Layered Robustness Framework for Intrusion and Malware Detection Under Adversarial Network Threats

Arien
Department of Computer Science and Engineering
Punjabi University Patiala, India
Email: ariensingh9@gmail.com

Ram Singh
Department of Computer Science and Engineering
Punjabi University Patiala, India
Email: bhankharz@gmail.com

***Abstract*—Intrusion detection systems (IDS) and automated malware analysis platforms increasingly rely on learning-based detection models to process high-dimensional network flow features and executable representations. While such models demonstrate strong empirical performance under controlled evaluation settings, their deployment in adversarial network environments exposes structural vulnerabilities that extend beyond classifier-level weaknesses. Adversaries can manipulate feature representations,exploit decision boundary instability, abuse confidence calibration mechanisms, and leverage transferability across distributed detection architectures.**

This paper presents a system-level analysis of adversarial risk in intrusion and malware detection pipelines and argues that robustness must be treated as an end-to-end architectural property rather than a purely algorithmic objective. We formalize adversarial capabilities under realistic network deployment assumptions and introduce a structured vulnerability taxonomy spanning feature-space manipulation, boundary instability, confidence exploitation, transfer-based evasion, and deployment-level interactions.

Building upon this taxonomy, we propose a layered robustness framework that integrates representation stabilization, boundary regularization, calibration-aware monitoring, transfer mitigation strategies, and deployment-aware consistency validation. The proposed framework emphasizes operational feasibility in enterprise and cloud infrastructures where latency constraints, traffic drift, and adaptive adversaries must be considered.

By aligning defensive mechanisms with systemic vulnerability classes, this work provides a conceptual foundation for designing resilient intrusion and malware detection systems capable of sustaining robustness under evolving adversarial pressures.

***Index Terms*—Adversarial Machine Learning, Intrusion Detection Systems, Malware Detection, Network Security, Robust Deep Learning**

I. INTRODUCTION

Intrusion detection systems (IDS) and automated malware analysis platforms increasingly rely on data-driven detection models to process high-dimensional network flow features and executable representations [1], [2].These approaches enable the identification of complex behavioral patterns that are difficult to capture using traditional signature-based techniques.Deep neural and statistical models have demonstrated improved performance in modeling sequential traffic behavior and structural characteristics of malicious binaries [3].

Despite these advances, modern detection systems introduce new attack surfaces. When classification decisions are based on learned feature representations, small, carefully constructed perturbations may significantly alter the output of the model. Adversarial examples, originally studied in other machine learning domains [4], [5], have since been shown to impact intrusion detection and malware classification models [6]. In network environments, adversaries can manipulate packet timing distributions, flow-level statistics, or encoded attributes while preserving malicious functionality.

The operational impact of such attacks is particularly concerning in enterprise and cloud infrastructures where IDS systems are deployed in real-time monitoring pipelines [7]. The transferability of adversarial samples across models increases the risk for many distributed and federated detection architectures [8], [9]. An attacker capable of approximating decision boundaries through probing or surrogate modeling may systematically evade detection mechanisms.

Existing robustness strategies, including adversarial training and regularization-based defenses [10], [11], provide measurable resilience in constrained evaluation settings.However, these approaches typically address specific perturbation assumptions rather than systemic weaknesses throughout the detection life cycle.In practice, detection pipelines consist of multiple interacting components, including feature extraction, normalization, classification, and alert calibration. Weaknesses at any stage can undermine overall robustness.

This work adopts a system-level perspective on adversarial resilience in intrusion and malware detection systems. Advanced studies have shown that robustness improvements achieved through adversarial training or regularization often degrade under adaptive or transfer-based attacks [6], [12]. Furthermore, recent analysis emphasize that vulnerabilities frequently arise from interactions between feature extraction, model optimization, and deployment constraints rather than from classification models alone [13], [14].

Instead of focusing solely on algorithmic defenses, we analyze vulnerabilities across representation, optimization, confidence estimation, and deployment stages. Based on this analysis, we introduce a layered robustness framework de-

signed to reduce sensitivity to adversarial perturbations while maintaining feasibility for real-world network environments.

