A Theoretical Framework for Adversarial Robustness in Real-Time ML-Based Intrusion Detection Systems (IDS)

Ogechukwu Scholastica Onyenaucheya

Computer Information Systems, Prairie View A&M University, 100 University Drive, Prairie View, Texas 77446, United States of America

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110100143

Received: 08 February 2025; Accepted: 11 February 2026; Published: 20 February 2026

ABSTRACT

Machine learning-based intrusion detection systems (IDS) are now commonly used in real-time cybersecurity to defend against quickly changing threats. However, recent developments in adversarial machine learning have shown that many IDS models are still very susceptible to adaptive attacks, especially in real-time conditions. Most existing research examines adversarial robustness in offline or static scenarios, missing the dynamic nature of live network traffic, ongoing data flows, and strict time requirements. This gap limits the effectiveness of current adversarial defense strategies in practical intrusion detection systems.

This paper aims to fill this gap by proposing a theoretical framework for adversarial robustness in real-time machine learning-based intrusion detection systems. The framework treats adversarial robustness as a characteristic that changes over time, influenced by detection delays, attacker strategies, concept drift, and ongoing interactions between models and adversaries. We introduce formal concepts such as time-to-evasion, detection stability, and robustness decay to describe how IDS reacts under lasting adversarial pressure.

Instead of creating a new detection algorithm, this study gives a theoretical viewpoint that clarifies why many adversarial defenses perform well in offline tests but struggle in real-time scenarios. The framework applies to various IDS designs and machine learning methods. By connecting adversarial machine learning theory with the needs of real-time intrusion detection, this work lays the groundwork for future testing, comparisons, and development of resilient IDS for challenging operational environments.

Keywords: Adversarial Machine Learning; Intrusion Detection Systems; Real-Time Cybersecurity; Adversarial Robustness; Machine Learning Security.

INTRODUCTION

Background of the Study

The rapid increase in network systems, cloud computing, Internet of Things (IoT) devices, and digital services has made modern cyber threats more complex and widespread. Traditional rule-based and signature-driven intrusion detection systems (IDS) are increasingly ineffective against sophisticated and evolving attack patterns, particularly zero-day exploits and polymorphic malware [1,2]. As a result, machine learning (ML)–based intrusion detection systems have been widely adopted to enhance detection accuracy, scalability, and adaptability in complex network environments [3,4].

ML-based IDS uses statistical learning and pattern recognition techniques to find unusual or harmful behaviors in network traffic, system logs, and user activity. These systems have demonstrated improved detection capabilities compared to conventional approaches, especially in handling high-dimensional and large-scale data [5]. Consequently, ML-driven IDS are now deployed in **real-time operational environments**, including security operations centers (SOCs), cloud platforms, and critical infrastructure networks, where decisions must be made within strict time constraints [6].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

However, recent advances in adversarial machine learning (AML) have revealed that ML models are inherently vulnerable to carefully crafted adversarial inputs designed to manipulate model predictions [7,8]. In intrusion detection contexts, adversaries can subtly modify network traffic features—such as packet timing, flow statistics, or protocol fields—to evade detection while maintaining malicious intent [9]. These adversarial attacks pose a significant threat to the reliability and trustworthiness of ML-based IDS, particularly when deployed in mission-critical environments.

While extensive research has investigated adversarial attacks and defenses in IDS, most studies evaluate robustness under offline or static conditions, relying on pre-collected datasets and batch inference [10,11]. Such evaluation settings fail to capture the operational realities of real-time intrusion detection, where data arrives continuously, network behavior evolves, and attackers adapt their strategies dynamically. This mismatch between research assumptions and deployment conditions motivates a deeper theoretical examination of adversarial robustness in real-time IDS.

Unlike earlier studies that suggest specific adversarial defense mechanisms or test IDS robustness in limited conditions, this work does not propose a new algorithm or defense method. Instead, it provides a unified theoretical viewpoint that clarifies why adversarial robustness weakens in real-time intrusion detection systems, regardless of the model architecture. By viewing adversarial robustness as a time-dependent property influenced by latency, adversarial adaptation, and concept drift, this study offers a foundational framework for evaluating existing and future IDS defenses. To our knowledge, no prior study has presented a theory-focused approach that links adversarial machine learning to real-time operational constraints in intrusion detection systems.

