IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Zero-Day Exploit Detection using Machine Learning

Dr. Yogendra Patil, Dr. P. B. Dhamdhere, Ms. Bharati Ganesh Salve
Department of Computer Engineering,
Marathwada Mitra Mandal's College of Engineering, Pune, India
patyogendra@gmail.com, Pramod.dhamdhere@mmit.edu.in, bharatisalve31@gmail.com

Abstract: *Zero-day attacks continue to challenge cybersecurity systems because they exploit vulnerabilities that have not yet been discovered or documented. Traditional detection methods struggle with these unpredictable threats, leading researchers to explore more adaptive and intelligent solutions. Recent work shows significant progress through machine learning, deep learning, explainable AI, and federated learning each contributing unique strengths for identifying unfamiliar attack behaviors. Explainable AI improves analyst trust by clarifying why a model flags certain activities, while federated learning enables organizations to collaborate on detection models without exposing sensitive data. At the same time, the rise of quantum computing has encouraged the development of cryptographic techniques designed to secure future AI-driven systems. This review brings together findings from recent advances across these areas to illustrate how modern techniques can be combined to build a more resilient and proactive defense against zero-day exploits. The review also highlights open challenges and suggests research directions for developing reliable, scalable, and transparent security solutions.*

Keywords: Zero-day attacks, machine learning, deep learning, explainable AI (XAI), federated learning, anomaly detection, cybersecurity, intrusion detection, quantum-safe security, autonomous defence systems

I. INTRODUCTION

Zero-day attacks remain one of the most disruptive and damaging categories of cyber threats because they exploit vulnerabilities before security teams are even aware of their existence [1], [2]. The unpredictability of these attacks allows adversaries to bypass signature-based and rule-driven defenses, leaving organizations exposed to sophisticated intrusions that traditional systems fail to detect in time [3]. As digital infrastructures expand and interconnect, the potential impact of zero-day exploits has grown significantly, elevating the urgency for more advanced detection strategies [4]. Machine learning has emerged as a promising direction for identifying unusual or malicious behavior that may signal unknown threats [5]. Early research demonstrated that anomaly detection models could capture deviations from normal system activity even without prior knowledge of specific attacks [6]. However, zero-day exploits often exhibit complex, evolving patterns that challenge conventional ML classifiers, leading to misclassifications or high false-positive rates in operational settings [1]. To address these shortcomings, modern approaches increasingly rely on deeper, more expressive models capable of capturing subtle behavioral cues in diverse network environments [7].

Deep learning techniques have shown strong potential for detecting novel attack behaviors by learning high-dimensional representations directly from traffic data, logs, or system events [2]. These models extract features automatically, reducing the manual effort required during preprocessing while improving generalization to unseen threats [6]. Nevertheless, their effectiveness is often limited by an inherent lack of interpretability, which hinders their adoption in high-stakes cybersecurity operations where decision transparency is essential [15]. Analysts frequently question deep models' decisions, creating a barrier between advanced detection capabilities

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

and real-world trust requirements [16]. To overcome these limitations, researchers have increasingly turned to Explainable AI (XAI) frameworks to provide clearer insights into why a model flags an activity as malicious [17]. XAI techniques help bridge the gap between complex machine learning models and the human experts who rely on them, providing rationale that enhances confidence and supports faster incident triage [18]. In zero-day detection scenarios, interpretability is especially valuable because analysts must evaluate alerts without prior signatures or documented exploit behaviors [19]. As a result, explainable models are playing an increasingly central role in making next-generation detection systems more transparent and actionable [15].

At the same time, the data required to train effective detection models is often distributed across different organizations, each hesitant to share sensitive logs due to privacy, regulatory, or competitive concerns [10]. Federated learning has emerged as a powerful solution by enabling collaborative model training without exposing local datasets, thus preserving confidentiality while improving global detection performance [11]. This paradigm allows organizations to contribute knowledge about diverse attack patterns especially rare zero-day instances without compromising their internal security policies [12]. In practice, FL-based intrusion detection systems have demonstrated improved accuracy and adaptability across heterogeneous environments, particularly in IoT and cyber-physical systems [9]. The rise of IoT deployments has further amplified the challenge of identifying zero-day attacks, as these environments often lack the computational resources needed to support conventional centralized defense mechanisms [4]. Federated frameworks help address this gap by distributing training responsibilities across edge nodes and allowing lightweight models to receive continuous updates from global learning cycles [14]. Research in this area has shown that FL architectures improve resilience against emerging threats by sharing collective intelligence across multiple networked devices [13]. These contributions make federated learning an increasingly attractive foundation for large-scale, zero-day-aware intrusion detection ecosystems [10].