II. Background and Related Work

A. Learning-Based Intrusion Detection Systems

Machine learning has significantly influenced the evolution of intrusion detection systems over the past decade. Early discussions highlighted limitations of applying learning techniques directly to network security due to distributional drift and adversarial behavior [1]. Nevertheless, subsequent work demonstrated that deep architectures can effectively model complex traffic patterns, especially in large-scale network environments [3], [15].

Recurrent and convolution models have been widely adopted for modeling sequential network flow behavior [16]. More recent approaches incorporate hybrid and hierarchical feature extraction pipelines to improve generalization under dynamic traffic conditions. Despite improved detection accuracy, these systems remain dependent on statistical regularities that adversaries may exploit.

B. Contributions of This Work

This paper makes the following primary contributions:

- **System-Level Threat Formalization:** We formalize an adversarial threat model for intrusion and malware detection pipelines that extends beyond classifier-level perturbations to include feature extraction, calibration, and deployment interactions.
- **Five-Class Vulnerability Taxonomy:** We introduce a structured taxonomy spanning feature-space manipulation, decision boundary instability, confidence exploitation, transfer-based evasion, and deployment-level interaction weaknesses.
- **Layered Robustness Architecture:** We propose a five-layer robustness framework aligning defensive mechanisms with identified vulnerability dimensions to promote architectural resilience.
- **Deployment-Aware Evaluation Blueprint:** We provide a structured robustness evaluation methodology incorporating adaptive adversaries, distributional drift, and cross-node transferability scenarios.

C. Deep Learning in Malware Detection

Learning-based malware detection techniques leverage behavioral logs, opcode sequences, and structural presentation of the executable to classify malicious binaries [6].Visualization based and embedding based methods have decreased the reliance on handcrafted signatures, enabling scalable detection in large datasets. However, these approaches introduce sensitivity to perturbations in feature space, particularly when classification relies on learned embeddings rather than rule-based heuristics.

D. Adversarial Attacks in Security Systems

Adversarial machine learning research demonstrated that small carefully constructed perturbations can induce misclassification in high-capacity models [4], [17]. While initially explored in image recognition, adversarial techniques have since been adapted to network intrusion detection and malware classifiers [18]. Attackers may manipulate traffic statistics, inject adversarial payload patterns, or exploit feature extraction pipelines to evade deployed systems [7].

Transferability of adversarial samples further complicates defensive strategies. Adversarial inputs crafted against surrogate models may generalize to unseen classifiers [8]. This is particularly concerning for distributed detection infrastructures where consistent robustness across nodes is required.

E. Robustness and Defense Mechanisms

Defense mechanisms such as adversarial training and ensemble-based strategies aim to improve model robustness under constrained perturbation settings [10], [11]. Although these approaches demonstrate measurable improvements, they typically focus on algorithm-level resilience rather than architectural robustness. As detection pipelines involve multiple interacting components, vulnerabilities may persist even when individual classifiers exhibit improved adversarial tolerance.

III. Adversarial Threat Model

This section formalizes the adversarial assumptions considered in intrusion and malware detection environments.Rather than limiting the analysis to isolated classifier perturbations,we model the interaction between adversaries and the full detection pipeline.

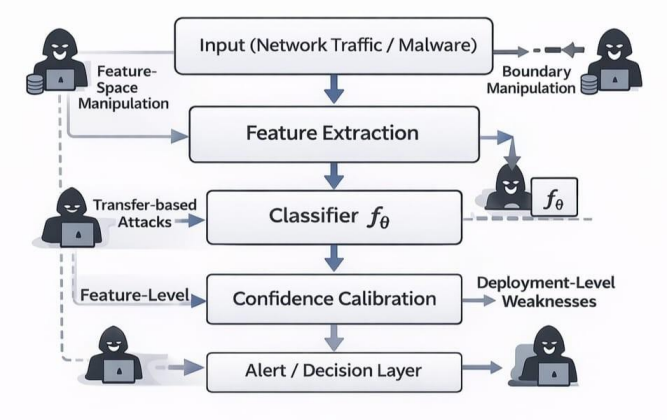

Fig. 1. System-level adversarial threat model illustrating attack surfaces across the intrusion detection pipeline.

