Advanced Machine Learning Approaches for Zero-Day Attack Detection: A Review

Fatema El Husseini[1], Hassan Noura[2], Ola Salman[3], and Ali Chehab[4]

[1]LISTIC – Polytech Annecy-Chambéry, Université Savoie Mont Blanc, France
[2]Univ. Franche-Comté (UFC), FEMTO-ST Institute, CNRS, Belfort, France
[3]DeepVU, USA
[4]American University of Beirut, Electrical and Computer Engineering

2024 8th Cyber Security in Networking Conference (CSNet) | 979-8-3315-3410-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSNET64211.2024.10851751

***Abstract*—Zero-day attacks provide an essential challenge in cybersecurity because of their unpredictability and absence of pre-existing defenses. To detect these threats, this paper thoroughly examines machine learning (ML) and artificial intelligence (AI) methodologies, encompassing supervised, unsupervised, and hybrid models. It underscores the capabilities of modern AI technologies, including deep learning, federated learning, and lightweight AI models, especially in real-time detection and resource-constrained environments. The research highlights the considerable deficiencies in the availability and uniformity of zero-day datasets, discusses the advantages and limitations of ML-based detection methods, and proposes directions for future inquiry, such as adversarial learning, privacy-preserving strategies, and the enhancement of real-time detection. The results intend to assist researchers and practitioners in formulating more resilient, scalable approaches to address zero-day vulnerabilities.**

***Keywords*—** Z ero-day attacks; supervised learning; unsupervised learning; hybrid learning; deep learning; federated learning; adversarial machine learning; lightweight AI models; real-time detection; resource-constrained environments; zero-day datasets; privacy-preserving methodologies; security strategies.

I. Introduction

Zero-day attacks exploit undisclosed vulnerabilities, making traditional signature-based detection methods ineffective against these novel threats. Machine learning (ML) and artificial intelligence (AI) have demonstrated their capabilities in examining patterns and anomalies in extensive datasets without relying on predetermined signatures. This study examines the efficacy of diverse machine learning techniques, encompassing supervised, unsupervised, and hybrid models, in conjunction with advanced artificial intelligence technologies such as federated learning and lightweight models tailored for resource-constrained contexts like the Internet of Things. Notwithstanding their potential, these technologies encounter problems such as elevated computing expenses, the necessity for real-time identification, and the absence of complete, current datasets. This paper enhances the construction of more resilient and scalable security systems for identifying and avoiding zero-day vulnerabilities by tackling these concerns and integrating adversarial resilience and explainable AI (XAI).

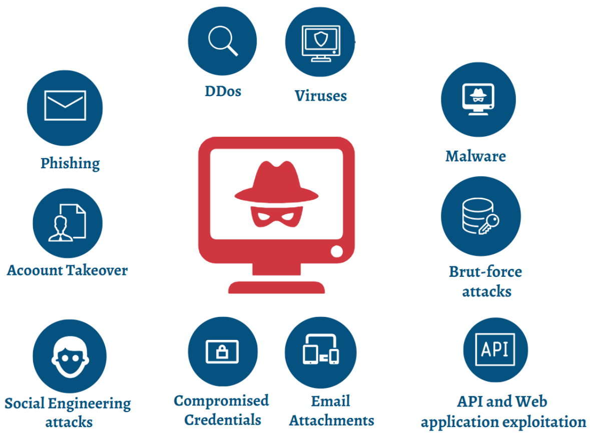

Fig. 1. Common Types of Attacks Vectors.

A. Problem Formulation

Zero-day attacks pose a significant issue in cybersecurity by exploiting undisclosed vulnerabilities that lack fixes or detection mechanisms, rendering conventional security measures ineffective. This problem is especially concerning for critical infrastructure, IoT devices, and distributed systems such as cloud and edge networks, where these attacks may remain undiscovered and cause considerable damage. The escalating complexity and interdependence of systems augment their susceptibility. Our research aims to utilize sophisticated machine learning and AI models, such as anomaly detection, federated learning, and lightweight AI frameworks, to identify zero-day attacks in real-time, particularly in resource-limited settings, while maintaining a balance between accuracy and computational efficiency.

B. Motivation

The escalation of cyber risks, notably zero-day attacks, has underscored the necessity for advanced detection techniques, as traditional methods are becoming progressively inadequate, particularly in real-time contexts. This paper addresses the disparity between advanced attacks and existing detection capabilities, emphasizing the utilization of AI and ML to identify previously unrecognized threats. AI-driven models, encompassing lightweight and federated learning methodologies, provide scalable and efficient solutions for resource-limited contexts such as IoT. The study seeks to assess and categorize AI-driven zero-day detection techniques, tackling existing difficulties and prospective opportunities in this advancing domain.

C. Contributions

This paper presents many contributions in detecting zero-day attacks, emphasizing AI and ML-based methodologies.

1) **Comprehensive Review of AI-Based Zero-Day Detection Techniques:** It offers a comprehensive analysis and categorization of contemporary machine learning (ML) and deep learning (DL) methodologies employed for the detection of zero-day attacks.
2) **Analysis of Emerging AI Technologies:** The paper examines new AI technologies such as federated learning, reinforcement learning, and lightweight AI models. It evaluates their efficacy in mitigating zero-day attacks, especially in resource-constrained environments.
3) **Evaluation of Zero-Day Attack Datasets:** It summarizes current datasets utilized for zero-day attack detection.
4) **Introduction of Lightweight AI Models for Zero-Day Detection:** A significant contribution involves introducing and assessing lightweight AI models tailored for resource-limited settings. It illustrates the significance of optimizing AI models for real-time detection in environments with constrained computational resources.
5) **Challenges and Future Directions:** It outlines the principal obstacles related to zero-day attack detection with AI and ML techniques. Furthermore, it proposes future research avenues that emphasize the necessity for explainable AI (XAI), privacy-preserving methodologies, and adaptive learning frameworks to address the dynamic characteristics of cyber threats.

