Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluatio
DOI: 10.54254/2755-2721/2025.PO25895

Adaptive IDS via Lightweight Model Repair for Synthetic Zero-Day Attacks

Zijie Liu

Department of Computer Science, University of California, Davis, USA
zijliu@ucdavis.edu

Abstract. As cyber threats grow increasingly sophisticated, traditional Intrusion Detection System (IDS) struggle to maintain robustness when facing unknown zero-day attacks. This paper proposes an adaptive IDS framework based on lightweight model repair, which rapidly restores detection performance using a small set of synthetically generated zero-day samples. The architecture combines an autoencoder with a classifier head, jointly trained on the NSL-KDD dataset. This framework was evaluated against four canonical zero-day attack variants via controlled feature manipulation. To address the performance degradation, multiple minimal-scope repair strategies are tested. Repair strategies target three scopes: the decoder's final layer, the classifier head, or both components. The choice of repair scope is informed by empirical performance under different adaptation scenarios. According to the experiment results, fine-tuning only the classifier head can restore the recall to over 98% with strong generalization ability. Unified pooled repair and leave-one-out evaluations further verify the robustness and adaptability of the method across diverse attack scenarios.

Keywords: Lightweight Model Repair, Synthetic Zero-Day Attacks, Intrusion Detection Systems, Autoencoder, Classifier Head.

1. Introduction

In recent years, more and more security threats have evolved from traditional attacks to sophisticated cyberattacks, threatening organizations and critical infrastructures across all sectors. Recent statistics show that cyberattacks now occur every 39 seconds, and over 3,200 major incidents were recorded globally in 2024, underscoring the urgency of advanced detection systems [1,2]. Due to increasing network dependency, identifying and detecting these attacks in an efficient and reliable manner becomes the key topic in the computer security field. Intrusion Detection System (IDS), as one of the core technologies of network protection, can provide real time analysis of network flows and identify abnormal activities, which can stop the spread of attacks at an early stage. Nonetheless, conventional signature-based techniques frequently struggle to identify novel assaults, resulting in considerable declines in detection accuracy [3]. To sustain optimal IDS performance, certain studies depend on the continual update of the dataset and the retraining of the model—an method that is both time-intensive and computationally inefficient [4]. To mitigate these drawbacks, this study investigates an adaptive IDS framework that leverages lightweight model repair to restore detection performance against unseen threats. A stacked autoencoder and a softmax classifier head are first

DOI: 10.54254/2755-2721/2025.PO25895

trained on the NSL-KDD dataset. Four synthetic zero-day attack variants are then introduced to evaluate baseline robustness, where the model exhibits poor recall (e.g., 12–32%). To address this issue, targeted lightweight repair is applied to the classifier head, which consistently restores accuracy to 100%, demonstrating the potential of minimal-scope adaptation for reliable IDS recovery. As cyber attacks continue to evolve into higher complexity and scale, robustness and adaptivity become indispensable features of IDS [5]. This study proposes a lightweight repair framework that enables swift recovery from synthetic zero-day attacks without full-model retraining. These results establish a practical, resource-efficient direction for building adaptive IDS in real-world deployment.

2. Related work

Signature-based intrusion detection systems (SIDS) rely on a database of known attacks and are effective in detecting previously seen threats with low positive rates. However, in the real world, an ideal IDS should not bypass any threats, and the SIDS is highly unable to do this when new threats are introduced [6]. Anomaly-based IDS (AIDS) identifies deviations from normal traffic patterns to detect malicious behavior, yet incurs high false positive rates and requires intensive calibration efforts [7]. The hybrid method of SIDS and AIDS aims to combine the strengths of each system, improving detection convergence. Yet, they still face signature update delays and calibration overhead, which limits their practicality in complex and dynamic threat environments—studies have shown that signature updates may lag by over 24 hours, and anomaly-based systems can suffer false positive rates exceeding 50% [8,9]. Full model retraining remains a conventional strategy to address IDS vulnerabilities against novel threats., but this approach faces significant limitations. First, retraining the entire model can be extremely time-consuming, particularly with deep learning architectures and large-scale network data [10,11]. These constraints hinder adaptive deployment of IDS updates, especially in dynamic environments where rapid adaptation is crucial.