A. System Model

We consider a detection pipeline D composed of four primary stages: feature extraction, representation encoding, classification, and decision calibration. Given an input sample $x \in X$, representing either network traffic flow features or malware representations, the system produces a prediction $f_\vartheta(x)$ parameterized by model parameters ϑ.

Unlike closed-world classification tasks, intrusion detection systems operate in dynamic environments where input distributions evolve over time [1]. Consequently, robustness must be evaluated under non-stationary and adversarial conditions.

B. Adversary Capabilities

We assume an adversary capable of manipulating observable input features while preserving malicious functionality. In network settings, this may involve modifying packet timing distributions, traffic statistics, or protocol-level attributes. In malware contexts, attackers may introduce structural or behavioral perturbations that alter feature embeddings without disrupting execution semantics [6].

We categorize adversarial knowledge into three levels:

- **White-box adversary:** Full knowledge of model parameters and gradients.
- **Black-box adversary:** Access limited to model queries and output responses.
- **Transfer-based adversary:** Ability to train surrogate models and exploit transferability properties [8], [19].

C. Attack Objectives

The adversary seeks to induce misclassification while minimizing detectable perturbations. Formally, given a loss function $L(f_\vartheta(x), y)$ and perturbation constraint set S, the adversarial objective can be expressed as:

$$\max_{\delta \in S} L(f_\vartheta(x + \delta), y) \quad (1)$$

where δ represents adversarial manipulation constrained by functional validity requirements. In intrusion detection contexts,constraints are not purely norm-based but often governed by protocol compliance and traffic plausibility.

D. Threat Surfaces in Deployment

Beyond classifier sensitivity,adversarial vulnerabilities emerge across multiple deployment stages. These include:

- **Feature extraction layer:** Manipulation of aggregation windows or encoding schemes.
- **Model optimization layer:** Exploitation of overfitting or gradient sensitivity.
- **Confidence estimation layer:** Abuse of overconfident predictions.
- **Distributed deployment layer:** Transferability across federated or multi-node detection systems [20], [21].

Recent studies emphasize that adversarial robustness failures often arise from systemic interactions between these components rather than from classification models alone [13]. Therefore, an effective defense strategy must consider architectural robustness across the entire detection life cycle.

E. Operational Assumptions and Constraints

In practical deployments, intrusion detection systems operate under computational and latency constraints. Real-time monitoring pipelines require bounded inference time and memory usage, limiting the feasibility of computationally intensive defenses. Additionally, network traffic distributions are non-stationary due to evolving user behavior and adaptive attackers.

We assume that adversaries may probe deployed systems through repeated interactions, observing output labels or confidence scores. This interaction model enables black-box approximation of decision boundaries over time. Consequently, robustness must account for iterative adversarial adaptation rather than single-shot perturbation attempts.

Furthermore, detection systems may be periodically updated or retrained using newly collected traffic data. Poisoning risks and distributional drift must therefore be considered as long-term threat vectors affecting system stability.

IV. Vulnerability Taxonomy

Based on the threat model defined above, we identify five systemic vulnerability classes that commonly affect intrusion and malware detection pipelines under adversarial conditions. Unlike prior work that focuses primarily on classifier robustness, this taxonomy characterizes weaknesses across the full detection life cycle.

A. Feature-Space Manipulation Vulnerability

Learning-based detection systems rely on high-dimensional feature representations derived from network flows or executable structures. Adversaries may introduce perturbations that alter statistical properties of feature vectors while preserving malicious behavior. Such manipulations exploit the reliance of classifiers on distributional regularities rather than semantic invariants [6], [7].

In network environments, attackers may modify packet timing intervals, padding behavior, or traffic burst patterns. In malware settings, perturbations may target opcode sequences or structural metadata. These vulnerabilities arise from the sensitivity of learned representations to small but strategically crafted feature changes.

In high-dimensional feature spaces typical of network flow representations, decision regions may form narrow margins around benign clusters. Adversaries can exploit these narrow margins by introducing statistically minor but directionally targeted perturbations. Such instability becomes amplified when training data insufficiently represents rare but legitimate traffic variations.

B. Decision Boundary Instability

Deep models often form highly non-linear decision boundaries in high-dimensional space. When training data does not sufficiently cover adversarial regions, small perturbations may shift inputs across class boundaries. Adversarial training mitigates this issue to some extent [10], yet empirical studies show that adaptive attacks can still exploit boundary irregularities [11].