D. Organization

The rest of this paper is organized as follows: Section III encompasses the background and preliminaries, detailing essential concepts related to zero-day attacks and the AI/ML algorithms employed in their detection.Section IV analyzes Zero-Day Attack Detection Techniques, exploring the various methods and approaches used to detect zero-day vulnerabilities, ranging from traditional to more advanced AI-driven approaches.Section V explores Emerging AI-Technology for Zero-Day Attacks, with a comparison between Deep Learning and Traditional Machine Learning. Section VI presents an overview of the Existing Zero-Day Datasets, discussing the datasets used for training and evaluating machine learning models in zero-day attack detection. Section VII evaluates the advantages, limitations, and challenges of existing solutions. Section VIII provides the lessons learned, suggestions, and recommendations from the paper, offering insights into the potential improvements and future directions in zero-day attack detection. Finally, Section IX encapsulates the principal results of the paper and proposes avenues for subsequent research.

II. Related Work

Zero-day attack detection has progressively concentrated on using machine learning (ML) and artificial intelligence (AI) methodologies to overcome the limitations of traditional approaches. While effective for recognized attacks, signature-based detection cannot address zero-day exploits due to the absence of established patterns. Anomaly detection has been thoroughly researched but it is plagued by elevated false positive rates in dynamic settings. Machine learning, especially supervised techniques such as decision trees and support vector machines, demonstrates potential but encounters difficulties in generalizing to novel attack vectors, as evidenced by studies like [1]. Unsupervised learning, utilizing models like autoencoders, identifies anomalies in unlabeled data but frequently necessitates fine-tuning and is susceptible to false positives. Deep learning models, such as CNNs and LSTMs, have exhibited superiority in processing large-scale, high-dimensional data, surpassing conventional methods. Nonetheless, their substantial computational expense restricts their utilization in real-time or resource-limited settings. Hybrid methodologies, exemplified by [2], integrate supervised and unsupervised learning to enhance detection precision, yet they necessitate substantial feature engineering, which constrains flexibility. As emphasized by [3], Federated learning mitigates privacy issues by facilitating decentralized detection without transmitting sensitive information; nonetheless, communication overhead and model variability present obstacles. Lightweight AI models, such as MobileNet, diminish processing requirements yet encounter accuracy constraints for intricate attacks. Our paper enhances current methodologies by optimizing lightweight models using federated learning to balance detection accuracy and efficiency, especially in resource-constrained situations like the Internet of Things (IoT). Furthermore, the paper tackles challenges like scalability, explainability, and privacy, intending to develop a more flexible and scalable system for zero-day attack detection.

III. Background & Preliminaries

Zero-day attacks represent a critical cybersecurity threat, leveraging undisclosed, undetected vulnerabilities. Advanced technologies, including artificial intelligence (AI) and machine learning (ML), play a vital role in identifying these complex threats by analyzing patterns and anomalies within data. This section delineates the fundamental attributes of zero-day attacks and specifies the AI/ML algorithms frequently employed for their detection.

A. Zero-Day Attacks: Properties, Characteristics and Specialties

Zero-day attacks pose a significant cybersecurity risk by exploiting unpatched vulnerabilities, enabling attackers to function without detection; below are the key properties, characteristics, and specialties that define zero-day attacks [1]:

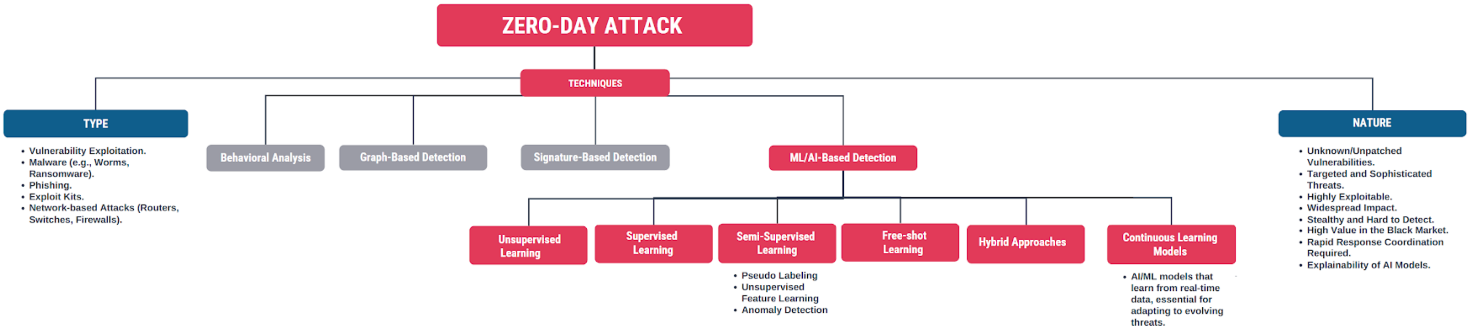
ZERO-DAY ATTACK
TECHNIQUES
TYPE
• Vulnerability Exploitation.
• Malware (e.g., Worms, Ransomware).
• Phishing.
• Exploit Kits.
• Network-based Attacks (Routers, Switches, Firewalls).
Behavioral Analysis
Graph-Based Detection
Signature-Based Detection
ML/AI-Based Detection
Unsupervised Learning
Supervised Learning
Semi-Supervised Learning
• Pseudo Labeling
• Unsupervised Feature Learning
• Anomaly Detection
Free-shot Learning
Hybrid Approaches
Continuous Learning Models
• AI/ML models that learn from real-time data, essential for adapting to evolving threats.
NATURE
• Unknown/Unpatched Vulnerabilities.
• Targeted and Sophisticated Threats.
• Highly Exploitable.
• Widespread Impact.
• Stealthy and Hard to Detect.
• High Value in the Black Market.
• Rapid Response Coordination Required.
• Explainability of AI Models.