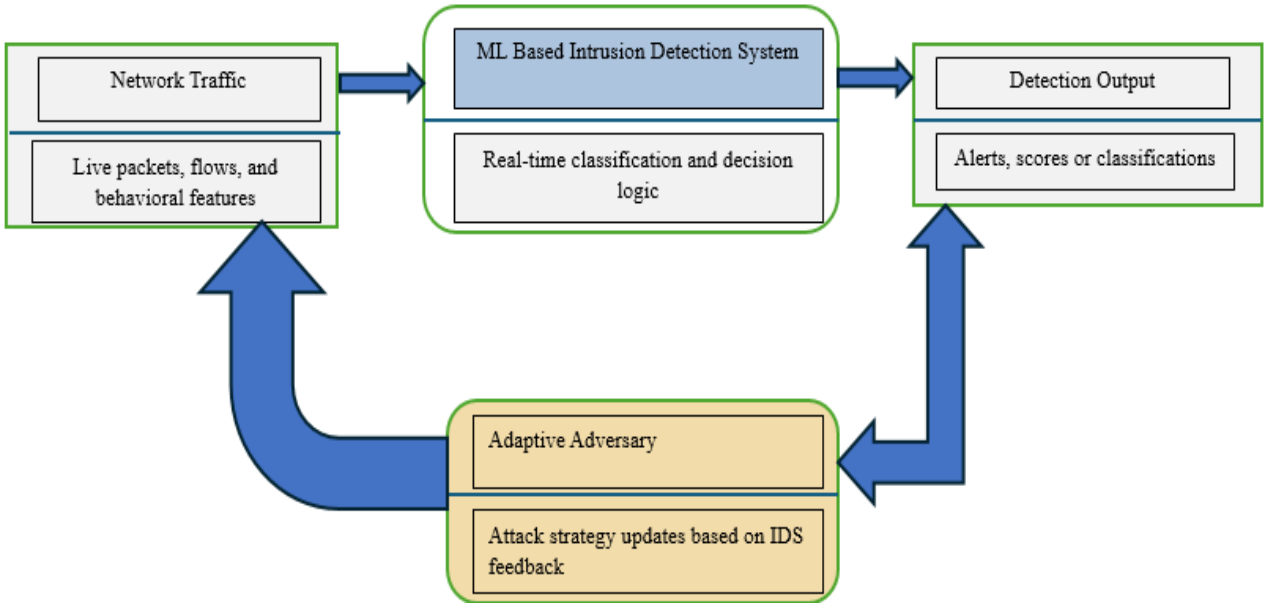

Figure 1: Real-time adversarial interaction loop between streaming network traffic, ML-based IDS, and adaptive attacker.

Problem Statement

Despite the growing body of literature on adversarial machine learning in intrusion detection systems, a fundamental problem persists: the lack of a unified theoretical understanding of adversarial robustness in real-time ML-based IDS. Existing research predominantly focuses on improving classification accuracy or attack resistance under static evaluation scenarios, overlooking the temporal dynamics and operational constraints inherent in real-time environments [12].

Real-time intrusion detection systems must process high-velocity data streams, deliver detection decisions within bounded latency, and remain effective under continuous adversarial pressure. In such environments, adversarial robustness is not a static property of a trained model but a dynamic characteristic influenced by detection latency,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

model adaptation rates, concept drift, and attacker learning behavior [13]. However, current metrics like accuracy loss or attack success rate do not adequately capture these time-related interactions.

As a result, many adversarial defense mechanisms that perform well in controlled experimental settings fail when deployed in operational IDS environments [14]. This failure exposes security infrastructures to undetected attacks and undermines confidence in ML-based detection technologies. Without a theoretical framework that models real-time constraints and adversarial interactions over time, it is challenging to assess, compare, or design strong IDS defenses suitable for real-world use.

Objectives of the Study

The primary objective of this study is to address the theoretical gap in understanding adversarial robustness within real-time ML-based intrusion detection systems. Specifically, the study aims to:

1. Analyze the weaknesses of current adversarial machine learning methods when applied to real-time intrusion detection.
2. Develop a theoretical framework that defines adversarial robustness as a time-dependent feature in streaming IDS environments.
3. Identify and outline robustness measures that take into account latency, time interactions, and adaptive adversarial behavior.
4. Examine the trade-offs between detection accuracy, computational costs, and adversarial resilience in real-time IDS.
5. Offer fundamental insights to guide the development and assessment of future IDS defenses under realistic operational limits.