Decision boundary instability is particularly problematic in evolving network environments where traffic distributions drift over time. Under such conditions, robustness guarantees established during training may not generalize to deployment settings.

C. Confidence Exploitation and Calibration Weakness

Many detection systems produce overconfident predictions even when inputs lie near adversarial regions. Attackers may exploit poorly calibrated confidence scores to bypass threshold-based alerting mechanisms. Miscalibrated outputs increase the risk of false negatives in real-time monitoring systems.

Confidence exploitation is not purely a classifier-level issue; it affects downstream decision processes such as alert prioritization and automated mitigation. Weak calibration amplifies the operational impact of adversarial evasion.

D. Transferability in Distributed Detection Architectures

Adversarial samples crafted against surrogate models often generalize to other models trained on similar data [8], [19]. In distributed intrusion detection systems, this property allows attackers to probe one node and deploy effective evasion strategies across multiple nodes.

Federated or multi-site deployments are especially vulnerable when models share similar architectures or feature extraction pipelines. Transferability therefore represents a systemic risk extending beyond single-model robustness.

E. Deployment-Level and Pipeline Interaction Vulnerability

Intrusion and malware detection systems consist of interacting components, including preprocessing, normalization, classification, and alert generation. Vulnerabilities may emerge from the interaction between these components rather than from the classifier alone. Recent analyses emphasize that adversarial robustness failures often arise from architectural and deployment-level weaknesses [13], [20].

For example, aggregation window choices, feature scaling methods, or asynchronous model updates may introduce exploitable inconsistencies. Addressing these vulnerabilities requires coordinated robustness strategies across the entire detection pipeline.

Fig. 2. Five-class vulnerability taxonomy spanning representation, boundary stability, calibration, transferability, and deployment interactions.

The proposed taxonomy highlights that adversarial risk in intrusion and malware detection is multi-dimensional. Effective defense mechanisms must therefore address representation sensitivity, boundary stability, calibration reliability, transferability, and deployment-level interactions in an integrated manner.

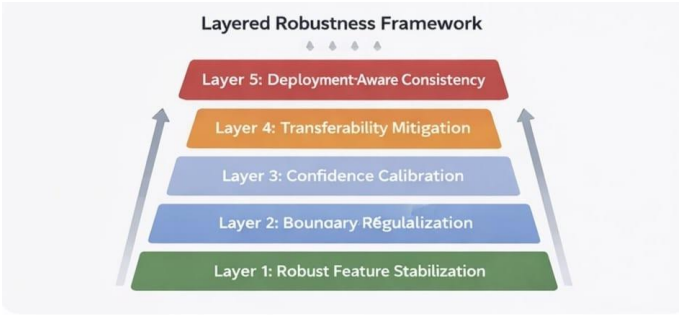

Fig. 3. Proposed five-layer robustness architecture aligned with identified vulnerability classes.

V. Proposed Layered Robustness Framework

Based on the vulnerability taxonomy defined in the previous section, we propose a five-layer robustness framework designed to strengthen intrusion and malware detection pipelines against adversarial manipulation. The framework emphasizes architectural resilience rather than isolated defensive techniques.

A. Robust Feature Stabilization Layer

To mitigate feature-space manipulation, the first layer focuses on stabilizing input representations before classification. This includes feature normalization strategies, aggregation consistency checks, and invariant representation learning. By reducing sensitivity to small statistical perturbations, the system limits adversarial leverage in high-dimensional feature space.

In network environments, this may involve enforcing consistency constraints on flow-level statistics across time windows. In malware analysis, structural embeddings may be regularized to preserve semantic invariants under minor perturbations.

B. Boundary Regularization Layer

To address decision boundary instability, the second layer incorporates robustness-aware training mechanisms such as adversarial training and margin regularization. Rather than relying solely on empirical accuracy, the objective is to smooth decision surfaces in regions likely to be exploited by adaptive adversaries [10].

This layer reduces susceptibility to gradient-based evasion and improves generalization under distributional shifts commonly observed in real-world deployments.