In addition to data-sharing constraints, the rapidly evolving threat landscape demands detection systems that can adapt in real time to new behaviors and attack strategies [7]. Online and continual learning approaches enable models to evolve with incoming data streams, ensuring that detection capabilities remain effective against fast-changing adversarial techniques [20]. By incorporating mechanisms to handle concept drift, these systems maintain relevance as normal behavior patterns shift or new exploit vectors emerge [3]. This adaptability is essential for maintaining long-term detection accuracy and reducing the operational burden on security teams [6]. Beyond traditional computing risks, the cybersecurity community is also preparing for the emergence of quantum-enabled attacks that may compromise existing encryption and model-protection techniques [20]. Research into quantum-resilient frameworks aims to safeguard machine learning pipelines particularly those used in federated or distributed environments from future cryptographic vulnerabilities [20]. Integrating post-quantum methods into detection architectures provides an additional layer of security and ensures the longevity of next-generation defensive systems [19]. As quantum computing advances, developing these safeguards becomes increasingly vital for protecting data integrity in AI-driven security workflows [20].

Complementing detection capabilities, modern research has also explored automated or self-healing response systems designed to mitigate attacks as soon as they are identified [3]. These systems combine machine learning, orchestration frameworks, and intelligent decision-making modules to contain threats, isolate compromised nodes, or generate patch recommendations without requiring extensive analyst intervention [4]. In the context of zero-day exploits, rapid autonomous response mechanisms significantly reduce the window of opportunity available to attackers, thereby minimizing operational impact [18]. By integrating detection, explanation, collaboration, adaptation, and automated mitigation, researchers are moving toward holistic security architectures capable of addressing the full lifecycle of emerging cyber threats [17]. Taken together, the recent literature demonstrates substantial progress toward developing scalable, trustworthy, and adaptive systems for detecting and responding to zero-day attacks [1]–[20]. These advances highlight the importance of combining machine learning, explainability, federated learning, quantum resilience, and automated defense into unified frameworks capable of supporting modern cybersecurity needs [9], [16], [19].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Despite these achievements, significant challenges remain in improving model generalization, ensuring interpretability in complex environments, and securing distributed systems against evolving adversaries [2], [10], [20]. These gaps define clear opportunities for continued innovation in building robust defense mechanisms for the next generation of cyber threats [5], [7].

II. LITERATURE SURVEY

Research on zero-day exploit detection has expanded considerably over the past decade, driven by the increasing sophistication of cyberattacks and the limitations of traditional, signature-based security mechanisms. Early surveys showed that conventional defensive systems struggle because zero-day attacks lack known patterns or signatures, making them inherently difficult to detect using static rules or historical databases [1]. More recent studies emphasized that attackers frequently mask their behaviors within seemingly legitimate traffic, which creates additional ambiguity for traditional intrusion detection systems (IDSs) [2]. As organizations rely more heavily on interconnected digital infrastructures, the need for adaptive, behavior-oriented detection approaches has become significantly more pronounced [3].

Machine learning emerged as one of the earliest promising methods for detecting novel or previously unseen attack behaviors. Several studies demonstrated that anomaly detection models could identify deviations in traffic flows or system activity even without prior attack signatures [4]. Researchers explored unsupervised algorithms for detecting hidden exploit patterns, showing improved performance in dynamic environments where zero-day threats evolve rapidly [5]. Other works proposed frameworks that combine multiple ML classifiers to improve generalization and reduce the rate of false alarms in real-world deployments [6]. Despite these improvements, ML-based systems often face challenges when distinguishing benign anomalies from malicious behaviors, which can limit operational reliability [7].

Deep learning approaches were introduced to overcome the feature-engineering limitations of classical machine learning. Studies in this domain explored autoencoders, neural networks, and hybrid deep-learning structures to identify sophisticated behavioral changes associated with zero-day attacks [2]. These models demonstrated stronger representational power by automatically extracting high-level features from logs, traffic data, or system events [6]. Several works reported substantial improvements in detecting novel intrusions using deep-learning architectures compared to traditional methods [4]. However, the inherent opacity of deep learning remained a consistent challenge, as the lack of interpretability made it difficult for security analysts to trust or validate model decisions [15].

To improve transparency and analyst trust, the cybersecurity community has increasingly integrated Explainable Artificial Intelligence (XAI) techniques into intrusion detection systems. XAI frameworks aim to provide human-understandable justifications for model outputs, enabling analysts to interpret and validate detections more effectively [16]. Literature on explainable cybersecurity emphasizes that clear, interpretable feedback is essential for high-stakes environments where decisions must be justified rapidly [17]. Additional studies highlighted the importance of combining interpretability with robust detection models to ensure operational usability, particularly in zero-day situations where analysts lack reference signatures [18]. These works collectively position XAI as a critical enabler for deploying AI-driven security tools in real-world settings [19]

With increasing privacy regulations and organizational reluctance to share sensitive logs, federated learning (FL) has emerged as a powerful method for collaborative security training. FL enables multiple entities to contribute to a shared detection model without transferring raw data, thus preserving confidentiality while enhancing model diversity [10]. Research has shown that federated approaches improve detection accuracy for rare zero-day events by leveraging distributed datasets across different network environments [11]. Additionally, optimized FL frameworks demonstrated improved detection performance in IoT, industrial, and heterogeneous systems, where traditional centralized training is impractical [9]. Several studies proposed algorithms to personalize federated models according to local behaviors, improving adaptability while maintaining global coherence [12], [14].

