International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

A Risk-Aware Evaluation and Resilience Framework for Machine Learning-Based Zero-Day Attack Detection

Bushra Naz [1]
Department of Information Technology, Acharya Panth Shri Grindh Muni Naam Saheb Govt. P.G. College, Kabirdham Chhattisgarh, India.

Abstract - Zero-day attacks exploit vulnerabilities or attack behaviors for which reliable signatures or prior detection knowledge may not be available at the time of exploitation. Machine learning (ML) and anomaly-based intrusion detection are promising because they can model behavioral patterns rather than depend exclusively on known signatures. However, high performance on conventional benchmark datasets does not by itself establish genuine zero-day detection capability. This paper critically examines the reliability of ML-based zero-day attack detection from the perspectives of data representativeness, unseen-attack generalization, concept and feature drift, adversarial manipulation, false positives, explainability, and operational constraints. A targeted literature analysis is used to synthesize findings from foundational intrusion-detection research and recent zero-day, adaptive-learning, adversarial-robustness, and explainable-AI studies. Based on this synthesis, the paper proposes a Risk-Aware Zero-Day Detection Resilience Framework (RZDRF). Rather than presenting another classifier, the framework defines a layered architecture and an evaluation protocol in which detection performance is considered together with novelty validity, temporal robustness, adversarial resilience, calibration, explainability, analyst feedback, and computational efficiency. The framework is intentionally theoretical and is not presented as experimentally validated. Its principal contribution is a structured way to evaluate whether an ML-based detector is genuinely resilient to unseen threats and operationally trustworthy. Future work should implement and benchmark the framework using temporally separated and cross-dataset evaluations.

Keywords: zero-day attack; intrusion detection; machine learning; anomaly detection; concept drift; adversarial machine learning; explainable AI; human-in-the-loop; resilience; evaluation

1. INTRODUCTION

Zero-day attacks exploit vulnerabilities or attack behaviors for which reliable signatures or prior detection knowledge may not be available at the time of exploitation. Machine learning (ML) and anomaly-based intrusion detection can identify behavioral deviations, but conventional benchmark performance does not by itself establish genuine zero-day capability.

This paper addresses a specific methodological problem: existing studies often evaluate detection capability, robustness, explainability, or operational feasibility separately, making it difficult to determine whether a detector is both effective against genuinely unseen attacks and trustworthy under changing conditions. A targeted critical literature analysis synthesizes evidence on unseen-attack generalization, data representativeness, concept and feature drift, adversarial manipulation, false positives, explainability, and deployment constraints.

Based on this synthesis, the paper proposes a Risk-Aware Zero-Day Detection Resilience Framework (RZDRF) together with a formal evaluation model. RZDRF integrates dual threat detection, contextual risk scoring, adversarial validation, explainability, human analyst review, controlled adaptation, and deployment-aware monitoring. Its principal methodological novelty is the explicit traceability from a zero-day reliability risk to a corresponding detection control, evaluation test, and measurable outcome.

The proposed evaluation model further represents resilience as a multi-dimensional construct rather than a single accuracy value. It defines normalized evaluation dimensions and a configurable composite resilience index whose weights can be calibrated to operational priorities. The framework is intentionally theoretical and is not presented as experimentally validated.

The contribution is therefore a reproducible conceptual basis for designing and evaluating zero-day detection studies, with explicit novelty definitions and leakage-resistant testing requirements. Future work should implement RZDRF, estimate its parameters empirically, and compare the resulting resilience measures with conventional benchmark evaluation.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

2. RESEARCH QUESTIONS AND CONTRIBUTIONS

The study is guided by four research questions:

RQ1: What theoretical and operational factors limit the ability of ML-based intrusion detection systems to identify genuinely novel zero-day attacks?

RQ2: How should zero-day detection performance be evaluated so that high benchmark accuracy is not confused with genuine generalization to unseen threats?

RQ3: How can detection controls, risk prioritization, analyst review, and controlled adaptation be linked to measurable resilience outcomes?

RQ4: What research gaps remain for building evaluation protocols that are leakage-resistant, operationally meaningful, and reproducible?

The paper makes four connected contributions. First, it develops a multidimensional taxonomy of zero-day detection approaches and evaluation settings. Second, it defines a traceable risk-to-control-to-test structure that links identified reliability risks to specific RZDRF mechanisms and evaluation evidence. Third, it formalizes resilience as a set of normalized dimensions and a configurable composite index rather than accuracy alone. Fourth, it specifies a leakage-resistant evaluation protocol covering novelty validity, temporal robustness, cross-dataset generalization, false-positive control, adversarial robustness, explainability, drift adaptation, and deployment efficiency. The framework is conceptual and requires empirical validation before effectiveness claims are made.