C. Confidence Calibration and Uncertainty Monitoring Layer

Given the risks associated with overconfident misclassification, the third layer introduces confidence calibration mechanisms. These include threshold adaptation, uncertainty estimation, and anomaly-aware scoring functions.

Instead of treating predictions as binary outputs,calibrated confidence values allow downstream systems to detect suspicious low-margin decisions and trigger additional verification procedures [22], [23].

D. Transferability Mitigation Layer

To reduce cross-model evasion risks, the fourth layer incorporates architectural diversity and ensemble-based strategies [11]. By introducing heterogeneity in model architectures, feature extraction pipelines, or training distributions, the system reduces the effectiveness of transfer-based adversarial attacks. This layer is particularly critical for distributed and federated intrusion detection infrastructures where attackers may exploit shared model characteristics.

E. Deployment-Aware Consistency Layer

The final layer addresses deployment-level vulnerabilities. This includes synchronization of feature preprocessing pipelines, periodic recalibration under traffic drift, and cross-node consistency monitoring in distributed systems.

Rather than assuming static operational conditions, the framework treats deployment as a dynamic environment requiring continuous robustness validation. Monitoring interactions between preprocessing, classification, and alert generation components reduces systemic weaknesses that adversaries may exploit.

F. Design Principles for Robust Deployment

The proposed architecture follows three guiding principles:

- **Layered Defense-in-Depth:** Robustness mechanisms should be distributed across preprocessing, model training, and deployment monitoring stages.
- **Adversary-Aware Evaluation:** Robustness must be validated under adaptive threat models rather than static perturbation assumptions.
- **Operational Feasibility:** Defensive techniques must respect latency and scalability constraints inherent to enterprise network monitoring systems.

Collectively, the proposed layered framework integrates representation stabilization, boundary regularization, calibration reliability, transfer resistance, and deployment-level validation into a unified robustness architecture. By aligning defensive mechanisms with identified vulnerability classes, the framework promotes systemic resilience in intrusion and malware detection systems operating under adversarial conditions.

VI. Robustness Analysis Perspective

A. Formal Definition of System-Level Robustness

Definition 1 (System-Level Robustness). Let D denote a detection pipeline composed of feature extraction F, representation encoding E, classifier f_ϑ, and calibration module C. Let S denote a structured perturbation set constrained by semantic validity and protocol compliance.

The pipeline D is said to be *system-level robust* under perturbation class S if:

$$\sup_{\delta \in S} \mathrm{L}(C(f_\vartheta(E(F(x+\delta)))), y) - \mathrm{L}(C(f_\vartheta(E(F(x)))), y) \leq \epsilon \quad (2)$$

for all admissible inputs x, where ϵ is a bounded robustness tolerance.

B. Multi-Dimensional Robustness Perspective

Unlike traditional adversarial formulations that rely solely on norm-bounded perturbations, robustness in network intrusion detection must be evaluated across multiple dimensions.

We define overall robustness R as a composite function:

$$\mathrm{R} = \alpha R_{rep} + \beta R_{bound} + \gamma R_{cal} + \delta R_{trans} + \eta R_{deploy} \quad (3)$$

where:

- R_{rep} denotes representation stability,
- R_{bound} denotes decision boundary smoothness,
- R_{cal} denotes calibration reliability,
- R_{trans} denotes transfer resistance,
- R_{deploy} denotes deployment consistency.

The coefficients $\alpha, \beta, \gamma, \delta, \eta$ reflect operational priorities in specific deployment environments.

The proposed layered framework is motivated by the observation that adversarial robustness in intrusion detection systems cannot be captured by a single robustness metric. Traditional adversarial robustness formulations typically assume bounded norm perturbations; however, network traffic manipulation is governed by protocol constraints, semantic validity, and operational plausibility.

From a theoretical standpoint, robustness in this context can be interpreted as reducing sensitivity of the detection function $f_\vartheta(x)$ to semantically preserving perturbations δ. Unlike image-based domains, perturbation constraints in network environments are structured and often non-convex. Therefore, robustness analysis must consider feasibility constraints rather than purely geometric norms.