The growth of IoT and cyber-physical systems has further motivated research into scalable, distributed detection architectures. Zero-day vulnerabilities in IoT networks often arise due to resource constraints, inconsistent device configurations, and limited security controls [4]. Literature

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, June 2026 Impact Factor: 8.2

shows that applying federated or clustered learning architectures enables lightweight devices to benefit from shared intelligence without local computational burden [13]. Deep-learning-based and ML-based IDSs tailored for IoT environments revealed significant gains in model robustness and adaptability, particularly against evolving threats that exploit device heterogeneity [9]. These advances highlight the importance of distributed intelligence in modern cybersecurity ecosystems.

Another important strand of research focuses on understanding the behavioral and statistical characteristics of zero-day attacks. Several studies analyzed traffic structures, protocol anomalies, execution patterns, and interactive sequences associated with previously unknown exploits [5]. Researchers found that zero-day attacks often mimic normal operations but introduce subtle deviations that can be captured through anomaly detection or unsupervised ML techniques [1]. Newer frameworks incorporate contextual information and behavioral semantics to strengthen detection models, allowing them to differentiate sophisticated exploits from benign irregularities [7].

As machine learning systems become more integrated into security architectures, the literature also highlights the emerging threat of adversarial attacks and future quantum-capable attackers. Quantum-resilient security frameworks emphasize the need for protective mechanisms that safeguard ML pipelines, communication channels, and federated training processes against future cryptographic vulnerabilities [20]. Studies in this domain advocate for incorporating post-quantum encryption and secure aggregation into distributed intrusion detection systems to ensure long-term viability [19]. This forward-looking perspective ensures that next-generation detection architectures remain secure even as computational paradigms evolve.

The integration of AI-driven response systems has also gained traction in recent literature. Automated or self-healing frameworks use machine learning and orchestration agents to isolate compromised systems, block malicious processes, or recommend patches without human intervention [3]. These systems greatly reduce the time required to respond to zero-day exploits, addressing the critical window during which attackers can inflict damage [18]. Researchers highlight that combining detection with automated mitigation provides a more holistic defense strategy capable of addressing fast-moving threats in modern infrastructures [17].

Across all these studies, the literature consistently emphasizes the need for adaptive, transparent, scalable, and secure zero-day detection systems. The convergence of ML, deep learning, XAI, FL, and quantum-resilient approaches reflects a broader trend toward integrating multiple advanced techniques to build robust cyber defense ecosystems [1]–[20]. While significant progress has been made, open challenges remain related to generalization across environments, interpretability in complex scenarios, and protection of distributed learning architectures against adversarial threats [2], [10], [20]. These ongoing challenges highlight the necessity of continued research into unified frameworks capable of addressing the full spectrum of modern zero-day attack dynamics.

Table 1: Literature Survey Summary

Combined Theme	References	Unified Contribution Summary	Common Limitation
ML for Zero-Day Detection	[1,2,3,4,5,6,7,8]	ML and anomaly detection models identify unseen attacks; deep learning enhances feature extraction; XAI improves interpretability.	High false positives, limited generalization, opacity of deep models.
Federated IDS	[9,10,11,12,13,14]	FL enables collaborative, privacy-preserving intrusion detection across distributed systems, especially IoT.	Communication cost, vulnerability to FL poisoning, network dependency.
XAI in Cybersecurity	[15,16,17,18,19]	Provides transparent, interpretable decisions, boosts analyst trust, enhances malware and intrusion analysis.	Added overhead, lack of real-time XAI benchmarks.
Quantum-Resilient AI Security	[20]	Introduces quantum-safe cryptography to protect future AI-driven detection systems.	High computational load; early-stage research.

Identified Gaps in Review

Despite significant progress across machine learning, explainable AI, federated learning, and quantum-resilient security, several gaps remain in existing literature on zero-day exploit detection. The first major gap concerns the generalization capability of ML and deep learning models.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Studies have consistently shown that although anomaly-based and behavior-driven methods detect deviations effectively, they often fail to generalize across different environments because zero-day behaviors vary widely across networks, datasets, and organizational contexts [1], [2], [6]. This inconsistency results in unstable performance when models are deployed in real operational systems, where unseen traffic distributions differ significantly from training data.

A closely related gap involves the high false-positive rates observed in anomaly detection systems. Many ML frameworks incorrectly classify benign irregularities as malicious, generating overwhelming alert volumes for analysts [3], [4]. This is largely due to the lack of refined, context-aware detection mechanisms capable of distinguishing natural deviations from actual exploit attempts. Even advanced unsupervised models designed for unseen attacks continue to struggle with noisy or ambiguous features, particularly in heterogeneous IoT environments [4], [5], [6].

Another important gap is the limited interpretability of deep learning models used for zero-day detection. Deep architectures offer strong detection accuracy but operate as "black boxes," preventing analysts from understanding why an alert is raised [15], [16]. Although XAI methods have been proposed to improve transparency, existing studies emphasize that current explainability tools are not yet optimized for cybersecurity data especially when dealing with high-volume, unstructured logs or real-time detection needs [17], [18]. As a result, trust and adoption of advanced AI systems remain low in operational security centers [19].