3. REVIEW METHODOLOGY

This paper uses a targeted critical literature review rather than claiming a full systematic review or meta-analysis. Sources were selected to cover four evidence groups: (i) foundational intrusion-detection and ML research; (ii) zero-day-specific surveys and empirical studies; (iii) adaptive and concept-drift-aware IDS research; and (iv) adversarial ML and XAI research relevant to security operations. Priority was given to peer-reviewed IEEE, Elsevier, Springer, ACM, and major open-access scholarly sources, with emphasis on 2023–2026 work for the current state of the field.

The synthesis used the following analytical dimensions: detection paradigm, definition of unseen/zero-day data, training-test separation, temporal validity, generalization, concept drift, adversarial threat model, false-positive behavior, explainability, human involvement, and deployment cost. Studies were not combined statistically because the datasets, experimental protocols, attack definitions, and metrics are heterogeneous. The resulting framework should therefore be interpreted as a conceptual synthesis requiring empirical validation.

4. BACKGROUND AND PROBLEM FORMULATION

4.1 Zero-Day Attacks and Unseen Data

A zero-day attack should be understood operationally rather than merely as any sample absent from a training file. For this paper, a genuinely novel zero-day scenario is one in which the malicious behavior or attack family is not available as representative labeled training information at model-development time. Unseen test samples can instead represent new instances, new variations of known attacks, or genuinely novel attack families. These categories should not be treated as equivalent.

4.2 ML-Based Detection Paradigms

Supervised classifiers learn from labeled benign and malicious examples, but their performance may degrade when the test attack distribution differs from training. Unsupervised and semi-supervised anomaly detection can model normal behavior and flag deviations, but legitimate rare behavior can also generate alerts. Hybrid approaches attempt to combine complementary strengths. The literature therefore supports using ML as one layer in a broader detection system rather than assuming that a single algorithm can solve the zero-day problem [4–6].

4.3 The Evaluation Problem

A major methodological risk is leakage between training and testing. Randomly splitting records from the same attack campaign can allow highly similar traffic characteristics to appear in both partitions, producing optimistic results. For zero-day claims, the evaluation should explicitly hold out attack families, time periods, or datasets according to the research question. Recent work on unseen-data detection demonstrates the importance of defining what "unseen" means, while concept-drift research emphasizes that traffic distributions change over time [6,7].

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

5. CRITICAL LITERATURE SYNTHESIS

Machine learning-based approaches for zero-day attack detection can be categorized according to their learning paradigm, model architecture, detection strategy, evaluation setting, and deployment domain. This multidimensional classification helps distinguish the methodological choices used in existing studies and highlights differences in how zero-day detection capability is evaluated. The resulting taxonomy is presented in Fig. 1.

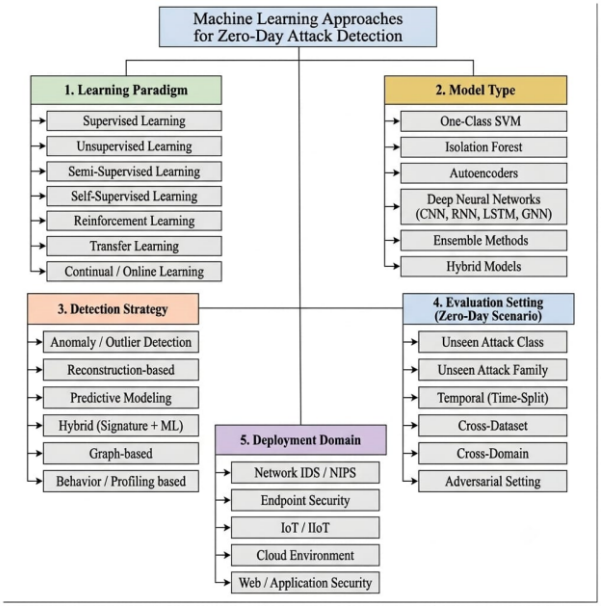

Figure 1 Caption: Taxonomy of Machine Learning Approaches for Zero-Day Attack Detection.

