pp. 694–699, doi: 10.1109/MPCON69668.2026.11508553.
[42] P. N. G, V. A. Lepakshi, K. M, M. M, and Deeksha, "A review of unsupervised learning techniques for zero-day attack detection," in Proc. 3rd Int. Conf. Innovations in Cybersecurity and Data Science (ICICDS), Pathum Thani, Thailand, 2026, pp. 199–204, doi: 10.1109/ICICDS70526.2026.11604639.
[43] G. Singh and S. Kumari, "AIDL-ZDA: An explainable AI and adaptive deep learning framework for zero-day attacks detection in high-speed data," in Proc. Int. Conf. Emerging Smart Computing and Informatics (ESCI), Pune, India, 2026, pp. 1–6, doi: 10.1109/ESCI68015.2026.11493413.
[44] S. A. Alansary, S. M. Ayyad, F. M. Talaat, and M. M. Saafan, "A robust framework for zero-day attack detection using novel oversampling, LASSO feature selection, and Optuna-based optimization techniques," in Proc. 43rd National Radio Science Conf. (NRSC), New Damietta, Egypt, 2026, pp. 25–34, doi: 10.1109/NRSC71102.2026.11556068.

[45] D. Santhadevi, B. J. Keerthi, and R. N., "Isolation forest based detection of evil twin-assisted zero-day attacks in network traffic," in Proc. 13th Int. Conf. Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2026, pp. 1–7, doi: 10.23919/INDIACom70271.2026.11526278.

[46] A. Al Siam, N. Faruqui, A. Azad, and M. A. Moni, "Securing the unseen: A comprehensive exploration review of AI-powered models for zero-day attack detection," Expert Systems, vol. 43, no. 3, p. e70217, 2026.

[47] S. Hussain, V. Kalpana, V. Raghavendran, N. Gangatharan, B. Mouleswararao, and V. RK, "Machine learning-based zero-day attack detection using deep reinforcement learning," Information Security Journal: A Global Perspective, pp. 1–23, 2026.