Fig. 2. Zero-Day Attack Taxonomy.

1) **Exploitation of Unknown Vulnerabilities:** Zero-day attacks take advantage of unpatched vulnerabilities that developers haven't yet identified, making them challenging to address due to the lack of existing defenses.
2) **Absence of Signatures:** Because zero-day exploits do not have predefined signatures, they are difficult to find using traditional security measures like firewalls and antivirus software that rely on well-known patterns.
3) **High Stealth and Sophistication:** These attacks use advanced polymorphism and code obfuscation to stay hidden for extended periods, making detection challenging until vulnerabilities are discovered and patched.
4) **Targeted and Highly Selective:** Zero-day attacks frequently target particular high-value organizations or vital infrastructure to cause the most disruption or steal the most data from the targeted operations.
5) **Widespread Impact:** When a widely used software vulnerability is exploited, it can compromise millions of systems, causing significant damage before the patch is applied.
6) **Short Window of Exploitation:** To maximize damage, attackers using zero-day attacks must maximize, since they have a limited opportunity to exploit vulnerabilities before they are patched.
7) **Advanced Evasion Techniques:** Attackers lack detection by using complex techniques like encryption and polymorphic malware, which makes it more difficult to recognize and neutralize the threat.
8) **Critical for High-Value Targets:** Because of their effectiveness against high-value targets, zero-day attacks are highly valuable in cyberwarfare and espionage and are frequently sold for high sums on the black market.

B. AI/ML Algorithms

Artificial intelligence (AI) and machine learning (ML) techniques are essential in identifying and mitigating zero-day threats. These algorithms can detect abnormalities and new risks by analyzing trends from past data, even without specified signatures or patches for the exploited vulnerabilities. This section examines the principal AI/ML techniques for detecting zero-day attacks.

1) **Supervised Learning Algorithms:**Supervised learning algorithms, including Decision Trees, Random Forests, and Support Vector Machines (SVM), have been widely employed for malicious and benign activity. These models depend on labeled datasets in which known attack types are pre-classified [4]. Random Forests and Decision Trees are proficient in pattern recognition and classification utilizing previous attack data such as CICIDS2017 and NSL-KDD [5].
2) **Unsupervised learning techniques:**Unsupervised learning techniques, like Clustering Algorithms and Autoencoders, are very effective in identifying anomalies without labeled data [4]. K-means clustering techniques can assist in identifying anomalous activity patterns that may indicate zero-day vulnerabilities. The Kitsune framework, an unsupervised model utilizing autoencoders, has demonstrated efficacy in identifying network irregularities associated with zero-day attacks [6].
3) **Deep Learning Algorithms:** Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, are very proficient in managing high-dimensional and complex data, such as network traffic records [7]. These models efficiently identify complex zero-day attacks encompassing several stages and methodologies. Convolutional Neural Networks (CNNs) are frequently utilized to discover abnormalities in network traffic. In contrast, Long Short-Term Memory (LSTM) networks are employed for sequence-based data, recognizing temporal patterns that may indicate active attacks [8].
4) **Federated Learning:** It emerges as a viable artificial intelligence paradigm for identifying zero-day attacks in distributed and decentralized settings [9]. In contrast to traditional models that necessitate centralized data collection, federated learning enables individual devices to train their models using their data locally, transmitting only the model updates to a central server. This approach preserves privacy and reduces risks linked to data transmission, especially in sensitive contexts like IoT networks.
5) **Reinforcement learning (RL):** Reinforcement learning

(RL) techniques, including Deep Q-Networks (DQN), have been explored for zero-day attack detection. These models learn optimal techniques for identifying and alleviating attacks through interaction with an environment and receiving input as rewards or penalties [10]. Reinforcement learning methods are especially effective for real-time adaptation to dynamic threats, facilitating proactive threat hunting and automated solutions to zero-day vulnerabilities.

6) **Hybrid approaches:** Hybrid approaches integrating supervised and unsupervised learning have been investigated to capitalize on the advantages of both approaches. Integrating clustering methods with decision trees or SVMs augments the model's capacity to identify zero-day threats by merging anomaly detection with classification [11].

IV. Zero-Day Attacks Detection Techniques

Detecting zero-day attacks can be challenging because they exploit undisclosed vulnerabilities lacking patches or unique signatures. This section outlines various methodologies for detecting zero-day threats, encompassing traditional and advanced AI-driven solutions.