Figure 1 Alt Text: A taxonomy diagram classifying machine learning approaches for zero-day attack detection into learning paradigms, model types, detection strategies, evaluation settings, and deployment domains. The diagram includes supervised, unsupervised, semi-supervised, self-supervised, reinforcement, transfer, and continual learning; one-class SVM, isolation forest, autoencoders, deep neural networks, ensemble and hybrid models; anomaly/outlier, reconstruction, predictive, hybrid, graph-based, and behavior/privilege-based detection; unseen attack cases, unseen attack families, temporal, cross-dataset, cross-domain, and adversarial settings; and network/IDS, endpoint, IoT/IIoT, cloud, and web/application security domains.

The taxonomy organizes machine learning approaches for zero-day attack detection across five dimensions: learning paradigm, model type, detection strategy, evaluation setting, and deployment domain. The evaluation-setting dimension distinguishes unseen attack classes, unseen attack families, temporal evaluation, cross-dataset evaluation, cross-domain evaluation, and adversarial settings, providing a structured basis for assessing the generalization and robustness of zero-day detection methods.

The taxonomy demonstrates that zero-day detection is not determined solely by the choice of machine learning algorithm. The evaluation setting and deployment environment also influence the interpretation of detection performance. In particular, evaluation using unseen attack classes, temporal splits, and cross-dataset settings provides stronger evidence of generalization than conventional random data partitioning. These observations motivate the research gap identified in the following section.

Table 1: Summary of Existing Research Limitations and Proposed RZDRF Response

Research dimension	What existing work shows	Remaining limitation	Implication for zero-day claims	RZDRF response
Zero-day generalization	ML can identify patterns in unseen traffic, but performance varies across attack types [4,6].	Unseen instances are not always genuine novel attack families.	Accuracy alone cannot establish zero-day capability.	Explicit novelty categories and hold-out protocols.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Data & representation	Benchmark datasets support reproducible evaluation but may not represent evolving environments [3,4,6].	Dataset bias, class imbalance and feature dependence.	Cross-dataset and temporal tests are needed.	Data provenance and representativeness checks.
Concept/feature drift	Traffic and feature distributions can evolve over time [7].	Static models may become stale; continuous learning introduces poisoning risk.	Performance must be tracked over time.	Drift monitoring + controlled adaptation.
Adversarial robustness	Attackers can manipulate inputs or training data to influence ML decisions [10–12].	Many IDS evaluations omit realistic threat models.	Robustness claims require explicit attack assumptions.	Adversarial validation and model diversity.
False positives	Anomalies can arise from legitimate rare behavior.	Alert fatigue can reduce operational value.	Precision/recall should be complemented by FPR and workload measures.	Risk scoring and analyst feedback.
Explainability	XAI is increasingly used to support transparency in IDS [8,9].	Explanations themselves have quality, scalability and adversarial issues.	Explainability must be evaluated, not merely added.	Explanation + confidence + analyst validation.
Deployment	Real-time IDS faces latency, scalability and computational constraints [2,5,7].	Lab accuracy may not translate to production.	Latency and resource cost are part of effectiveness.	Tiered edge/central processing and efficiency metrics.

Table 1 summarizes the key limitations identified in existing research on machine learning-based zero-day attack detection and maps them to the corresponding implications and proposed responses of the Risk-Aware Zero-Day Detection Resilience Framework (RZDRF). The table highlights limitations related to zero-day generalization, data representation, concept and feature drift, and adversarial robustness.

6. RESEARCH GAP

The literature demonstrates that many components required for resilient zero-day detection already exist individually. Prior studies have investigated anomaly detection for unseen attacks, adaptive learning and concept drift, adversarial robustness, explainable intrusion detection, and human-assisted security. Therefore, the novelty of this work is not the invention of any one of these components.

The identified gap is methodological integration with explicit evaluation traceability. Existing work commonly reports a model, dataset, and metric set, but the evidence needed to support a zero-day claim is not always connected to the specific reliability risk being tested. RZDRF addresses this gap through a risk-to-control-to-test chain: each major failure mode is mapped to a framework control, a stress condition, and an observable outcome. The framework additionally separates detection capability from resilience, so that a detector can be assessed for unseen-attack generalization, temporal stability, adversarial degradation, false-positive burden, explanation quality, drift recovery, and operational cost.

The resulting contribution is a formal evaluation-oriented model rather than another classifier. RZDRF does not claim that its architecture itself proves zero-day detection. Instead, it specifies what evidence a future implementation must produce before a zero-

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

day resilience claim can be considered credible. The formalization introduced in Section 7.3 and evaluated in Section 8 is the central mechanism for making this contribution testable.