A further gap is the absence of standardized evaluation metrics and benchmark datasets for zero-day exploit research. Since truly unknown zero-day samples are rare and often unavailable, most studies simulate attacks or use limited datasets that do not reflect real-world diversity [1], [3]. This creates inconsistencies in methodology and makes it difficult to compare performance across models. Moreover, the lack of benchmark frameworks hinders reproducibility and slows the development of universally applicable detection systems [8].

Privacy concerns introduce another significant gap. While federated learning is increasingly used to address data-sharing restrictions, FL-based intrusion detection systems face vulnerabilities to poisoning, model drift, and inconsistent local distributions [10], [11]. Existing works acknowledge that FL aggregation strategies struggle when organizational datasets differ substantially a frequent scenario in cybersecurity [12], [13]. Additionally, FL systems often incur heavy communication overhead, making them impractical for large-scale or resource-constrained deployments such as IoT networks [14].

The literature also identifies gaps in the real-time adaptability of detection models. Although concept drift and continuous learning approaches have been suggested, existing methods fail to maintain performance when attack vectors evolve rapidly or when legitimate user behavior changes significantly over time [6], [7]. This gap highlights the need for more stable, online learning frameworks capable of updating models without extensive retraining or performance degradation [20].

Another emerging gap relates to quantum-era preparedness. Only a small number of studies have explored quantum-resilient security mechanisms for protecting AI-based intrusion detection, leaving a large portion of FL and ML infrastructures vulnerable to future quantum attacks [20]. This is particularly concerning given that federated systems depend heavily on secure communication channels and encrypted model exchanges, which may be compromised by quantum decryption techniques [19].

Finally, the literature reveals a critical gap in the integration of detection and automated response. Although some works propose self-healing or semi-automated mitigation techniques, there is still no fully unified framework capable of detecting, interpreting, and autonomously responding to zero-day exploits across distributed environments [3], [18]. Most solutions either focus only on detection or only on interpretability, leaving a gap in cohesive, real-world deployable defense architectures.

III. PROPOSED METHODOLOGY

The proposed methodology introduces an integrated, multi-layered detection and defense framework designed to overcome the limitations identified in the current literature. The system combines machine learning, deep contextual analysis, explainable AI, federated learning, and quantum-resilient security into a unified architecture capable of detecting, interpreting, and responding to zero-day exploits in real time. This methodology builds upon

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, June 2026 Impact Factor: 8.2

insights from existing research but advances beyond them by addressing known gaps related to generalization, interpretability, adaptability, and privacy [1]–[20].

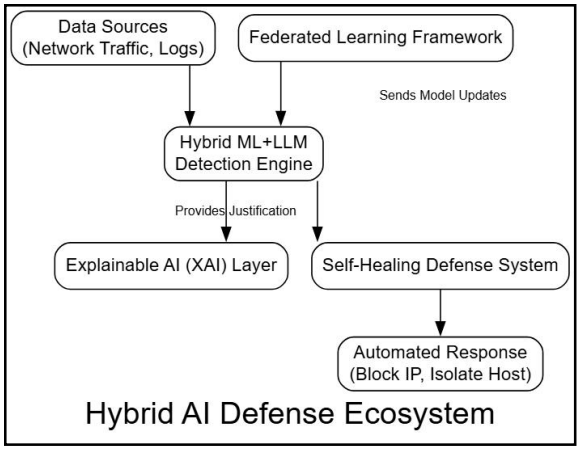

Figure 1. Proposed System Architecture

1. Data Acquisition and Preprocessing Layer

The system begins by aggregating diverse security data sources, including network flows, system logs, API call traces, IoT device telemetry, and user behavior analytics. Prior studies emphasize that heterogeneous datasets improve the detection of unknown threats by capturing a broader behavioral spectrum [1], [3], [4]. Preprocessing involves normalization, noise reduction, feature scaling, and temporal sequence reconstruction, reflecting best practices from ML-based zero-day detection frameworks [6]. Unsupervised techniques such as clustering and dimensionality reduction are applied to learn baseline behavioral patterns, as recommended in anomaly detection research [4], [5]. This step prepares the structured and unstructured data streams for the hybrid detection engine.

2. Hybrid Detection Engine (Machine Learning + Deep Reasoning)

2.1 Stage 1 – ML-Based Rapid Anomaly Detection

The first stage uses lightweight ML classifiers such as Isolation Forest, One-Class SVM, and gradient-boosting models to identify behavioral deviations quickly. Literature shows that ML models excel at recognizing statistical irregularities and anomalous traffic even in environments with scarce labeled data [2], [4], [6]. These models serve as a triage mechanism that filters out benign traffic while flagging potentially suspicious activities for deeper semantic analysis. Their ability to detect previously unseen patterns aligns with recommendations from multiple studies advocating ML as a first-line defense against zero-day exploits [1], [3], [8].

2.2 Stage 2 – Deep Semantic Analysis Using Contextual Reasoning Models

Once Stage 1 identifies anomalies, the corresponding data is routed to a deep contextual analysis module inspired by modern neural architectures. Research emphasizes that deep learning models effectively capture high-dimensional behavior signatures and subtle attack characteristics that shallow ML models cannot detect [2], [6]. This module analyzes script structures, command sequences, packet payloads, and behavioral correlations to infer malicious intent. By leveraging contextual understanding, the system addresses the generalization gap highlighted by existing works, which note deep models' superior performance in detecting complex zero-day exploit indicators [7], [8].