1) **Behavioral Analysis:** Behavioral analysis entails observing programs, processes, and network traffic to identify anomalies or suspicious actions that diverge from established baselines. These anomalies frequently indicate zero-day attacks seeking to exploit undiscovered vulnerabilities [6].
2) **Heuristic Analysis:** Heuristic analysis employs rule-based techniques to evaluate code and investigate its potential for malicious use. It can identify probable zero-day malware even without matching known malware signatures. This approach, while effective, may occasionally yield false positives due to its broad criteria for detecting suspicious behavior [12].
3) **Sandboxing:** Sandboxing entails executing potentially malicious files or code within an isolated, controlled environment to monitor their activity without endangering the actual system [13]. It is an essential method for identifying zero-day malware that seeks to exploit undisclosed vulnerabilities, as it offers a secure environment for testing without affecting the live system.
4) **Machine Learning and Artificial Intelligence:** Artificial intelligence and machine learning models have become essential in detecting zero-day attacks due to their ability to detect patterns within extensive datasets. These models are trained on extensive historical data to detect anomalous behaviors, indicators of compromise (IOCs), and subtle patterns that may indicate zero-day attacks [14]. Machine learning techniques, including deep learning, support vector machines (SVM), and decision trees, can detect complex attack patterns and zero-day vulnerabilities that traditional signature-based systems may overlook.
5) **Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS):** Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) are solutions, either network-based or host-based, that monitor network traffic and system activity for indications of malicious conduct [15]. Intrusion Detection and Prevention Systems (IDS/IPS) integrate signature-based and anomaly-based detection methodologies. They can identify anomalous behavior or deviations from standard processes, perhaps signifying a zero-day attack [16].
6) **Reputation-Based Detection:** Reputation-based detection utilizes databases that monitor recognized benign and malicious files, domains, IP addresses, and URLs. This approach is especially effective for detecting zero-day threats originating from recognized malicious actors or suspicious sources [17].
7) **Endpoint Detection and Response (EDR):** EDR systems continuously monitor endpoint devices such as PCs, servers, and mobile devices for anomalous activity in real-time. EDR systems can identify and respond to zero-day attacks by examining processes, network connections, and system activity to highlight unusual patterns, processes, or unauthorized activities [18].
8) **Threat Intelligence Feeds:** Threat intelligence streams deliver real-time data on new threats, encompassing zero-day vulnerabilities and exploits. Its feeds would allow organizations to proactively apply mitigation techniques against newly identified zero-day attacks [19].

V. Emerging AI-Technology for Zero-Day Attacks

This section compares deep learning with traditional machine learning, examines federated learning-based solutions, and highlights lightweight AI models that balance effectiveness and efficiency in resource-constrained environments.

A. Deep Learning versus Traditional Machine Learning

Deep learning (DL) and classical machine learning (ML) methodologies are pivotal in zero-day attack detection, yet they markedly differ in architecture, data processing, and detection accuracy [1].

1) **Traditional Machine Learning:** Traditional machine learning models, including decision trees, support vector machines (SVM), and random forests, are generally constructed to identify established patterns or anomalies in pre-labeled datasets. These models depend on manually produced features and significantly rely on labeled data for training. Although proficient in identifying established attack patterns, traditional machine learning methods frequently falter in addressing the unpredictability of zero-day attacks due to their inability to generalize from insufficient or nonexistent past instances.
2) **Deep Learning**: Deep learning models, such as CNNs and LSTMs, proficiently detect zero-day threats by autonomously detecting complex patterns within extensive amounts of unlabeled data, eliminating manual feature engineering. Their capacity for ongoing learning and

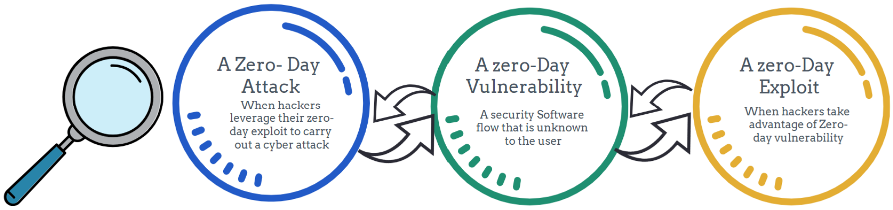
A Zero- Day Attack
When hackers leverage their zero-day exploit to carry out a cyber attack
A zero-Day Vulnerability
A security Software flow that is unknown to the user
A zero-Day Exploit
When hackers take advantage of Zero-day vulnerability

Fig. 3. A Zero-Day Definitions.

adaptation to new tactics for attack makes them exceptionally proficient in detecting zero-day vulnerabilities, particularly within high-dimensional data environments [6].

B. Federated Learning-based Solutions

Federated learning is an emerging AI paradigm with significant potential for zero-day attack detection where data privacy and security are critical [20].

1) **Decentralized Learning:** Federated learning enables training machine learning models across numerous decentralized devices without transferring sensitive data to a central server. This facilitates ongoing education in decentralized settings while protecting sensitive data, thus enhancing privacy [21].
2) **Application in Zero-Day Detection:** Federated learning is particularly beneficial for contexts characterized by large data distribution, such as IoT networks, mobile devices, and cloud systems, in the context of zero-day attacks [22]. It can discover zero-day attacks by aggregating information from devices that have experienced similar anomalies, resulting in a more adaptive and responsive detection framework [23].