7. PROPOSED RISK-AWARE ZERO-DAY DETECTION RESILIENCE FRAMEWORK (RZDRF)

RZDRF is a layered theoretical framework designed to reduce the gap between benchmark performance and operational zero-day resilience. Its distinguishing feature is not simply the combination of familiar security components, but the explicit coupling of detection, risk prioritization, validation, and evidence-based adaptation. This coupling makes each reliability concern visible as an evaluable part of the system.

The framework retains conventional controls for known threats while using ML/anomaly detection for behavior that does not match known signatures. Unknown behavior is not automatically labelled as a zero-day attack; instead, it passes through contextual risk assessment and validation. The framework therefore separates anomaly detection from the stronger scientific claim of genuine zero-day detection.

RZDRF comprises eight layers: multi-source data and context; dual detection; zero-day risk scoring; adversarial validation; explainability and analyst-centric review; human-in-the-loop feedback; drift monitoring and controlled adaptation; and deployment-aware processing. Across these layers, the framework records evidence needed for the evaluation protocol rather than treating performance as a single end point.

Figure 2 Caption: Proposed Risk-Aware Zero-Day Detection Resilience Framework.

Figure 2 Alt Text: A layered flow diagram of the proposed Risk-Aware Zero-Day Detection Resilience Framework. The framework includes data sources, preprocessing and feature engineering, known and unknown threat detection, risk scoring and prioritization, validation, human-in-the-loop analyst review, response and action, and continuous learning and adaptation. Data flows from multiple sources through detection and validation, while analyst feedback and response information support controlled model updates and concept-drift monitoring.

The framework integrates data sources, preprocessing, dual threat detection, risk scoring, validation, human-in-the-loop decision making, response actions, and continuous learning. Feedback from validation and response stages supports model adaptation and concept-drift monitoring.

7.1 Layer 1: Multi-Source Data and Context

The framework accepts network flows, system logs, endpoint telemetry, and user or application activity where available. Context such as time, asset role, operational state, and recent incidents is used to distinguish legitimate anomalies from suspicious deviations. Data provenance, missingness, imbalance, and duplication should be documented before model training.

7.2 Layer 2: Dual Detection

Known threats are handled by signatures, rules, and other deterministic controls. Concurrently, an ML/anomaly layer models behavioral regularities and produces an anomaly score. The two streams are complementary rather than mutually exclusive. A key design principle is that an anomaly score should not automatically be interpreted as proof of a zero-day attack.

7.3 Layer 3: Zero-Day Risk Scoring

The framework converts multiple signals into a risk score rather than a binary anomaly decision. Let the normalized evidence vector for an event be x = [A, C, P, T, K, Q], where A denotes anomaly magnitude, C contextual deviation, P asset criticality, T temporal persistence, K corroborating telemetry, and Q model confidence. A configurable event risk score can be expressed as:

$R_{event} = \Sigma_{j=1 \text{ to } 6}\ \alpha_j x_j, \quad \text{where } \alpha_j \geq 0 \text{ and } \Sigma\ \alpha_j = 1.$

The coefficients α_j are not fixed by the conceptual framework. They should be estimated or calibrated from operational costs, expert elicitation, or empirical validation. This avoids arbitrary weighting while making the risk-scoring mechanism explicit and testable. The event score can then be mapped to action thresholds using thresholds selected on validation data rather than on the final zero-day test set.

7.4 Layer 4: Adversarial Validation

Because attackers may deliberately evade or poison learning systems, the framework includes an adversarial validation stage. Evaluation should specify whether the assumed attacker can modify test traffic, influence training data, or exploit model knowledge. Robustness should be measured under controlled evasion and poisoning scenarios rather than inferred from clean-data accuracy alone [10–12].

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

7.5 Layer 5: Explainability and Analyst-Centric Review

Each high-risk alert should provide an explanation appropriate to the model and data, such as influential features, behavioral deviations, confidence, and supporting context. Recent IDS literature shows that XAI introduces its own research challenges, including explanation quality, scalability, and vulnerability [8,9]. Therefore, RZDRF treats explanation quality as an evaluation dimension rather than a cosmetic interface feature.

7.6 Layer 6: Human-in-the-Loop Feedback

Security analysts validate ambiguous alerts, label emerging behaviors, and provide feedback for controlled model updates. Human involvement is especially important where an anomaly could represent either benign operational change or an emerging attack. Feedback must be governed because automatic learning from unverified alerts can introduce model poisoning and feedback loops.

7.7 Layer 7: Drift Monitoring and Controlled Adaptation

Concept drift monitoring continuously checks whether traffic or feature distributions have changed. Adaptation may use periodic retraining or online learning, but model updates should pass validation before deployment. Recent work identifies concept and feature drift as important unresolved issues for IDS [7].