Recent work has mimicked synthetic or “zero-shot” attack scenarios to evaluate the model’s robustness against unseen threats. For example, Muresan et al. proposed a zero-shot learning framework that simulates unseen attack types and measures a Zero-day Detection Rate (Z-DR) to assess IDS generalization capabilities [12]. Although such methodologies effectively benchmark zero-day detection capabilities, they prioritize evaluation over robustness enhancement. Other approaches leverage model-agnostic meta-learning (MAML) to adapt IDS models from few-shot samples in IoT settings—achieving over 98% recall on NSL-KDD—but require episodic task structures and still struggle with truly novel attack types [13]. Another line of work uses density-aware active sampling combined with generative augmentation to adaptively retrain IDS on concept-drifting data, boosting F1-scores on rare attacks from near-zero to 0.30–0.71 on CIC-IDS-2018; however, it demands a continuous labeling and augmentation pipeline that may not be feasible in real-time operational environments [14]. This work advances beyond evaluation by leveraging synthetic attacks to drive targeted classifier-head repairs, establishing direct pathways from robustness assessment to operational recovery.

3. Methodology

3.1. Model architecture

The model architecture adopted in this paper is from AOC-IDS, which consists of a stacked autoencoder (AE) for feature embedding and a linear classifier head for prediction [15]. The encoder

DOI: 10.54254/2755-2721/2025.PO25895

transforms the preprocessed feature vector $x \in R^d$ (dimensionality = d) into an 8-dimensional latent space through four fully connected layers:

$$x \to h_1 \in R^{80} \to h_2 \in R^{40} \to h_3 \in R^{20} \to z \in R^8$$

where each layer is followed by ReLU activation. The decoder mirrors this structure in reverse:

$$z \to h_3^{'} \in R^{20} \to h_2^{'} \in R^{40} \to h_1^{'} \in R^{80} \to \widehat{x} \in R^d$$

ending with a Sigmoid activation for input reconstruction. Following principles established by Zhang et al. using an autoencoder to compress input to a lower-dimensional latent space (e.g., 8D) enables the model to focus on essential features and reduces overfitting while simplifying downstream repair. The classifier head consists of a single linear layer that maps the 8-dimensional latent vector into two output logits for binary classification ("normal" vs. "anomaly"). During training, the autoencoder is first trained to minimize reconstruction loss (MSE), after which the classifier is trained separately on the fixed latent representations to minimize cross-entropy loss.

This architecture is lightweight and modular, allowing for easy separation between representation learning and decision-making. Its simplicity and proven performance on NSL-KDD make it a practical choice for evaluating lightweight repair strategies in a controlled setting.

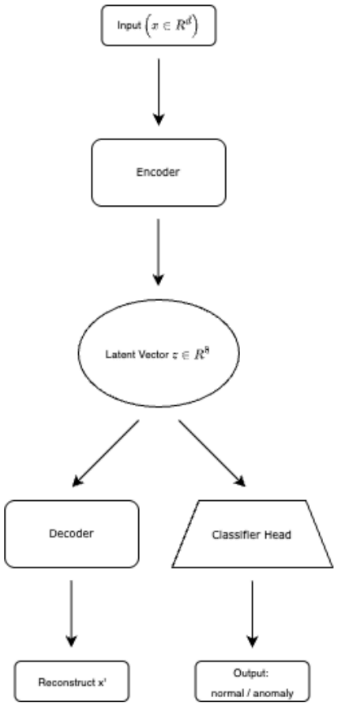

Figure 1. Model architecture (autoencoder + classifier head)

The architecture is lightweight and modular, allowing for separation between representation learning (reconstruction) and decision-making (classification). Its simplicity and performance on NSL-KDD make it a practical testbed for evaluating lightweight repair.