C. Lightweight AI-based Models

Lightweight AI models are engineered explicitly for situations with constrained computational resources. These models seek to effectively detect zero-day attacks while reducing memory usage, computing power, and energy consumption. Due to the growing prevalence of resource-limited contexts and the escalating risk of zero-day attacks, lightweight models are becoming crucial for ensuring robustness. Federated learning is especially advantageous in environments marked by extensive data dispersion, like IoT networks, mobile devices, and cloud systems, notably for zero-day threats. These devices can collaboratively learn from each other's experiences with new and emerging threats, thereby enhancing the network's overall detection capabilities. Federated learning systems can identify zero-day attacks by consolidating data from multiple devices with similar anomalies [24]. As a result, this leads to a more adaptive and responsive detection framework to enhance security without imposing excessive demands on the systems they protect.

VI. Existing Zero-Day Datasets

TABLE I shows the summary of machine learning-based zero-day attack detection methods by approach. The table compares machine learning methodologies for identifying zero-day cyber-attacks, classified into unsupervised, supervised, reinforcement learning, and hybrid techniques. Unsupervised methods, including outlier detection via SVMs and autoencoders, generally train on benign traffic from datasets such as CIC-IDS2017 and NSL-KDD [25]. These models have different recall rates (27%–99%); however, they frequently encounter challenges in maintaining consistency across diverse attack types. Kitsune [26], an ensemble-based IP camera video monitoring approach, demonstrates effective performance; nevertheless, it has a restricted number of test cases. Supervised techniques, including models such as random forests, GANs, and convolutional neural networks (CNN), are assessed on datasets including CSE-CIC-IDS2018 and AWID [27], demonstrating high detection accuracy (up to 98%). Nonetheless, they have constraints due to restricted attack scenarios and the necessity for extensive labeled datasets, rendering them less adaptable. Reinforcement learning approaches, illustrated as DQN-based detectors [28], achieve a high detection rate of 90%. However, they are computationally intensive and need considerable time for training. Hybrid methodologies integrating supervised and unsupervised techniques show significant potential for perfect detection in controlled scenarios.

VII. Advantages, Limitations, and Challenges of Existing Solutions

Detecting zero-day attacks is a significant challenge in cybersecurity because the vulnerabilities they exploit are unpredictable. Different approaches, especially those using artificial intelligence (AI) and machine learning (ML), have their own benefits, limitations, and inherent challenges. First, discussing advantages and benefits, AI and ML methodologies, encompassing supervised, unsupervised, and hybrid models, have demonstrated notable success in detecting zero-day threats by recognizing uncharted patterns within large datasets. These methods have several key advantages. The first one is Adaptability, where the Machine learning models, particularly deep learning architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, show significant adaptability. They can analyze substantial data sets and identify complex patterns essential for managing zero-day attacks' complexity and evolving nature [1]. The second

TABLE I
SUMMARY OF MACHINE LEARNING (ML)-BASED ZERO-DAY ATTACK DETECTION METHODS BY APPROACH

Detector Name	ML Model	Training Data Set	Zero-day Testing Data Set	Evaluation Results	Challenges & Issues
Unsupervised Approaches					
Outlier detector (2020) [25]	One-class SVM	CIC-IDS2017, NSL-KDD (benign traffic only)	Attacks withheld from training and used in testing	Recall varies from 27% to 99% based on attack types	Varying accuracy, inconsistent against different attacks
Outlier detector (2020) [25]	Autoencoder	CIC-IDS2017, NSL-KDD (benign traffic only)	Attacks withheld from training and used in testing	Out-performs One-class SVM on complex attacks	Varying accuracy, inconsistent against different attacks
Kitsune (2018) [26]	An ensemble of autoencoders	IP camera video surveillance test-bed (benign traffic only)	Emulated attacks on test-bed	Out-performs offline platforms; very good accuracy against backdoor attack types	Varying accuracy, limited test cases
Supervised Approaches					
Comparison of six Supervised ML detectors (2019) [27]	Random forest, Gaussian naive Bayes, Decision tree, MLP, K-nearest neighbors, Quadratic discriminant analysis	CSE-CIC-IDS2018	Eight new attacks collected in real networks	Vary greatly based on attack types; decision tree model performs best	Varying accuracy, inconsistent against different attacks
Generative Adversarial Networks (GAN)-based detector (2018) [29]	GAN, Autoencoder	Kaggle Microsoft Malware Classification Challenge	Generated by introducing noise to existing malware	98% detection accuracy; robust against different noise levels	Limited test cases
Feature-based Transfer Learning (TL) detector (2019) [30]	Spectral transformation-based approach, decision tree, SVM, KNN	NSL-KDD; three source-target domain pairs used in training	Attack labels withheld in training and used in testing	70% accuracy, 0.75 F1 score	Low detection accuracy, limited test cases
CNN-based anomaly detection (2018) [31]	Convolutional Neural Network (CNN)	AWID dataset, CICIDS2017	IoT-related zero-day attacks from internal data	92% accuracy; low false positive rate (3.5%)	Requires large amounts of labeled data for training, potential overfitting
Reinforcement Learning Approaches					
Reinforcement Learning-based zero-day detection (2020) [28]	Deep Q-network (DQN) based reinforcement learning	KDDCup99, CICIDS2018	Novel attack vectors generated via adversarial networks	90% detection rate for zero-day variants	High training time, difficult in high-dimensional spaces
Hybrid Approaches					
Two-level supervised and unsupervised learning method (2013) [32]	Binary random forest model and SVM	Private data set from an ISP	Attacks withheld from training and used in testing	88.54% detection at top level; AUC reaches 99%	Limited test cases, evaluated using private data
Hybrid learning method (2017) [33]	K-means cluster, Random Forest, SVM	Data from CA Technologies VET Zoo	Attack labels withheld in training and used in testing	Perfect detection – AUC 1, accuracy 100%	Limited test cases, data set not public anymore
Domain adaptation-based TL detector (2020) [34]	Manifold alignment-based domain adaptation	NSL-KDD and CIDD data sets	Zero-day attacks withheld from training	Vary according to domain pairs; out-performs feature-based TL detector	Limited test cases, varying accuracy
LSTM-based zero-day attack detector (2024) [35]	LSTM	UNSW-NB15, CICIDS2017	Attacks from APT datasets	Detects zero-day attacks with F1-score of 0.82	High computational cost and slower processing
Autoencoder-based anomaly detection (2020) [28]	Autoencoder, LSTM	CICIDS2017, Bot-IoT dataset	Zero-day attack variants withheld from training	¿ 93% accuracy; suitable for real-time monitoring	Requires careful tuning, prone to false positives
Deep Belief Network (DBN)-based detector (2020) [25]	DBN	NSL-KDD, CICIDS2017	Zero-day attacks from internal benchmarks	90% detection accuracy, robust in complex networks	High computational cost, complex model training
Hybrid CNN-RNN zero-day detector (2019) [36]	CNN-RNN hybrid model	CICIDS2018, UNSW-NB15	Attacks simulated via adversarial learning	94% detection rate; effective for high-dimensional data	Complex architecture, slow training time