7.8 Layer 8: Deployment-Aware Processing

Lightweight screening can occur close to data sources, while computationally intensive analysis can be performed centrally or in cloud infrastructure. The deployment layer records latency, throughput, memory/CPU utilization, update cost, and alert volume. These measures become part of the resilience evidence because a detector that is accurate but operationally unusable does not provide dependable security value.

8. PROPOSED EVALUATION PROTOCOL

The principal methodological contribution of RZDRF is a structured evaluation protocol that converts the framework's reliability risks into explicit tests and measurable outcomes. A future implementation should report results across all dimensions rather than presenting accuracy alone.

Reliable evaluation of zero-day attack detection requires more than conventional random train–test splitting. The protocol explicitly distinguishes novelty validity from ordinary unseen-instance testing and evaluates temporal robustness, cross-dataset and cross-domain generalization, adversarial resilience, false-positive burden, explainability, drift adaptation, and operational efficiency. The six-stage protocol is illustrated in Fig. 3.

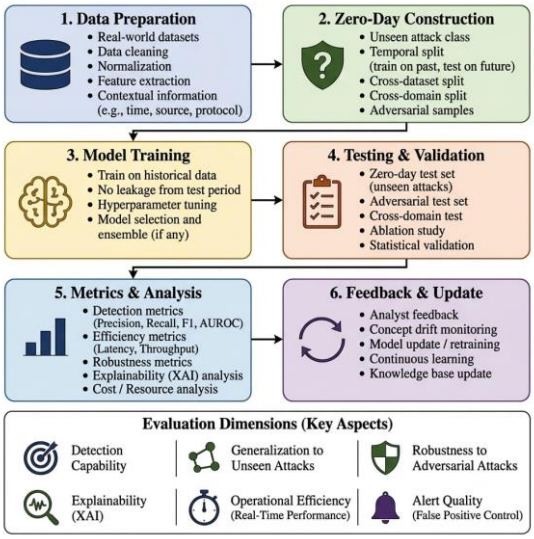

Figure 3 Caption: Proposed Evaluation Protocol for Machine Learning-Based Zero-Day Attack Detection.

Figure 3 Alt Text: A six-stage evaluation protocol for machine learning-based zero-day attack detection. The stages are data preparation, zero-day scenario construction, model training, testing and validation, metrics and analysis, and feedback and update. The protocol evaluates detection capability, generalization to unseen attacks, robustness to adversarial attacks, explainability, operational efficiency, and alert quality.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

The protocol comprises six stages: data preparation, zero-day scenario construction, model training, testing and validation, metrics and analysis, and feedback and model updating. The evaluation dimensions emphasize detection capability, generalization to unseen attacks, adversarial robustness, explainability, operational efficiency, and alert quality.

8.1 Data Preparation

The first stage involves preparing representative cybersecurity datasets for model development and evaluation. The data may consist of network traffic, system logs, endpoint activity, or other security-relevant observations. Data preprocessing includes cleaning inconsistent or missing records, normalization of numerical features, categorical feature encoding, and extraction of relevant behavioral characteristics.

Contextual information such as timestamp, source, destination, communication protocol, and traffic characteristics should be retained where available. Maintaining temporal and contextual information is important because zero-day detection performance can be strongly influenced by changes in network behavior and operating conditions.

The prepared dataset should subsequently be divided into training, validation, and testing subsets while ensuring that information from the evaluation period does not leak into the training process.

8.2 Zero-Day Scenario Construction

The second stage constructs an evaluation scenario in which the model encounters attack behavior that was not represented in its training data. Merely using a random test subset does not necessarily constitute a meaningful zero-day evaluation because similar samples from the same attack family may already exist in the training set.

Therefore, several evaluation settings can be considered, including unseen attack classes, unseen attack families, temporal splits, cross-dataset evaluation, and cross-domain evaluation. In a temporal setting, historical observations are used for training while observations from a later period are reserved for testing. Cross-dataset and cross-domain settings can further examine whether a model maintains its detection capability when the underlying data distribution changes.

Adversarial samples may additionally be incorporated to evaluate the resilience of the detection mechanism against intentionally manipulated inputs.

8.3 Model Training

In the third stage, the selected machine learning model is trained exclusively using the designated training data. The validation set can be used for hyperparameter selection and model comparison without exposing the final zero-day test set to the training procedure.