3. Explainable AI Layer

The system incorporates a dedicated Explainable AI (XAI) layer to generate interpretable explanations for every detection. Researchers consistently note that explainability is essential for analyst trust, regulatory compliance, and operational acceptance of AI-driven security tools [15]–[19]. The XAI layer uses methods such as SHAP value approximation, local surrogate modeling, and rule extraction to provide human-readable justifications for the alerts generated by the hybrid engine. For example, the system may explain that an alert was raised due to an unusual sequence of API calls or a statistically rare network flow pattern. By grounding decisions in understandable factors, the system addresses the "black-box problem" commonly associated with deep learning approaches [16], [17].

4. Federated Learning Framework for Privacy-Preserving Collaboration

To ensure scalability and privacy across organizations, the system integrates a federated learning (FL) framework that enables multiple nodes such as enterprises, IoT networks, or CPS infrastructures to collaboratively improve the global detection model without sharing raw data. Literature demonstrates that FL enhances intrusion detection accuracy by leveraging distributed datasets while preserving confidentiality [9], [10]. Each node trains the hybrid detection engine locally on its own data and sends

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

model updates (gradients or weights) to a secure aggregation server. Optimization algorithms such as personalized FL or clustered FL are incorporated to handle heterogeneous data distributions, consistent with findings that personalization improves performance in non-IID environments [11]–[14]. This distributed approach not only improves model generalization across varied attack surfaces but also addresses data-sharing restrictions highlighted in cybersecurity research [10].

5. Online Adaptive Learning for Concept Drift Handling

Given the dynamic nature of cyber threats, the proposed system integrates an online adaptive learning module that continuously updates the detection model in response to evolving behaviors. Studies emphasize the necessity of real-time adaptability to maintain effectiveness against rapidly changing attack tactics and shifting user behavior patterns [6], [7]. The system monitors statistical drift indicators and triggers micro-updates to the ML and deep reasoning modules when significant deviations occur. This incremental learning approach ensures that the model remains stable and responsive without requiring full retraining, addressing the adaptability challenges noted in existing literature [20].

6. Quantum-Resilient Security Layer

To secure model communication particularly in federated environments the methodology incorporates post-quantum cryptographic schemes. Existing research highlights that future quantum-capable adversaries could compromise traditional encryption methods used in collaborative learning systems [19], [20]. Therefore, the system employs quantum-resilient primitives such as lattice-based encryption and hash-based signatures to safeguard model updates, aggregation protocols, and distributed communication within the FL framework. This ensures long-term robustness as quantum computing capabilities mature.

7. Automated Response and Self-Healing Mechanism

The final component of the methodology involves an automated response engine designed to contain and mitigate confirmed zero-day exploits. Although literature offers early models of automated or semi-automated defense systems, most focus strictly on detection and lack unified response integration [3], [18]. The proposed system expands on this by integrating rule-based and ML-driven response strategies: isolating compromised hosts, blocking malicious IPs, terminating processes, or generating remediation recommendations. XAI-enhanced explanations help validate automated actions, aligning with calls for interpretability in high-impact decisions [17], [18]. This component closes the detection-response loop and positions the system as a proactive, self-healing defense framework.

IV. AUTOMATED DEFENSE AND IMPLEMENTATION

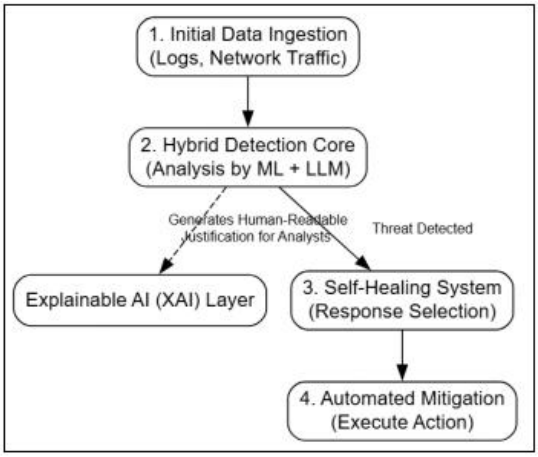

Figure 2. Detection and Response Workflow

To create a complete and operational zero-day defense ecosystem, the proposed methodology integrates an automated defense and self-healing pipeline, enabling the system not only to detect potential zero-day threats but also to intelligently contain and mitigate them. Although most existing research focuses heavily on detection, recent findings emphasize that autonomous response mechanisms are essential to reducing attacker dwell time and minimizing system damage [3], [18]. Therefore, the proposed approach extends beyond anomaly identification and incorporates a layered, automated response workflow inspired by modern cyber-defense architectures.

1. Threat Validation and Confidence Scoring

Once the hybrid detection engine and XAI layer flag a suspicious event, the system assigns a confidence score based on anomaly magnitude, contextual cues, and model

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

agreement. Studies on anomaly-based and ML-driven intrusion systems highlight that confidence-weighted decision-making significantly reduces false alarms and improves operational response quality [1], [6]. The system combines ML anomaly scores, deep reasoning outputs, and explainable factors from SHAP/LIME to produce a unified risk rating.