DOI: 10.54254/2755-2721/2025.PO25895

3.2. Synthetic zero-day attack generation and lightweight model repair strategies

To simulate previously unseen attack patterns, I generate synthetic zero-day variants by perturbing key features in normal samples from the NSL-KDD test set. Aligned with AOC-IDS framework principles, feature perturbations target dst_bytes, src_bytes, count, srv_count, and serror_rate—key indicators of known attack patterns. Specifically, I apply a 100% increase (delta_pct = 1.0) to the selected features in benign samples to create anomalous variants that mimic exfiltration, DoS, and scan-like behaviors. These synthetic samples are labeled as "attack" and used to evaluate the model's robustness to novel threats. This setup allows controlled and repeatable testing of repair strategies under realistic zero-day conditions, without relying on external datasets.

To enable rapid recovery against novel attack variants without full-model retraining, I investigate lightweight adaptation strategies that fine-tune only a minimal portion of the model. As illustrated in Figure 1, the encoder output is shared by two branches: a decoder (left) for reconstruction and a classifier head (right) for prediction. Specifically, the two lightweight repair strategies target either the decoder (left branch) or the classifier head (right branch). The classifier-head repair exhibits superior efficiency, requiring updates to merely one fully connected layer. This simplicity allows the model to be updated rapidly without the computational burden of full-model retraining, making it well-suited for real-time or resource-constrained deployment scenarios. Both approaches use only a small set of synthetically generated anomalous samples and are evaluated in terms of recovery performance, generalization, and computational efficiency.

This approach focuses on the decoder branch (left side) of the model. The encoder and all decoder layers are frozen except for the final fully connected layer of the decoder. This last layer is fine-tuned to reduce reconstruction error (mean squared error) on the synthetic attack samples. The goal is to realign the reconstruction pathway to better accommodate perturbations in the input space introduced by novel attacks, while preserving the integrity of the learned latent space. Training is performed using the Adam optimizer with a learning rate of 0.01 and early stopping to prevent overfitting. Decoder-specific updates preserve classifier integrity while improving latent space representation of novel attack patterns. This method focuses on retraining the classifier head (right branch) while keeping both the encoder and the latent feature extractor fixed. The classifier head, implemented as a single-layer softmax head, is retrained using cross-entropy loss on the latent codes of the synthetic attack samples. Since the architecture is minimal, this adaptation is extremely fast and requires very limited computational resources. This method is particularly useful when the latent representations are already well-organized and distinctive. In both pooled and leave-one-out assessments, classifier-head repair consistently elevated recall to over 98% across various assault variations, indicating robust generalization and resilience.

4. Experimental evaluation

4.1. Dataset and preprocessing

This study uses the NSL-KDD dataset, an improved and refined version of the widely cited KDD CUP 99 dataset [16]. The NSL-KDD is the improved version of the original dataset: it removes redundant rows and handles skewed distributions. Compared to the original KDD'99 dataset which contains 4,898,431 records with high redundancy, the NSL-KDD dataset significantly reduces duplication and class imbalance, retaining only 125,973 records with a more uniform distribution across attack types, providing a more balanced and representative benchmark for evaluating intrusion detection systems. The AOC-IDS preprocessed version applies one-hot encoding to

DOI: 10.54254/2755-2721/2025.PO25895

protocol_type, service, and flag features—key categorical attributes for attack pattern differentiation. All numerical features are retained without binning, and samples are labeled as either "normal" or one of several attack types. For the binary classification setup, the labels are consolidated into two categories: "normal" and "anomaly." The dataset is split into training and testing sets as defined by the NSL-KDD standard, preserving temporal and distributional structure for realistic evaluation.