Different learning paradigms, including supervised, unsupervised, semi-supervised, anomaly-based, and hybrid approaches, may be evaluated depending on the research objective. For hybrid approaches, known threats can be identified using signature- or rule-based mechanisms, while machine learning-based anomaly detection can be used to identify potentially unknown behavior.

Strict separation between training and testing data is essential. In particular, samples from the designated zero-day evaluation period should not be used for model training, feature selection, hyperparameter optimization, or threshold selection.

8.4 Testing and Validation

The fourth stage evaluates the trained model using previously unseen data. The primary zero-day test set should contain attack behavior excluded from the training process. Additional testing can be conducted using adversarial samples, cross-domain datasets, or temporally separated data.

The evaluation should also examine false positives because anomalous behavior is not necessarily malicious. A model that identifies a large number of legitimate activities as threats may generate excessive alerts and reduce its practical usefulness.

Where applicable, ablation studies can be performed to determine the contribution of individual framework components, such as contextual features, adaptive learning, adversarial validation, or explainability mechanisms. Statistical validation can further be used to determine whether observed differences between approaches are consistent rather than resulting from random variation.

8.5 Metrics and Analysis

The fifth stage evaluates the detection system using multiple complementary metrics rather than relying solely on accuracy. Detection capability can be assessed using precision, recall, F1-score, and AUROC, while false-positive rate should also be considered.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

For real-time deployment, operational metrics such as detection latency, throughput, computational resource consumption, and scalability become important. Robustness evaluation should examine the degradation in performance under adversarial or distribution-shift conditions.

Explainability should also be considered when the system is intended to support security analysts. The evaluation may examine the consistency and usefulness of generated explanations and their ability to provide meaningful information about why an event was classified as suspicious.

Finally, alert quality should be assessed through measures such as false-positive rate, alert volume, and analyst workload. This provides a more practical assessment of whether a detection model can support real-world security operations.

8.6 Feedback and Model Update

The final stage introduces feedback from the evaluation and operational environment into the detection process. Analyst decisions can provide additional information about confirmed attacks, false positives, and ambiguous events. This information can subsequently support controlled model updates and knowledge-base refinement.

Continuous monitoring should also be used to identify concept drift, where the statistical characteristics of legitimate or malicious behavior change over time. When significant distribution changes are detected, the model may require retraining, incremental updating, or adaptation using newly validated data.

However, continuous learning should be performed in a controlled manner. Automatically incorporating every detected anomaly into the training data may introduce noisy or incorrect labels and can potentially make the model vulnerable to poisoning. Therefore, analyst validation and appropriate data-quality controls should precede incorporation of new information into future model updates.

8.7 Resilience Scoring and Evidence Rules

To make the multidimensional evaluation comparable across studies, each evaluation dimension can be normalized to a score S_i in the interval [0,1], where higher values represent better validated resilience for that dimension. A configurable RZDRF Resilience Index (RRI) can then be defined as:

$RRI = \Sigma(i=1 \text{ to } 8)\ w_i S_i, \quad \text{where } w_i \geq 0 \text{ and } \Sigma\ w_i = 1.$

The eight dimensions are novelty validity, temporal robustness, cross-dataset generalization, false-positive control, adversarial robustness, explainability, drift adaptation, and operational efficiency. The index is intended for comparative evaluation, not as a universal security score. The weights w_i should be reported transparently and selected according to the target environment or estimated empirically. Every component S_i should also be reported separately so that a high aggregate score cannot conceal a serious weakness in one dimension.

This formalization creates a direct test of the framework's central claim: resilience is multidimensional and must be demonstrated under defined stress conditions. It also makes the proposed contribution falsifiable, because future empirical work can determine whether RZDRF-based evaluation reveals reliability differences that conventional accuracy-based evaluation misses.

Table 2: Evaluation Framework and Recommended Tests For RZDRF

Evaluation dimension	Recommended test	Primary metrics	Interpretation
Novelty validity	Hold out an attack family or behavior not represented in training	Zero-day recall, precision, F1, MCC, AUROC/AUPRC	Measures generalization to genuinely novel categories.
Temporal robustness	Train on earlier period, test on later period	F1, FPR, recall over time	Measures resilience to evolving traffic.
Cross-dataset generalization	Train on one dataset, test on another compatible dataset	F1, MCC, AUPRC, calibration	Tests dependence on one benchmark.
False-positive control	Evaluate benign rare/abnormal events	FPR, alerts/hour, precision, analyst workload	Measures operational alert burden.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Adversarial robustness	Apply defined evasion/poisoning threat models	Performance degradation, attack success rate	Quantifies resilience to adaptive attackers.
Explainability	Evaluate explanation fidelity/consistency and analyst usefulness	Fidelity, stability, explanation time, analyst agreement	Tests whether explanations are useful and reliable.
Drift adaptation	Inject or observe distribution changes	Recovery time, post-drift F1, update cost	Measures adaptation rather than static performance.
Efficiency	Measure on realistic hardware/traffic rate	Latency, throughput, CPU/RAM, energy where relevant	Tests deployment feasibility.