By concentrating on theory rather than implementation, this study aims to create a conceptual foundation that can support various IDS designs and defense strategies.

Significance of the Study

This study is significant for its potential to connect adversarial machine learning research with real-world intrusion detection applications. By presenting a theoretical framework designed for real-time IDS environments, this work provides several important advantages.

First, it gives researchers a way to understand why current adversarial defenses often fail in operational settings, despite performing well in offline tests. Second, it introduces time-sensitive robustness considerations that go beyond traditional accuracy-based metrics. Third, the framework assists in developing evaluation methods for IDS that more accurately reflect the real-time challenges faced by security operations.

From a practical standpoint, this study helps cybersecurity experts and system developers make informed decisions about implementing ML-based IDS in adversarial scenarios. Knowing the trade-offs between detection speed and robustness enables more realistic risk evaluations and defense strategies. Finally, by establishing basic theoretical concepts, this work paves the way for future empirical research and algorithmic and standards-focused studies in robust intrusion detection.

LITERATURE REVIEW

Overview of Machine Learning–Based Intrusion Detection Systems

Intrusion Detection Systems (IDS) are critical parts of cybersecurity infrastructures, set up to monitor network and system activities for harmful behavior. Traditional IDS approaches rely on signature-based or rule-based techniques, which are effective against known attacks but struggle to detect novel or evolving threats such as

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

zero-day exploits and polymorphic malware [1,2]. To address these limitations, machine learning (ML) techniques have been increasingly adopted in IDS design.

ML-based IDS employ supervised, unsupervised, and semi-supervised learning algorithms to model normal and malicious behavior from data.

Techniques such as support vector machines, decision trees, ensemble learning, and deep neural networks have demonstrated improved detection accuracy and scalability compared to traditional approaches [3–5]. Deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been particularly effective in capturing complex patterns in high-dimensional network traffic data [6].

With the expansion of high-speed networks, cloud computing, and IoT ecosystems, ML-based IDS are increasingly deployed in **real-time environments** where they must process continuous streams of data with minimal latency [7].

In such settings, IDS are required not only to be accurate but also computationally efficient and stable under dynamic network conditions. These operational requirements introduce new challenges that go beyond conventional offline IDS evaluation.

To clarify the key differences between traditional offline evaluations and operational IDS environments, Table 1 compares assumptions underlying offline and real-time assessments of adversarial robustness.

Dimension	**Offline ML-Based IDS Evaluation**	**Real-Time ML-Based IDS Evaluation**
Data Characteristics	Static, pre-collected datasets	Continuous, streaming network traffic
Adversary Model	Fixed or pre-defined attacker	Adaptive adversary reacting to IDS outputs
Temporal Dynamics	Ignored	Central to detection and robustness
Latency Consideration	Not modeled	Critical operational constraint
Model Updates	Batch retraining	Online or incremental adaptation
Robustness Measurement	Accuracy under attack	Sustained detection over time
Deployment Fidelity	Low	High

Table 1: Key differences between offline and real-time assumptions in adversarial robustness evaluation for intrusion detection systems.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