one is anomaly Detection. Behavioral analysis and anomaly detection methods enable the identification of anomalies in network traffic or system behavior. These models can discover novel attack patterns by recognizing anomalies from the standard without requiring predefined signatures [1]. Last but not least, the real-time Detection AI-driven systems can integrate with real-time monitoring solutions, including Intrusion Detection Systems (IDS) and Endpoint Detection and Response (EDR) systems, to identify zero-day attacks as they occur. This substantially decreases the interval between detection and response, mitigating damage from active attacks [37]. Machine learning techniques and intense learning models demonstrate remarkable scalability across extensive distributed networks. This is particularly advantageous in Internet of Things (IoT) networks, where attacks can propagate rapidly across interconnected devices.

AI/ML-based zero-day attack detection systems have inherent limitations despite their advantages. The Data Dependency, such as supervised models, rely heavily on labeled data for training, which is frequently limited or absent in zero-day attacks. Although unsupervised techniques such as autoencoders can be beneficial, their effectiveness is limited by the quality and diversity of the training data [28]. Using machine learning models, especially deep learning architectures, often leads to overfitting and heightened false positive rates. Overfitting may result in elevated detection rates during training, although it compromises generalization in practical applications. Moreover, these algorithms frequently generate elevated false-positive rates, overloading security personnel with benign alerts [38]. In addition, resource-intensive deep learning models such as LSTM and CNN also require substantial computer resources for training and real-time operation. Resource-constrained contexts, such as IoT networks or edge devices, restrict their feasibility for continuous real-time monitoring [25]. Furthermore, many AI models, particularly intense learning systems, lack transparency in their decision-making processes, making them appear as "black boxes".The lack of transparency can complicate the interpretation and trustworthiness of outcomes for cybersecurity experts, particularly in critical security scenarios where rapid decision-making is crucial [39].

Zero-day attack detection faces several challenges related to AI/ML [6], including the lack of training, which can be mitigated through Few-shot learning and transfer to detect new threats with minimal labeled data. Another issue is high false positives, which can be reduced by hybrid models that combine rule-based, heuristic, and AI/ML techniques to improve detection accuracy. Adversarial Attacks can be addressed by creating robust adversarial models using defensive distillation and adversarial training to prevent manipulation. In the face of Evasion and Polymorphism, reinforcement learning creates a dynamic defense that constantly changes to counter new infections and attack techniques. Explainability and Interpretability Develop are also crucial for interpretable insights and increase trust in AI-based security solutions by utilizing explainable AI (XAI) methodologies [40], [41].

Moreover, Data Quality and Noise can be addressed using Automated feature engineering (AutoML) to improve model accuracy, reduce noise, and select pertinent features [25]. Real-time processing constraints are a challenge as well, but Model pruning, quantization, and knowledge distillation can support real-time detection in constrained environments. Continuous learning and adaptation are needed to tackle concept drift so that AI/ML systems can handle idea drift and learn from real-world data. In terms of adversarial robustness, strategies like defensive distillation, adversarial training, and model hardening can strengthen defenses, while Graph-based learning limitations can be overcome by utilizing knowledge graphs and graph-based learning techniques to capture complex attack patterns better.

Federated learning is another area of focus, as it improves detection capacities without exchanging raw data among organizations through safe, collaborative training. Enriching the feature space with user behavior, network topology, and threat intelligence feeds to address the lack of threat intelligence integration can lead to more context-aware detection. Finally, collaboration with human experts through human-in-the-loop techniques allows for refinement and provides feedback on models to enhance decision-making.