Moreover, robustness should be evaluated across multiple dimensions: representation stability, boundary smoothness, calibration reliability, and cross-node consistency. A system may exhibit local classifier robustness while remaining vulnerable at the pipeline level. The layered framework addresses this multi-dimensional perspective by distributing defensive mechanisms across stages of the detection lifecycle.

This perspective shifts robustness evaluation from isolated adversarial accuracy metrics toward architectural resilience assessment.

VII. Security Implications and Future Research

The increasing reliance on learning-based intrusion and malware detection systems in enterprise, cloud, and critical infrastructure environments amplifies the consequences of adversarial failures. Unlike laboratory evaluation settings, operational deployments must contend with adaptive attackers, traffic distribution drift, and evolving threat landscapes. Robustness, therefore, cannot be treated as a static training objective but must be continuously validated under realistic adversarial conditions.

In large-scale enterprise and cloud infrastructures, adversarial evasion may enable persistent threats to bypass automated monitoring pipelines. In distributed intrusion detection systems, transferability across nodes increases systemic risk,

particularly when shared feature extraction pipelines are used. Federated or cross-organizational detection architectures introduce additional synchronization and consistency challenges that adversaries may exploit.

Future research should investigate certified robustness mechanisms tailored to network traffic constraints rather than purely norm-based perturbation models. Additionally, integrating uncertainty estimation with automated response systems remains an open challenge, as overly conservative thresholds may increase false positives while insufficient calibration increases false negatives.

Another promising direction involves deployment-aware robustness evaluation frameworks that simulate adaptive probing, traffic drift, and multi-stage adversarial strategies. Bridging the gap between theoretical robustness guarantees and real-world operational feasibility remains a central challenge in securing next-generation intrusion and malware detection systems.

Ultimately, advancing adversarial resilience in cybersecurity requires coordinated improvements across representation learning, system architecture design, calibration reliability, and deployment monitoring. Robustness must be treated as an end-to-end system property rather than an isolated classifier attribute.

VIII. Comparison with Existing Defense Strategies

Existing adversarial defense approaches in intrusion and malware detection primarily focus on adversarial training, feature denoising, or ensemble modeling. While these techniques improve robustness under constrained perturbation assumptions, they often neglect deployment-level interactions and system-wide consistency requirements.

Adversarial training improves boundary robustness but increases computational cost and may fail under adaptive or transfer-based attacks. Ensemble-based defenses reduce single-model sensitivity yet do not inherently address calibration weaknesses or preprocessing inconsistencies. Certified robustness methods offer theoretical guarantees but are often difficult to scale to high-dimensional network flow representations.

In contrast, the proposed layered framework does not replace algorithmic defenses but organizes them within a broader architectural structure. By aligning defenses with identified vulnerability classes, the framework integrates algorithmic robustness with deployment-aware validation, providing a more comprehensive resilience strategy [5], [24].

IX. Evaluation Blueprint

Although this work focuses on architectural design, practical validation of the proposed framework requires structured evaluation under adversarial conditions. A comprehensive assessment should consider multiple robustness dimensions.

First, representation robustness may be evaluated by measuring classification stability under statistically constrained perturbations in network flow features. Second, boundary robustness can be assessed using adaptive adversarial attacks under both white-box and black-box assumptions. Third, calibration reliability should be measured using confidence-error correlation metrics and uncertainty-aware detection thresholds.

In distributed deployment scenarios, robustness evaluation must incorporate transfer-based attack simulations across heterogeneous detection nodes. Furthermore, robustness should be tested under traffic distribution drift to assess long-term operational stability.

This evaluation blueprint provides a foundation for future empirical validation and establishes measurable criteria for assessing system-level adversarial resilience.

X. Conclusion

This paper presented a structured analysis of adversarial vulnerabilities in intrusion detection and malware analysis systems operating under realistic network deployment conditions. Rather than treating adversarial robustness as a narrow optimization objective confined to classifier training, we argued that resilience in cybersecurity systems must be approached as a multi-layer architectural property.

Through a formalized threat model, we identified that adversarial risk extends beyond feature perturbation and boundary manipulation. Vulnerabilities emerge from interactions between feature extraction mechanisms, statistical decision surfaces, confidence calibration processes, and distributed deployment constraints. The proposed five-class vulnerability taxonomy—spanning feature-space manipulation, decision boundary instability, confidence exploitation, transfer-based evasion, and deployment-level interaction weaknesses—provides a structured lens for understanding these systemic risks.