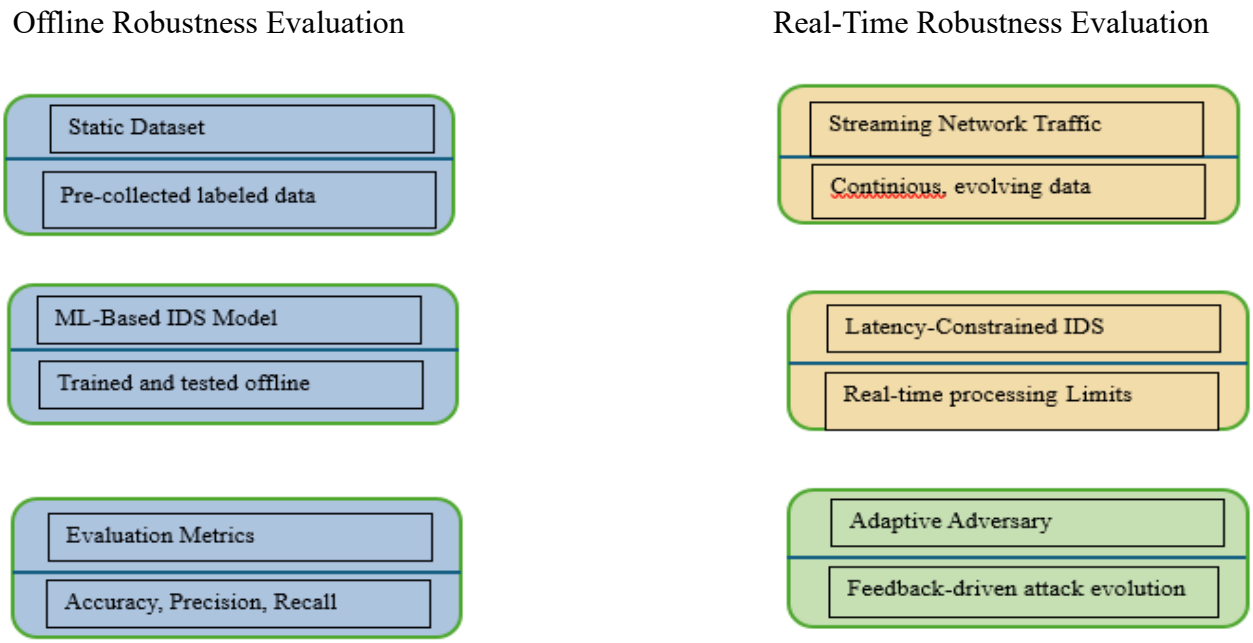

Figure 2: Structural comparison between offline adversarial evaluation and real-time streaming robustness assessment.

These differences explain why robustness guarantees from offline tests often do not apply to real-time intrusion detection systems.

Adversarial Machine Learning in Intrusion Detection

Adversarial machine learning (AML) studies how ML models are vulnerable to inputs designed to manipulate predictions. In the context of intrusion detection, adversarial attacks aim to evade detection by subtly altering network traffic features while preserving the underlying malicious functionality [8,9]. Common attack strategies include evasion attacks at inference time, poisoning attacks during model training, and model extraction attacks [10].

Many studies have shown that ML-based intrusion detection systems (IDS) are very prone to adversarial evasion attacks. Gradient-based methods, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), have been adapted to network traffic features to generate adversarial samples that significantly reduce detection accuracy [11,12]. More advanced approaches use generative adversarial networks (GANs) to synthesize realistic malicious traffic that bypasses IDS detection [13].

These results raise important concerns about the trustworthiness of ML-driven IDS in adversarial situations. As a result, numerous defense mechanisms have been proposed, including adversarial training, ensemble learning, feature squeezing, and input preprocessing techniques [14–16]. While these defenses improve robustness under certain conditions, their effectiveness is often evaluated using static datasets and offline testing methodologies.

Limitations of Existing Adversarial Defense Approaches

Despite promising results in controlled experiments, existing adversarial defenses exhibit significant limitations when applied to real-world IDS deployments. One main problem is the computational demand of many defense mechanisms. Adversarial training, for example, increases training and inference complexity, making it difficult to deploy in high-throughput real-time environments [17]. Similarly, ensemble-based defenses improve robustness by combining multiple classifiers but often introduce unacceptable detection latency [18].

Another limitation is the lack of adaptability to evolving adversarial strategies. Most defenses are designed to counter specific attack types and assume a fixed threat model. In practice, attackers continuously adapt their methods based on observed system behavior, rendering static defenses ineffective over time [9,19].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Furthermore, many studies overlook the impact of concept drift, where the statistical properties of network traffic change over time due to evolving user behavior, network configurations, or attack techniques [20].

Concept drift can degrade IDS performance even in non-adversarial settings, and its interaction with adversarial attacks further complicates robustness assessment. Existing AML defenses rarely consider this combined effect.

Real-Time Constraints and Streaming Data Challenges

Real-time intrusion detection has constraints that change the basic assumptions of most AML research. In streaming environments, data arrives continuously, and IDS must process and classify inputs within strict latency bounds to be operationally useful [7,21]. Delayed detection can render alerts ineffective, particularly for fast-moving attacks.

Most AML studies assume batch processing and unlimited inference time, which is unrealistic in real-time IDS contexts. Robustness metrics such as accuracy, precision, or attack success rate fail to capture **temporal performance**, such as detection delay or stability over time [22].

As a result, there is limited understanding of how adversarial robustness evolves during prolonged interactions between attackers and IDS.