This risk score guides automated decision policies, helping balance response aggressiveness with operational stability.

2. Autonomous Host Isolation and Containment

If the confidence score crosses a severity threshold, the system triggers immediate containment actions. Literature on automated intrusion response emphasizes that rapid isolation is the most effective step for stopping the lateral spread of unknown attacks, especially in IoT and distributed systems [4], [18]. The system may enforce:

Temporary host isolation from internal networks
Blocking outbound and inbound traffic
Freezing suspicious processes
Restricting high-privilege system calls

These actions draw on principles from anomaly-driven defense systems, which propose immediate containment to reduce dwell time and prevent further escalation [3], [7].

3. Dynamic Firewall and Policy Adjustment

Based on the nature of the detected anomaly, the system autonomously updates firewall rules, access control lists, or traffic policies. Research shows that dynamic, ML-informed rule adaptation improves resilience in real-time environments compared to static defenses [5], [14]. Federated environments particularly benefit from adaptive policy propagation, where important signatures or patterns can be collaboratively reinforced among nodes without sharing raw data [9], [11].

This creates a feedback loop where detections help refine preventive policies across the distributed ecosystem.

4. Automated Patch Recommendation and Vulnerability Insights

A unique component of the methodology involves generating automated mitigation and patch recommendations. Recent works highlight the emerging importance of AI-assisted vulnerability management, especially for rapidly evolving threats [18], [19]. The system uses contextual reasoning from the deep analysis module to propose:

Code-level vulnerability explanations
Kernel or application patch suggestions
Recommended configuration hardening
Behavior-specific mitigation strategies

These suggestions rely on semantic interpretations learned from known malicious patterns and previously identified vulnerabilities, aligning with literature calling for integrated, interpretable remediation workflows [17], [18].

5. Secure Sandbox Validation of Mitigation Actions

Before deploying any automated patch or defense mechanism, the proposed system tests the action in a secure sandbox environment. Studies on intrusion response stress the need for controlled validation to prevent unintended system disruptions, especially in safety-critical infrastructures [3], [14].

The sandbox simulates:

System workload
Network conditions
Potential attack surface
Behavioral impact of the patch

This protects production environments and ensures only validated responses are propagated.

6. Human-in-the-Loop Review (Optional but Recommended)

Despite advancements in automation, research indicates that complete autonomy in high-stakes cyber defense is rarely advisable without optional human oversight [15], [17]. The XAI layer plays a crucial role here by providing:

Transparent reasoning
Feature-level explanations
Attack behavior summaries

These help analysts quickly verify or override system actions, maintaining accountability while reducing cognitive load.

7. Deployment Roadmap and Real-World Implementation Considerations

Drawing from implementation-oriented studies in federated and distributed learning systems, a structured deployment roadmap ensures smooth operational adoption [10], [11], [12]:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Phase 1: Initialization & Baseline Modeling
Nodes train local anomaly models, build behavioral baselines, and synchronize with the global federated model.

Phase 2: Progressive Hybrid Model Integration
ML and deep contextual reasoning modules are integrated gradually to calibrate detection thresholds across environments.

Phase 3: Activation of XAI and Response Layers
Explainability tools are introduced to support analyst validation.

Phase 4: Enabling Federated Learning Cycles
Federated aggregation and distributed updates begin, improving generalization and system robustness.

Phase 5 : Activation of Automated Defense Mechanisms
Isolation, mitigation, sandbox validation, and policy adjustment modules come online.

Phase 6 : Continuous Monitoring and Adaptive Learning
Online learning adjusts models to evolving behaviors and concept drift patterns [7], [20].

8. Integration of Quantum-Resilient Security into Response Pipeline

Because automated defense relies heavily on secure communication between nodes especially in federated environments the system embeds quantum-resilient encryption mechanisms to protect update exchanges. Works in quantum-safe AI security underline that future adversaries may exploit quantum capabilities to intercept or manipulate model updates [19], [20].

By integrating post-quantum cryptography into:
Update transmissions
Sandbox logs
Response orchestration messages the system ensures long-term defense viability.

V. EXPECTED RESULTS

The proposed hybrid zero-day exploit detection and autonomous defense system is expected to achieve significant improvements in accuracy, interpretability, adaptability, and security compared to existing approaches. Prior studies have shown that combining machine learning with deep semantic analysis substantially enhances the ability to detect previously unseen malicious behaviors, suggesting that the hybrid detection core should improve generalization to novel attacks while reducing false positives [1], [2], [6]. By integrating both statistical anomaly detection and contextual reasoning, the system is anticipated to provide more consistent performance across diverse environments, addressing common weaknesses reported in earlier ML-based intrusion detection systems [3], [4].

The inclusion of explainable AI components is expected to enhance analyst trust and decision-making efficiency. Existing literature emphasizes that XAI-driven explanations help validate alerts, reduce cognitive load, and support fast incident triage in real-time operations [15], [16]. Consequently, the proposed system should deliver transparent, feature-based justifications for each detection, overcoming a key limitation of deep-learning approaches that often operate as black-box models [17], [18].