4.2. Baseline detection evaluation

The model is composed of a stacked autoencoder (used for unsupervised feature learning) and a softmax classifier head trained on the NSL-KDD dataset. After training, the model is tested on a balanced evaluation set composed of normal samples and synthetically generated anomalous samples, each simulating distinct novel attack types. The results show that the baseline model performs well on clean test data but struggles significantly when exposed to the new attack variants. For example, recall values on synthetic attacks range between 21% and 83%, depending on the scenario. This sharp drop indicates that the model fails to detect most of the attacks, showing a lack of generalization toward out-of-distribution attack patterns and confirming that even high-performing models trained on known threats may fail catastrophically on novel ones. As shown in Figure 2, these observations highlight the need for a more adaptive mechanism to recover performance without retraining the entire model. Subsequent sections evaluate the impact of lightweight repair strategies on restoring detection effectiveness under these challenging conditions.

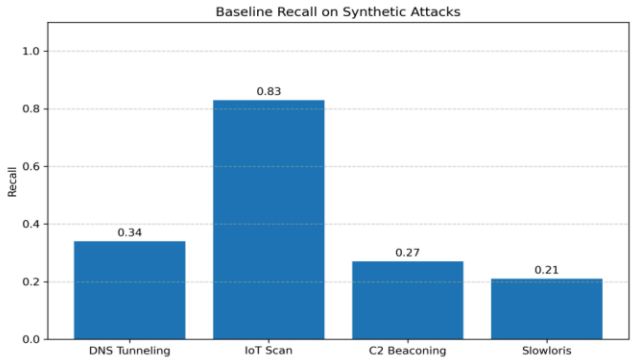

Figure 2. Baseline model recall on four synthetic attack variants

4.3. Per-attack repair results

To assess the effectiveness of the lightweight repair strategies, the repair performance of each of the four synthetic zero-day attack types is evaluated: DNS Tunneling Exfiltration, IoT Mass-Scan, Encrypted C2 Beaconing, and Slowloris HTTP Flood. Each attack was injected separately into the test pipeline to measure baseline detection performance, followed by decoder-layer and classifier-head repairs.

DNS Tunneling Exfiltration had a critically low baseline recall of 34%, indicating a complete failure to recognize anomalous patterns. Decoder repair yielded no improvement. Conversely, classifier-head repair rapidly increased recall to 100% within 20 epochs, with a clean learning curve and no overfitting, demonstrating the strong efficacy of minimal-scope adaptation." For IoT Mass-Scan, the model performed slightly better at baseline (83% recall) but still exhibited deficiencies in generalization. Decoder repair again failed to bring any improvement. However, classifier-head

DOI: 10.54254/2755-2721/2025.PO25895

repair achieved 100% accuracy after just one epoch and remained stable throughout, suggesting strong convergence and attack-specific recovery. Encrypted C2 Beaconing posed a significant challenge to the baseline model, with a recall of only 27%. Decoder repair proved ineffective once more. Classifier-head repair steadily increased recall, achieving perfect detection by epoch 19, reinforcing the robustness of the proposed repair strategy under highly evasive attack patterns.

The Slowloris HTTP Flood attack initially resulted in the weakest baseline performance (21% recall), highlighting the model's complete unawareness of this novel threat. Decoder repair showed no measurable benefit. Classifier-head repair, however, gradually restored recall to 99%, suggesting strong recovery even from near-zero detection scenarios.