Additionally, real-time IDS operate under resource constraints that limit the feasibility of complex defense mechanisms. High-dimensional feature extraction, frequent model updates, and continuous adversarial adaptation create trade-offs between robustness, scalability, and latency that are rarely addressed in existing literature [23].

Identified Research Gaps

Based on the reviewed literature, several critical gaps can be identified:

1. **Absence of a Unified Theoretical Framework:**

   There is no clear theoretical model that explains adversarial robustness in ML-based IDS specifically designed for real-time streaming environments. Current work mainly focuses on algorithmic defenses without outlining robustness as a dynamic property that changes over time.

2. **Lack of Temporal Robustness Metrics:**

   Current evaluation metrics are static and fail to measure latency, detection stability, and long-term decline under ongoing adversarial pressure.

3. **Limited Integration of Concept Drift and AML:**

   Few studies investigate how concept drift interacts with adversarial attacks in real-time IDS, despite both being common issues in operational settings

4. **Disconnect Between Research and Deployment:**

   Many defenses that work well in offline tests do not perform in real-time environments due to computational demands, delayed detection, and the adaptability of adversaries.

These gaps highlight the need for a theoretical approach that models adversarial robustness in real-time intrusion detection systems. Addressing this gap is crucial for both academic progress and practical deployment of resilient IDS technologies.

Based on the literature reviewed, Table 2 summarizes the key unresolved limitations that motivate this study.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Research Dimension	State of Existing Research	Observed Limitation
Robustness Evaluation	Predominantly offline testing	Fails to capture real-time dynamics
Temporal Modeling	Rarely considered	No time-dependent robustness theory
Adversary Adaptation	Mostly static attack models	Ignores feedback-driven attackers
Latency Awareness	Minimally addressed	Unrealistic deployment assumptions
Robustness Metrics	Accuracy and F1-score	No endurance or stability metrics
Theoretical Foundations	Fragmented, model-specific	No unified theoretical framework

Table 2: Summary of unresolved gaps motivating a theoretical framework for adversarial robustness in real-time IDS.

Notably, the absence of a unified, time-aware theoretical framework remains a consistent gap across existing studies.

METHODOLOGY

This study uses a theoretical and analytical research methodology aimed at creating a unified framework for understanding adversarial robustness in real-time machine learning-based intrusion detection systems (IDS). Instead of proposing or testing a specific detection algorithm, the methodology emphasizes formal modeling, conceptual analysis, and synthesis of existing literature on adversarial machine learning and real-time systems. This approach is fitting given the study's goal of establishing foundational insights relevant across different IDS architectures and deployment contexts.