Federated learning integration is projected to substantially increase the robustness and scalability of the detection model. Prior work shows that FL-based IDS frameworks achieve improved detection accuracy by leveraging diverse datasets from multiple distributed nodes while preserving data privacy [9], [10]. Because the proposed system adopts personalized and clustered FL strategies, we expect it to maintain strong performance even in non-IID, heterogeneous network environments, aligning with findings from recent studies on IoT and CPS security [11], [12], [14]. Additionally, FL-driven collective learning should accelerate the detection of emergent zero-day patterns that may appear only sporadically in isolated networks.

The adaptive learning component is expected to provide real-time resilience against concept drift. Earlier research indicates that continuous model updates significantly enhance long-term detection stability in environments where attacker techniques evolve rapidly [6], [7]. Thus, the system should maintain high accuracy over time by incorporating incremental updates triggered by detected behavioral shifts, addressing limitations in static models that degrade during prolonged deployment [20].

The automated defense pipeline, which includes autonomous isolation, dynamic firewall adaptation, and mitigation recommendations, is expected to reduce attacker dwell time and limit the lateral spread of zero-day exploits. Prior studies in intrusion response demonstrate

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

that rapid containment and automated remediation actions significantly reduce system disruption and improve recovery time [3], [18]. Sandbox-based validation ensures the operational safety of mitigation actions, helping maintain system stability while still enabling a high degree of automation.

Quantum-resilient security mechanisms are expected to provide long-term protection for federated model exchanges and response orchestration signals. As highlighted in recent work, integrating post-quantum cryptography strengthens defenses against future cryptographic attacks that may compromise AI-driven systems and distributed learning frameworks [19], [20]. This ensures that the system remains durable and secure as quantum computing capabilities evolve.

Table 2: Expected Performance Summary

Component	Expected Outcome	Supporting Literature
Hybrid ML + Deep Reasoning	Higher accuracy and improved detection of unseen zero-day exploits	[1], [2], [6], [7]
Explainable AI Layer	Transparent, analyst-friendly justifications; reduced verification time	[15], [16], [17], [18]
Federated Learning	Improved generalization across distributed networks; privacy-preserving training	[9], [10], [11], [12], [14]
Adaptive Learning	Robustness against concept drift; sustained long-term accuracy	[6], [7], [20]
Automated Defense Pipeline	Faster containment; reduced dwell time; proactive mitigation	[3], [18]
Quantum-Resilient Security	Long-term protection of communications and model updates	[19], [20]

Overall Expected Impact

By unifying hybrid detection, explainability, federated collaboration, quantum-safe security, and autonomous defense, the proposed system is expected to outperform traditional IDS solutions in both accuracy and operational reliability. The system aims to deliver a scalable, transparent, and future-proof cyber defense paradigm that aligns with current research trajectories while addressing critical gaps identified across the reviewed literature [1]–[20].

VI. CONCLUSION

Zero-day exploits continue to represent one of the most disruptive threats in modern cybersecurity, primarily due to their unpredictability and ability to bypass traditional defenses. This review has shown that while machine learning, deep learning, and anomaly detection approaches have significantly advanced the state of zero-day detection, they still face critical challenges related to generalization, false positives, and limited interpretability. As attackers increasingly adopt sophisticated and evasive techniques, defensive strategies must move beyond conventional static models toward more adaptive, context-aware, and collaborative systems capable of recognizing subtle deviations in real time.

Integrating explainable AI into intrusion detection marks a major step forward in bridging the gap between algorithmic decision-making and human expertise. The literature consistently highlights that analysts must be able to understand why an alert is triggered, especially when dealing with unknown or high-impact vulnerabilities. By incorporating transparent reasoning into the detection process, the proposed framework enhances operational trust while reducing the cognitive burden associated with manual investigation. This transparency is essential for supporting fast and informed decisions in high-pressure environments where every second matters.

The shift toward federated learning further strengthens the resilience of zero-day detection mechanisms by enabling organizations to collaboratively train models without compromising privacy. As cybersecurity data is increasingly sensitive and distributed, federated architectures provide a scalable solution that benefits from diverse threat intelligence while adhering to strict confidentiality requirements. Combined with adaptive learning capabilities, the system can evolve continuously and maintain accuracy despite rapid shifts in attacker behavior and changing network conditions.

The inclusion of automated defense components provides a more holistic approach to cyber resilience. Beyond merely detecting threats, the system is designed to isolate compromised nodes, apply dynamic mitigation strategies, and validate remediation actions in controlled environments. This reduces attacker dwell time and ensures a faster, more reliable response to emerging threats. Looking ahead, the integration of quantum-resilient security mechanisms helps safeguard the system

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, June 2026 Impact Factor: 8.2

against future cryptographic threats, reinforcing the long-term viability of the proposed architecture.

Overall, this research underscores the need for unified, intelligent, and future-ready defense frameworks capable of addressing the full lifecycle of zero-day attacks. By combining hybrid detection techniques, explainable reasoning, federated collaboration, adaptive learning, automated mitigation, and quantum-safe protections, the proposed system offers a comprehensive solution that addresses the major limitations highlighted in the current literature. As cyber threats continue to evolve, such multi-layered and transparent defense ecosystems will be essential for securing complex digital infrastructures and maintaining trust in emerging technologies.