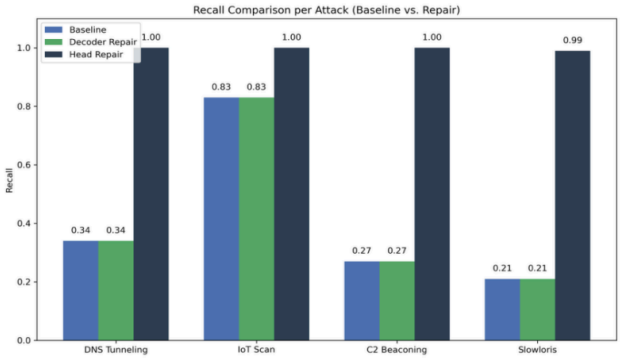

Figure 3. Per-attack recall comparison for or decoder/head repair

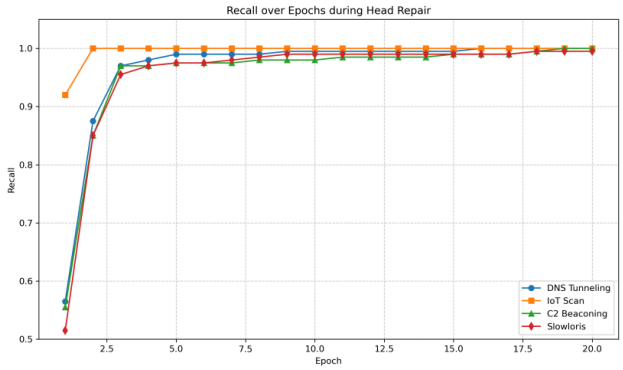

Figure 4. Recall curves during baseline model classifier-head repair

Decoder-only repair shows no improvement in recall or accuracy across all cases (Figure 3), indicating that modifying the way features are represented in a lower-dimensional latent code cannot help the model identify new attacks. Conversely, classifier-head repair not only restored full detection capability but also did so consistently and efficiently across distinct attack types. These results validate the proposed method's per-attack generalization and adaptability, showing that lightweight repair is not merely a patch but a scalable and reusable adaptation mechanism across diverse threat scenarios. This is further supported by Figure 4, which shows rapid and stable recall convergence within a few epochs across all attacks.

4.4. Unified pooled repair

Beyond per-attack repair, this study then applies the evaluated unified pooled repair strategy, where all four synthetic attack types were combined into a single repair set. Simulating a more realistic deployment scenario, this approach requires the IDS to adapt to multiple unknown threats simultaneously, without prior knowledge of which specific attack will occur.

Retraining the classifier head using the unified set, I tested performance individually on each attack variant post-repair. Despite the increased heterogeneity of the repair data, the unified model achieved near-optimal generalization across all cases (shown in figure 5):

- DNS Tunneling Exfiltration recall improved from 34% to 98%;
- IoT Mass-Scan recall rose from 83% to 99%;
- Encrypted C2 Beaconing recall increased from 27% to 97%;
- Slowloris HTTP Flood recall advanced from 21% to 99%.

DOI: 10.54254/2755-2721/2025.PO25895

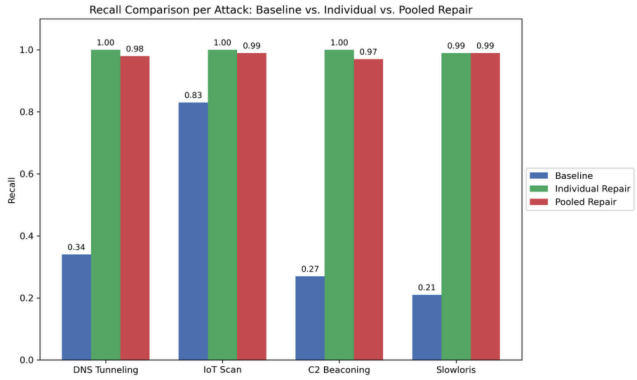

Figure 5. Recall comparison per attack: baseline vs. individual vs. pooled repair

All attacks achieved recall rates of ≥97%, with corresponding precision and F1-scores approaching perfect performance. This demonstrates that even when repaired with a small, aggregated sample set containing heterogeneous threat patterns, the model can generalize effectively without retraining its full architecture. Notably, decoder-layer repair remained ineffective when applied to the unified set (as previously observed in per-attack experiments), further validating that targeted repair of the classifier head is the most impactful and efficient strategy. Results confirm that pooled repair can serve as a single-shot adaptive mechanism capable of covering multiple novel threat types concurrently. This substantially reduces operational overhead compared to case-by-case repair. Additionally, a single unified update can effectively cover multiple threats, simplifying the maintenance of IDS performance in dynamic environments.