VIII. Lessons Learned, Suggestions & Recommendations

This section highlights key findings derived from examining AI-based zero-day attack detection methodologies, providing ideas and recommendations for subsequent research endeavors [24]. A significant finding is the adaptability of AI models, particularly those utilizing machine learning (ML) and deep learning (DL); nonetheless, their efficacy depends upon the availability of current and high-quality training data, hence requiring ongoing data collecting and retraining [24]. Real-time detection, enabled by behavioral and heuristic analysis, along with Intrusion Detection Systems (IDS) and Endpoint Detection and Response (EDR), is essential for reducing detection time [42] Nonetheless, overfitting and false positives continue to pose considerable issues, frequently overwhelming security personnel with superfluous alarms, requiring careful hyperparameter tuning and training data selection to achieve a balance in accuracy [43]. Ultimately, the absence of access to extensive and modern public datasets, such as CICIDS2017 and NSL-KDD, hinders the generalizability of models, highlighting the necessity for more up-to-date datasets to represent evolving attack strategies [25].

IX. Conclusion and Future Work

This study offers an in-depth examination of the challenges faced with detecting zero-day attacks, emphasizing the use of advanced artificial intelligence (AI) and machine learning (ML) models in improving detection and protection strategies. The research assesses diverse AI-driven approaches, encompassing supervised, unsupervised, hybrid techniques, deep learning, federated learning, and lightweight models designed for resource-constrained environments such as IoT [24]. Despite significant progress in scalability and real-time detection, challenges remain, including inadequate datasets, lack of explainability, and privacy issues. Further research into zero-day attack detection must include enhancing the robustness of AI models against adversarial attacks via techniques such as adversarial training, defensive distillation, and model hardening [6]. Few-shot and transfer learning must be investigated to mitigate the deficiency of labeled data, facilitating improved model generalization across domains [44]. Explainable AI (XAI) is crucial for maintaining transparency and trust in machine learning systems, particularly in critical security environments [45]. Federated learning can mitigate privacy problems by enabling decentralized devices to contribute to model training without disclosing sensitive data; nevertheless, additional research is required to enhance communication and consistency [46]. Moreover, graph-based learning can identify complex attack routes, and lightweight models are essential for real-time detection in resource-constrained environments [47]. Continuous learning techniques that accommodate idea drift will enable models to adjust to changing attack patterns, and more extensive public datasets are essential to enhance benchmarking and stimulate innovation. These initiatives will improve AI-driven zero-day attack detection systems' resilience, adaptability, and transparency.

References

[1] R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, "Zero-day attack detection: a systematic literature review," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 10 733–10 811, 2023.

[2] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[4] D. K. Gasu, "Threat detection in cyber security using data mining and machine learning techniques," in *Modern theories and practices for cyber ethics and security compliance*. IGI Global, 2020, pp. 234–253.

[5] A. Alshamrani, "Cyber attacks detection and mitigation in sdn environments," Ph.D. dissertation, Arizona State University, 2018.

[6] Y. Guo, "A review of machine learning-based zero-day attack detection: Challenges and future directions," *Computer communications*, vol. 198, pp. 175–185, 2023.

[7] M. B. Sarwar, M. K. Hanif, R. Talib, M. Younas, and M. U. Sarwar, "Darkdetect: Darknet traffic detection and categorization using modified convolution-long short-term memory," *IEEE Access*, vol. 9, pp. 113 705–113 713, 2021.

[8] M. Rithani, R. P. Kumar, and S. Doss, "A review on big data based on deep neural network approaches," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14 765–14 801, 2023.

[9] P. Verma, N. Bharot, J. G. Breslin, D. O'Shea, A. Vidyarthi, and D. Gupta, "Zero-day guardian: A dual model enabled federated learning framework for handling zero-day attacks in 5g enabled iiot," *IEEE Transactions on Consumer Electronics*, 2023.

[10] S. Shen, C. Cai, Z. Li, Y. Shen, G. Wu, and S. Yu, "Deep q-network-based heuristic intrusion detection against edge-based siot zero-day attacks," *Applied Soft Computing*, vol. 150, p. 111080, 2024.

[11] R. Kaur and M. Singh, "A hybrid real-time zero-day attack detection and analysis system," *International Journal of Computer Network and Information Security*, vol. 7, no. 9, pp. 19–31, 2015.

[12] M. Zakeri, F. Faraji Daneshgar, and M. Abbaspour, "A static heuristic approach to detecting malware targets," *Security and Communication Networks*, vol. 8, no. 17, pp. 3015–3027, 2015.

[13] E. Debas, N. Alhumam, and K. Riad, "Unveiling the dynamic landscape of malware sandboxing: A comprehensive review," 2023.

[14] P. Parrend, J. Navarro, F. Guigou, A. Deruyver, and P. Collet, "Foundations and applications of artificial intelligence for zero-day and multi-step attack detection," *EURASIP Journal on Information Security*, vol. 2018, pp. 1–21, 2018.

[15] K. Scarfone and P. Mell, "Intrusion detection and prevention systems," in *Handbook of information and communication security*. Springer, 2010, pp. 177–192.

[16] J. M. Kizza, "System intrusion detection and prevention," in *Guide to computer network security*. Springer, 2024, pp. 295–323.

[17] K. Gerrigagoitia, R. Uribeetxeberria, U. Zurutuza, and I. Arenaza, "Reputation-based intrusion detection system for wireless sensor networks," in *2012 Complexity in Engineering (COMPENG). Proceedings*. IEEE, 2012, pp. 1–5.

[18] H. Kaur, D. S. SL, T. Paul, R. K. Thakur, K. V. K. Reddy, J. Mahato, and K. Naveen, "Evolution of endpoint detection and response (edr) in cyber security: A comprehensive review," in *E3S Web of Conferences*, vol. 556. EDP Sciences, 2024, p. 01006.