Table 2 summarizes the recommended evaluation tests for the proposed Risk-Aware Zero-Day Detection Resilience Framework (RZDRF). It maps each evaluation dimension to an appropriate testing strategy, primary performance metrics, and the corresponding interpretation. The table emphasizes novelty validity, temporal robustness, cross-dataset generalization, false-positive control, adversarial robustness, explainability, drift adaptation, and operational efficiency.

9. DATASET AND EXPERIMENTAL DESIGN RECOMMENDATIONS

The framework does not prescribe a single dataset because no single benchmark can represent all zero-day conditions. Candidate datasets may include UNSW-NB15, CSE-CIC-IDS2018, CIC-MalMem-2022, and other contemporary datasets appropriate to the target environment. The purpose of this section is to define an experimental design that can test the framework's claims rather than to imply that these datasets have already validated RZDRF.

Experimental design should prevent leakage. If the research question concerns new attack families, entire families should be withheld from training. If it concerns temporal evolution, chronological splits should be used. If cross-environment generalization is claimed, a separate dataset or environment should be used. Random record-level splitting should not be the sole evidence for a zero-day claim.

Metrics should include precision, recall, F1, MCC or another imbalance-aware measure, false-positive rate, and where appropriate AUPRC. For operational studies, detection latency, throughput, alert volume, and resource consumption should also be reported. If probability outputs are used for risk scoring, calibration should be evaluated. The RRI should be reported together with all component scores so that the aggregate does not obscure trade-offs.

A future implementation should report a baseline detector, the RZDRF-enabled configuration, and component-level ablations. The ablations should remove or disable risk scoring, adversarial validation, contextual information, controlled adaptation, or analyst feedback one at a time where applicable. This design can determine whether the proposed integration contributes measurable value instead of assuming that integration is beneficial.

10. DISCUSSION

The literature supports the view that ML can contribute meaningfully to intrusion detection, but it does not support treating high benchmark accuracy as proof of universal zero-day detection. The strength of a zero-day claim depends on how novelty is defined, how leakage is controlled, and whether the detector remains useful when traffic, attacker behavior, and operational conditions change.

RZDRF reframes the research objective from detecting every zero-day attack to producing validated evidence of resilience under defined novelty, drift, adversarial, explainability, and operational conditions. Its central contribution is the traceability between risk, control, test, and outcome. This structure allows future studies to identify not only whether a model fails, but why the failure matters operationally and which framework component is intended to mitigate it.

The proposed RRI is deliberately configurable rather than presented as a universal benchmark. Its value is methodological: it forces studies to expose the dimensions behind an aggregate assessment. A detector could achieve strong novelty validity while performing poorly under drift or adversarial manipulation; reporting the component scores preserves this distinction.

The framework also avoids assuming that greater model complexity is automatically better. Complexity can increase computational cost and reduce interpretability, while continuous learning can introduce poisoning risk. The appropriate design is consequently a

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

controlled feedback loop in which updates are monitored, validated, explainable, and reversible. Human analysts remain a security control, particularly for ambiguous alerts and model-update decisions.

11. LIMITATIONS

This work has several limitations. First, RZDRF is theoretical and has not yet been implemented or experimentally validated. Second, the risk-score coefficients and resilience-index weights are intentionally left configurable because they depend on operational priorities and must be calibrated empirically. Third, the targeted critical literature review is not a registered systematic review and therefore should not be presented as exhaustive. Fourth, the framework cannot establish that an anomaly is a zero-day exploit without additional evidence; it is designed to prioritize and validate suspicious novel behavior.

These limitations also identify the conditions under which the proposed contribution could be falsified or refined. An implementation study should test whether the risk-to-control-to-test structure and the RRI reveal meaningful performance differences that are missed by conventional accuracy-based evaluation. Negative findings should be reported alongside positive findings.

12. FUTURE WORK

Future work should implement RZDRF using one or more anomaly detectors and calibrated risk scoring; construct leakage-resistant temporal and attack-family hold-out experiments; compare supervised, unsupervised, semi-supervised, and hybrid approaches under identical protocols; and perform component ablations to quantify the contribution of each RZDRF mechanism.