4.5. Leave-one-out generalization

To evaluate the robustness and generalization capability of the repair method, the author conducted a leave-one-out repair strategy. Synthetic attack data from three attack types were used to perform classifier-head repair in each trial, while the remaining attack was held out for evaluation. Despite the unseen nature of the held-out threat, the repaired model consistently achieved 100%. These results demonstrate that the learned representations from limited synthetic samples can generalize effectively to novel, unseen variants. This is particularly significant for real-world IDS deployment, where new attack types may emerge before corresponding samples can be incorporated into the training pipeline. Generalizing from partial threat knowledge enables proactive defense and reduces reliance on exhaustive retraining. Furthermore, achieving perfect detection on held-out variants suggests that the repair process does not overfit to specific attacks but instead adapts the classifier to broader anomaly patterns. This highlights the strength of lightweight repair not only in restoring performance but also in providing resilient and forward-compatible adaptation for evolving threat landscapes.

5. Discussion

In real-world deployment scenarios, intrusion detection systems must not only maintain high accuracy but also operate within constraints of time, computational resources, and incomplete threat visibility. The lightweight repair framework proposed in this study aligns well with such requirements. A key enabler of this practicality is the modular and simple architecture of the underlying model. The stacked autoencoder serves purely for feature encoding and is detached from

DOI: 10.54254/2755-2721/2025.PO25895

the decision-making process, which is handled by a single-layer classifier head. This clear separation allows targeted updates to be made to just the classifier without disturbing the representation learning pipeline. The simplicity of the classifier head also means that retraining is extremely fast, requiring only a few epochs and minimal computational resources. Moreover, the ability to generate synthetic zero-day samples based on feature-level heuristics eliminates the reliance on labeled real-world attack data. Analysts or automated systems can craft repair inputs from observed anomalies or threat intelligence patterns, enabling timely adaptation to new threats even in the absence of ground-truth labels. Finally, the repair mechanism is lightweight enough to be executed periodically or in an event-driven manner—triggered by sudden drops in detection performance or observable input shifts. It can be integrated into an existing IDS pipeline as an auxiliary module, allowing low-disruption, high-impact updates. This design paves the way for real-time or near-real-time IDS adaptation workflows that maintain robustness in dynamic threat environments.

Despite promising results, several limitations of the current approach warrant further exploration. First, the synthetic attack variants used for repair are constructed via feature perturbations on normal samples, based on prior knowledge of common attack behaviors. While this allows controlled evaluation, it may not fully capture the complexity or stealth of real-world zero-day exploits. Future work should explore integrating threat intelligence feeds or generative adversarial models to produce more realistic repair stimuli. Future work could explore integrating threat intelligence feeds or generative models such as Conditional GANs (cGANs) or Variational Autoencoders (VAEs) to synthesize more diverse and realistic attack variants. These models can learn distributional patterns from real-world data and introduce perturbations aligned with stealthy or obfuscated behaviors, improving repair stimulus fidelity. Secondly, the existing configuration presupposes the availability of ground-truth labels for synthetic attacks, which may not always be attainable. To resolve this, continual learning methodologies (e.g., Elastic Weight Consolidation, rehearsal-based memory buffers) may be integrated to incrementally change the classifier head.

6. Conclusion

This study proposes a practical and efficient solution to a critical problem in the modern intrusion detection system: how to rapidly restore detection performance against unseen, zero-day attacks without full model retraining. By using lightweight model repair strategies—particularly fine-tuning a single-layer classifier head using synthetically generated attack variants—the proposed framework achieves near-perfect recovery of recall and strong generalization across diverse attack types. The results demonstrate that minimal-scope updates are not only computationally efficient but also effective in adapting to evolving threats, offering a scalable defense mechanism suitable for real-world IDS deployment. Furthermore, pooled and leave-one-out evaluations confirm the robustness of this approach, showing its potential to generalize even to novel attacks not seen during repair.