[19] A. Niakanlahiji, "Discovering zero-day attacks by leveraging cyber threat intelligence," *Order*, no. 22592314, 2019.

[20] M. A. Ayed and C. Talhi, "Federated learning for anomaly-based intrusion detection," in *2021 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2021, pp. 1–8.

[21] I. Hegedűs, G. Danner, and M. Jelasity, "Decentralized learning works: An empirical comparison of gossip learning and federated learning," *Journal of Parallel and Distributed Computing*, vol. 148, pp. 109–124, 2021.

[22] L. Y. Por, Z. Dai, S. J. Leem, Y. Chen, J. Yang, F. Binbeshr, K. Y. Phan, and C. S. Ku, "A systematic literature review on the methods and challenges in detecting zero-day attacks: Insights from the recent crowdstrike incident," *IEEE Access*, 2024.

[23] A. A. Korba, A. Boualouache, B. Brik, R. Rahal, Y. Ghamri-Doudane, and S. M. Senouci, "Federated learning for zero-day attack detection in 5g and beyond v2x networks," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 1137–1142.

[24] S. Ali, S. U. Rehman, A. Imran, G. Adeem, Z. Iqbal, and K.-I. Kim, "Comparative evaluation of ai-based techniques for zero-day attacks detection," *Electronics*, vol. 11, no. 23, p. 3934, 2022.

[25] H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Utilising deep learning techniques for effective zero-day attack detection," *Electronics*, vol. 9, no. 10, p. 1684, 2020.

[26] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: an ensemble of autoencoders for online network intrusion detection," *arXiv preprint arXiv:1802.09089*, 2018.

[27] Q. Zhou and D. Pezaros, "Evaluation of machine learning classifiers for zero-day intrusion detection–an analysis on cic-aws-2018 dataset," *arXiv preprint arXiv:1905.03685*, 2019.

[28] A. Thakkar and R. Lohiya, "A review of the advancement in intrusion detection datasets," *Procedia Computer Science*, vol. 167, pp. 636–645, 2020.

[29] J.-Y. Kim, S.-J. Bu, and S.-B. Cho, "Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders," *Information Sciences*, vol. 460, pp. 83–102, 2018.

[30] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9259–9266.

[31] R. Panigrahi and S. Borah, "A detailed analysis of cicids2017 dataset for designing intrusion detection systems," *International Journal of Engineering & Technology*, vol. 7, no. 3.24, pp. 479–482, 2018.

[32] P. M. Comar, L. Liu, S. Saha, P.-N. Tan, and A. Nucci, "Combining supervised and unsupervised learning for zero-day malware detection," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 2022–2030.

[33] S. Huda, S. Miah, M. M. Hassan, R. Islam, J. Yearwood, M. Alrubaian, and A. Almogren, "Defending unknown attacks on cyber-physical systems by semi-supervised approach and available unlabeled data," *Information Sciences*, vol. 379, pp. 211–228, 2017.

[34] N. Sameera and M. Shashi, "Deep transductive transfer learning framework for zero-day attack detection," *ICT Express*, vol. 6, no. 4, pp. 361–367, 2020.

[35] P. Li, H. Wang, G. Tian, and Z. Fan, "A cooperative intrusion detection system for the internet of things using convolutional neural networks and black hole optimization," *Sensors*, vol. 24, no. 15, p. 4766, 2024.

[36] O. Faker and E. Dogdu, "Intrusion detection using big data and deep learning techniques," in *Proceedings of the 2019 ACM Southeast conference*, 2019, pp. 86–93.

[37] A. Arfeen, S. Ahmed, M. A. Khan, and S. F. A. Jafri, "Endpoint detection & response: A malware identification solution," in *2021 International Conference on Cyber Warfare and Security (ICCWS)*. IEEE, 2021, pp. 1–8.

[38] P. Pitre, A. Gandhi, V. Konde, R. Adhao, and V. Pachghare, "An intrusion detection system for zero-day attacks to reduce false positive rates," in *2022 International Conference for Advancement in Technology (ICONAT)*. IEEE, 2022, pp. 1–6.

[39] T. D. Krafft, M. Reber, R. Krafft, A. Coutrier, and K. A. Zweig, "Crucial challenges in large-scale black box analyses," in *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 2021, pp. 143–155.

[40] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[41] V. Kumar and D. Sinha, "A robust intelligent zero-day cyber-attack detection technique," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2211–2234, 2021.

[42] S. Reardon, M. D. Hssayeni, and I. Mahgoub, "Detection of zero-day attacks on iot," in *2024 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE, 2024, pp. 1–5.

[43] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 453–563, 2022.

[44] M. Gamal, H. M. Abbas, N. Moustafa, E. Sitnikova, and R. A. Sadek, "Few-shot learning for discovering anomalous behaviors in edge networks," *Computers, Materials and Continua*, vol. 69, no. 2, pp. 1823–1837, 2021.

[45] T. Rahman and M. Sayduzzaman, "Zero-day attacks detection in smart community through interoperability and explainable ai," 2024.

[46] T. Ohtani, R. Yamamoto, and S. Ohzahata, "Detecting zero-day attack with federated learning using autonomously extracted anomalies in iot," in *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*. IEEE, 2024, pp. 356–359.

[47] M. Swathy Akshaya and G. Padmavathi, "Zero-day attack path identification using probabilistic and graph approach based back propagation neural network in cloud," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 3s2, pp. 1091–1106, 2022.