Building upon this taxonomy, we introduced a layered robustness framework designed to align defensive mechanisms with specific vulnerability dimensions. The framework integrates representation stabilization, boundary regularization, calibration-aware monitoring, transfer mitigation strategies, and deployment-aware consistency validation. Importantly, the architecture emphasizes operational feasibility, recognizing that intrusion detection systems must satisfy latency, scalability, and adaptability constraints in enterprise and cloud environments.

A central insight of this work is that adversarial robustness in cybersecurity cannot be reduced to adversarial accuracy metrics alone. Robustness must instead be evaluated as an end-to-end system property that accounts for adaptive adversaries, traffic distribution drift, and long-term deployment dynamics. Strengthening detection systems therefore requires coordinated improvements across representation learning, training methodology, confidence estimation, and architectural design.

Future research should focus on developing deployment-aware robustness benchmarks, certified defense mechanisms tailored to structured network perturbations, and adaptive monitoring strategies capable of detecting multi-stage adversarial campaigns. Bridging the gap between theoretical robustness guarantees and operational security requirements remains a critical open challenge.

By reframing adversarial resilience as an architectural design problem rather than a purely algorithmic one, this work aims to contribute toward the development of intrusion and malware detection systems capable of sustaining reliability under evolving adversarial pressures.

REFERENCES

[1] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *IEEE Symposium on Security and Privacy*, 2010.

[2] S. Rezaei and X. Liu, "Deep learning-based intrusion detection systems in cybersecurity: A review," *Computers & Security*, 2022.

[3] W. Wang *et al.*, "Hast-ids: Learning hierarchical spatial-temporal features using deep neural networks," *IEEE Access*, vol. 6, pp. 1792–1806, 2018.

[4] I. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[5] T. Nguyen *et al.*, "Adversarial machine learning in network security: Emerging threats and defenses," *IEEE Communications Surveys & Tutorials*, 2024.

[6] L. Demetrio *et al.*, "Explaining vulnerabilities of deep learning to adversarial malware binaries," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[7] B. Biggio *et al.*, "Evasion attacks against machine learning at test time," in *ECML PKDD*, 2013.

[8] N. Papernot *et al.*, "Transferability in machine learning," in *USENIX Security Symposium*, 2016.

[9] M. Rahman *et al.*, "Transfer-based adversarial attacks on network intrusion detection systems," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[10] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[11] F. Tramer *et al.*, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations*, 2018.

[12] W. Han *et al.*, "Robust malware detection in adversarial environments using feature consistency regularization," *Computers & Security*, 2023.

[13] S. Chen *et al.*, "Adversarial attacks and defenses in network intrusion detection systems: A survey," *IEEE Communications Surveys & Tutorials*, 2021.

[14] M. Liu *et al.*, "Deployment challenges of robust machine learning in security-critical systems," in *ACM CCS*, 2023.

[15] C. Yin *et al.*, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21 954–21 961, 2017.

[16] A. Javaid *et al.*, "A deep learning approach for network intrusion detection system," in *EAI SecureComm*, 2016.

[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.

[18] M. Rigaki and S. Garcia, "Adversarial deep learning against intrusion detection classifiers," in *Deep Learning and Security Workshop*, 2018.

[19] Q. Wang *et al.*, "On the transferability of adversarial examples in network intrusion detection," in *NDSS Symposium*, 2022.

[20] R. Sheatsley *et al.*, "Evaluating intrusion detection robustness against adaptive adversaries," in *USENIX Security Symposium*, 2023.

[21] K. Zheng *et al.*, "Federated intrusion detection under adversarial conditions," *IEEE Transactions on Network and Service Management*, 2023.

[22] P. Xu *et al.*, "Uncertainty-aware deep learning for security-critical applications," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[23] B. Liang *et al.*, "Certified adversarial robustness for deep learning-based security systems," *IEEE Transactions on Information Forensics and Security*, 2022.

[24] H. Qiu *et al.*, "Adaptive evasion attacks against learning-based intrusion detection," in *NDSS Symposium*, 2023.