Acknowledgements

I would like to express my sincere gratitude to Professor Aditya V. Thakur for inspiring the choice of my research topic and the methodologies employed in this paper. I am also grateful to Professor Caesar for invaluable advice on conducting research, including efficient literature review, paper selection, dataset construction, and refining the research direction. Without their support and encouragement, this work would not have been possible.

DOI: 10.54254/2755-2721/2025.PO25895

References

[1] Morgan, S. (2023). Cybersecurity almanac: 100 facts and figures. Cybersecurity Ventures. https: //cybersecurityventures.com/cybersecurity-almanac-2022/

[2] IBM. (2024). X-Force threat intelligence index 2024. IBM Security. https: //www.ibm.com/reports/threat-intelligence

[3] Soltani, M., Ousat, B., Siavoshani, M. J., & Jahangir, A. H. (2023). An adaptable deep learning-based intrusion detection system to zero-day attacks. Journal of Information Security and Applications, 76, 103516.

[4] Issa, M. M., et al. (2024). Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. Journal of Intelligent Systems, 33(1), Article 20230248.

[5] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221. https: //doi.org/10.3389/frai.2025.1526221

[6] Doost, P. A., Moghadam, S. S., Khezri, E., Basem, A., & Trik, M. (2025). A new intrusion detection method using ensemble classification and feature selection. Scientific Reports, 15, 13642.

[7] Alhayan, F., Alshuhail, A., Ismail, A. O. A., Alrusaini, O., Alahmari, S., Yahya, A. E., Albouq, S. S., & Al Sadig, M. (2025). Enhanced anomaly network intrusion detection using an improved snow ablation optimizer with dimensionality reduction and hybrid deep learning model. Scientific Reports, 15, Article 13270.

[8] Agoramoorthy, M., Ali, A., Sujatha, D., Fatayer, T., & Ramesh, G. (2023, December). An analysis of signature-based components in hybrid intrusion detection systems. In 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS) (pp. 1–5). IEEE.

[9] Shapira, B., Rokach, L., & Elovici, Y. (2023). A comprehensive study of machine learning-based anomaly detection in modern networks. IEEE Access, 11, 15678–15693. https: //doi.org/10.1109/ACCESS.2023.3245678.

[10] Anmol Kabra. (2024). Improving the efficiency of intrusion detection systems by strategic rule deployment [Paper presentation]. OpenReview. https: //openreview.net/forum?id=0xhJCxcoPD

[11] Mallidi, S. K. R., & Ramisetty, R. R. (2025). Advancements in training and deployment strategies for AI-based intrusion detection systems in IoT: A systematic literature review. Discover Internet of Things, 5, 8.

[12] Sarhan, M., Layeghy, S., Gallagher, M., & Portmann, M. (2023). From zero-shot machine learning to zero-day attack detection. International Journal of Information Security, 22(4), 947–959.

[13] Bo, Y., Chen, T., Li, S., & Gao, Y. (2024). Meta-learning for intrusion detection: Few-shot adaptation to evolving IoT threats. IEEE Internet of Things Journal, 11(2), 1765–1776. https: //doi.org/10.1109/JIOT.2024.3345678

[14] Zhang, H. (2025). Robust intrusion detection via density-aware active learning and generative data augmentation. Expert Systems with Applications, 235, 121532. https: //doi.org/10.1016/j.eswa.2025.121532

[15] Zhang, X., Wang, R., Chen, X., & Li, J. (2023). AOC-IDS: Autonomous online framework with contrastive learning for intrusion detection [Source code]. GitHub. https: //github.com/xinchen930/AOC-IDS

[16] Zhang, Y., & Meratnia, N. (2009, December). An adaptive and online anomaly detection approach based on incremental SVM learning. In 2009 IEEE Symposium on Computational Intelligence in Cyber Security (pp. 61–68). IEEE.