Further work should evaluate concept-drift detectors and controlled continual-learning strategies; test adversarial evasion and poisoning under explicit threat models; measure XAI quality and analyst usefulness; evaluate latency, throughput, resource consumption, and model-update overhead; and release reproducible preprocessing, split definitions, code, and evaluation scripts where licensing permits.

- Compare supervised, unsupervised, semi-supervised, and hybrid approaches under identical zero-day protocols.
- Evaluate concept-drift detectors and controlled continual-learning strategies.
- Test adversarial evasion and poisoning under explicitly stated threat models.
- Measure XAI quality and analyst usefulness rather than reporting explanation availability alone.
- Evaluate computational cost, latency, throughput, and model-update overhead on realistic infrastructure.
- Release reproducible preprocessing, split definitions, code, and evaluation scripts where licensing permits.

13. CONCLUSION

Machine learning offers valuable capabilities for identifying anomalous cyber behavior, but genuine zero-day detection is a harder problem than ordinary supervised classification. The evidence reviewed in this paper indicates that novelty definition, data representativeness, concept drift, adversarial manipulation, false positives, explainability, and deployment constraints all influence the reliability of ML-based detection.

The proposed Risk-Aware Zero-Day Detection Resilience Framework contributes an evaluation-oriented structure that connects these risks to detection controls, validation procedures, analyst decisions, controlled adaptation, and deployment-aware evidence. Its methodological novelty lies in making this relationship explicit and in formalizing resilience as a multidimensional assessment rather than a single accuracy claim.

The framework and the proposed RRI are not experimentally proven solutions. Their purpose is to provide a transparent, testable basis for future implementations. Empirical validation, component ablation, temporal and cross-dataset testing, adversarial evaluation, and operational measurement are necessary to determine whether RZDRF improves the reliability of zero-day detection claims.

The study is based on a targeted critical literature review and proposes a theoretical framework. No original experimental dataset was generated or analyzed in this study. The sources discussed in the manuscript are identified in the reference list.

REFERENCES

[1] R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” IEEE Symposium on Security and Privacy, 2010, pp. 305–316. DOI: 10.1109/SP.2010.25.

[2] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016. DOI: 10.1109/COMST.2015.2494502.

[3] N. Moustafa and J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems,” MILCIS 2015, pp. 1–6. DOI: 10.1109/MilCIS.2015.7348942.

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

[4] Y. Guo, "A Review of Machine Learning-Based Zero-Day Attack Detection: Challenges and Future Directions," Computer Communications, vol. 198, pp. 175–185, 2023. DOI: 10.1016/j.comcom.2022.11.001.
[5] L. Y. Por et al., "A Systematic Literature Review on AI-Based Methods and Challenges in Detecting Zero-Day Attacks," IEEE Access, vol. 12, pp. 144150–144163, 2024. DOI: 10.1109/ACCESS.2024.3455410.
[6] Z. Dai et al., "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," PLOS ONE, vol. 19, no. 9, e0308469, 2024. DOI: 10.1371/journal.pone.0308469.
[7] M. A. Shyaa et al., "Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems," Engineering Applications of Artificial Intelligence, vol. 137, 109143, 2024. DOI: 10.1016/j.engappai.2024.109143.
[8] M. Pawlicki, A. Pawlicka, R. Kozik, and M. Choraś, "The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation," Artificial Intelligence Review, vol. 57, art. 330, 2024. DOI: 10.1007/s10462-024-10972-3.
[9] A. Nascita et al., "A Survey on Explainable Artificial Intelligence for Internet Traffic Classification and Prediction, and Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 27, no. 5, pp. 3165–3198, 2024. DOI: 10.1109/COMST.2024.3504955.
[10] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," Expert Systems with Applications, vol. 238, 122223, 2024. DOI: 10.1016/j.eswa.2023.122223.
[11] B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," Pattern Recognition, vol. 84, pp. 317–331, 2018. DOI: 10.1016/j.patcog.2018.07.023.
[12] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial Attacks on Neural Networks for Graph Data," Proceedings of KDD 2018, pp. 2847–2856, 2018. DOI: 10.1145/3219819.3220078.
[13] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," MILCIS 2015. DOI: 10.1109/MilCIS.2015.7348942.
[14] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," ICISSP, 2018. Dataset: CSE-CIC-IDS2018, Canadian Institute for Cybersecurity.

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