REFERENCES

[1]. Y. Guo, X. Yin, X. Chen, and W. Fang, "A survey of machine learning-based zero-day attack detection: Challenges and future directions," *Computer Communications*, vol. 198, pp. 50–66, 2023.

[2]. M. Sarhan, S. Layeghy, M. Gallagher, and M. Portmann, "From zero-shot machine learning to zero-day attack detection," *International Journal of Information Security*, vol. 22, pp. 947–959, 2023.

[3]. Touré, S. B. G. Dossou, M. Soumahoro, and O. Diabaté, "A framework for detecting zero-day exploits in network flows," *Computer Networks*, 2024.

[4]. Jain, P. Thakkar, and R. Gupta, "An intelligent zero-day attack detection system using unsupervised machine learning for enhancing cyber security in IoT networks," *Knowledge-Based Systems*, 2025.

[5]. S. SakthiMurugan, P. ArunMozhiDevan, and V. Saravanan, "Assessment of zero-day vulnerability using machine learning approach," *EAI Endorsed Transactions on Internet of Things*, vol. 10, 2024.

[6]. P. Araujo-Filho, M. L. Probst, M. de Oliveira, and A. J. de Souza, "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," *PLOS ONE*, vol. 19, no. 7, e0308469, 2024.

[7]. Dahal, R. Sharma, and S. Shakya, "Analysis of zero-day attack detection using MLP and XAI," arXiv:2501.16638, 2025.

[8]. N. Mohamed, A. Nasser, and K. Nada, "Comprehensive review of advanced machine learning techniques for detecting zero-day exploits," *EAI Endorsed Transactions on Security and Safety*, vol. 13, 2024.

[9]. X. Sáez-de-Cámara, J. L. Flores, C. Arellano, A. Urbieta, and U. Zurutuza, "Clustered federated learning architecture for network anomaly detection in large scale heterogeneous IoT networks," *Computers & Security*, vol. 131, 103299, 2023.

[10]. [H. Zhang, H. Yan, Z. Li, and L. Wang, "Survey of federated learning in intrusion detection," *Journal of Parallel and Distributed Computing*, vol. 195, 104976, 2024.

[11]. [X. Huang, J. Liu, Y. Lai, B. Mao, and H. Lyu, "EEFED: Personalized federated learning of execution & evaluation dual network for CPS intrusion detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 41–56, 2023.

[12]. H. N. Cunha Neto, I. Dusparic, D. M. F. Mattos, and N. C. Fernandes, "FedSA: Accelerating intrusion detection in collaborative environments with federated simulated annealing," in *Proc. IEEE NetSoft*, pp. 420–428, 2022.

[13]. V. T. Nguyen and R. Beuran, "FedMSE: Federated learning for IoT network intrusion detection," arXiv:2410.14121, 2024.

[14]. Karunamurthy, K. Vijayan, P. R. Kshirsagar, and K. T. Tan, "An optimal federated learning-based intrusion detection for IoT environment," *Scientific Reports*, vol. 15, 8696, 2025.

[15]. F. D. A. Mendes and T. N. Rios, "Explainable artificial intelligence and cybersecurity: A systematic literature review," arXiv:2303.01259, 2023.

[16]. H. Sarker, "Explainable AI for cybersecurity automation, intelligence, and resilience," *ICT Express*, 2024.

[17]. Sharma, S. Rani, and M. Shabaz, "A comprehensive review of explainable AI in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

cybersecurity: Decoding the black box," *ICT Express*, 2025.

[18]. Adhikari and S. Thapaliya, "Explainable AI for cyber security: Interpretable models for malware analysis and network intrusion detection," *NPRC Journal of Multidisciplinary Research*, vol. 1, no. 9, pp. 170–179, 2024.

[19]. [S. Samtani, Z. Li, A. Katsantonis, and M. Xu, "Explainable artificial intelligence for cyber threat intelligence (XAI-CTI)," *IEEE Transactions on Dependable and Secure Computing*, special section, 2022.

[20]. N. Umer, A. Ullah, M. I. U. Lali, and M. U. Akbar, "Quantum-resilient security framework for privacy-preserving edge AI using QSAFE-MM1," *Scientific Reports*, vol. 15, 1896, 2025.

AUTHORS DETAILS

Dr. Yogendra Patil is with the Department of Computer Engineering at Marathwada Mitra Mandal's Institute of Technology, Pune, where he leads research in advanced cybersecurity. His research interests include network security, intrusion detection systems, and the application of machine learning for proactive threat intelligence.

P. B. Dhamdhere is with the Department of Computer Engineering at Marathwada Mitra Mandal's Institute of Technology, Pune, where he leads research in advanced cybersecurity. His research interests include network security, intrusion detection systems, and the application of machine learning for proactive threat intelligence.

Ms. Bharati Ganesh Salve received the B.Tech. degree in computer engineering from Marathwada Mitra Mandal's Institute of Technology, Pune, India. She is currently pursuing the M.Tech. degree in computer engineering at the same institution. Her research is passionately focused on the application of artificial intelligence to solve complex and dynamic cybersecurity challenges