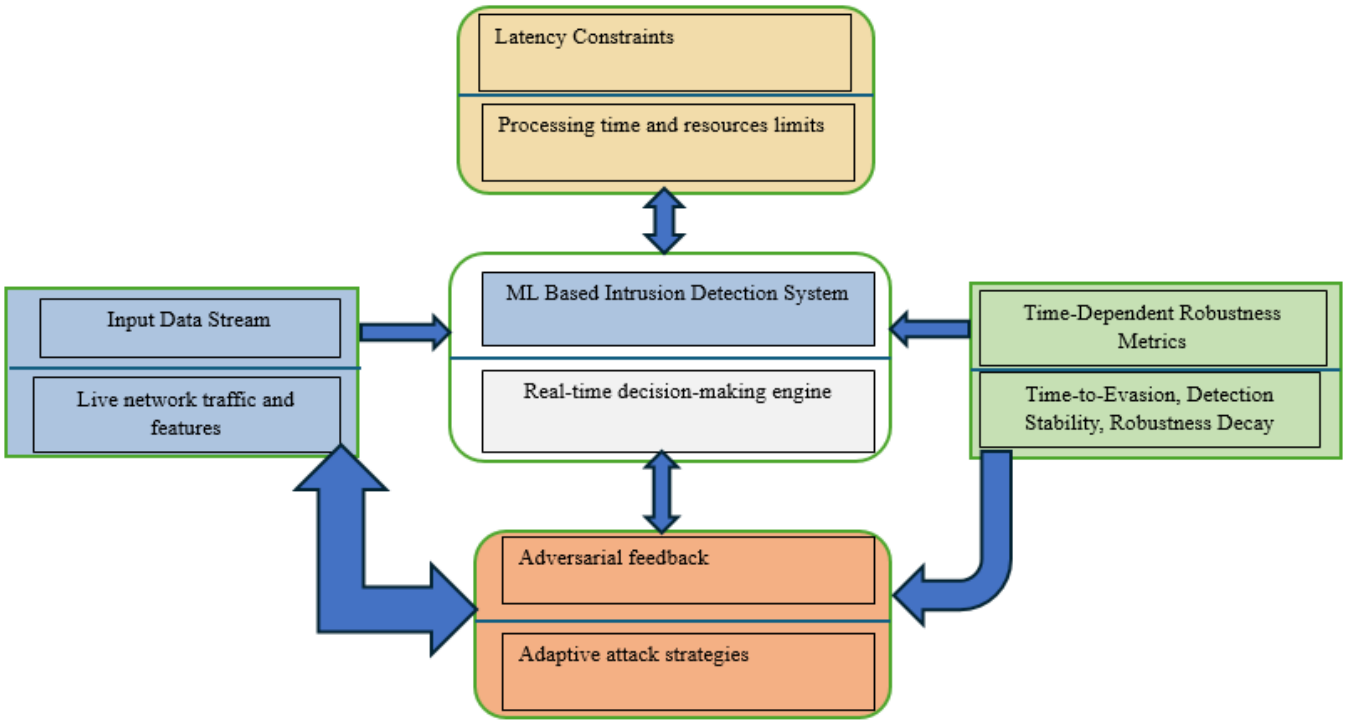

Figure 3: Time-dependent adversarial robustness framework incorporating latency, adaptation, and concept drift.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Research Design

The research design follows a conceptual-theoretical framework development approach, which is commonly used in cybersecurity and machine learning research to address foundational gaps where empirical benchmarks or standardized datasets are insufficient [24]. The study integrates principles from adversarial machine learning, streaming data analytics, and real-time system design to construct a coherent analytical model.

The methodology consists of four sequential phases:

1. Formalization of real-time IDS operational constraints
2. Definition of an adversarial threat model tailored to streaming environments
3. Development of time-aware robustness metrics
4. Analytical examination of adversarial–latency trade-offs

Each phase builds upon the previous one to ensure internal consistency and logical progression.

The absence of empirical experimentation is intentional, aligning with the study's aim of establishing theoretical foundations rather than making system-specific performance claims.

Similar theory-driven approaches have been widely used in adversarial machine learning and security research to define threat models, robustness properties, and evaluation principles before empirical validation. The framework proposed here is meant to guide and shape future experimental studies rather than replace them.

The proposed theoretical framework introduces a set of latency-aware and time-dependent robustness constructs, summarized in Table 3, to formalize adversarial behavior in real-time IDS environments.

Theoretical Construct	**Formal Description**	**Role in Real-Time IDS Robustness**
Time-to-Evasion	Duration before an adversary successfully bypasses detection	Measures robustness endurance
Detection Stability	Consistency of detection outcomes over time	Indicates reliability under attack
Robustness Decay	Rate of performance degradation during adversarial interaction	Captures sustained attack impact
Latency Impact	Effect of processing delay on detection effectiveness	Links theory to operational constraints
Adversarial Feedback Loop	Continuous interaction between attacker and IDS	Models adaptive adversary behavior

Table 3: Theoretical constructs defining adversarial robustness in real-time ML-based intrusion detection systems.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Formal Real-Time IDS Model

In contrast to conventional offline classifiers, a real-time intrusion detection system (IDS) operates on a continuous stream of network observations and must produce decisions within bounded latency constraints. Accordingly, the IDS is modeled as a time-indexed decision function defined over a streaming stochastic process.

Let the incoming network traffic be represented as a sequence:

$$\{(x_t, y_t)\}_{t=1}^{\infty}$$

where $x_t \in \mathbb{R}^d$denotes the feature vector extracted from network activity at time t, and $y_t \in \{0,1\}$represents the corresponding ground-truth label. The IDS classifier is defined as:

$$f_{\theta_t}: \mathbb{R}^d \to \{0,1\}$$

where $\theta_t$denotes model parameters that may evolve over time due to incremental learning, retraining, or environmental drift.

A key operational constraint in real-time IDS is bounded detection latency. Let $\delta_t$denote the time elapsed between data arrival and decision output. The system must satisfy:

$$\delta_t \leq \Delta_{max}$$

where Δ_{max}represents the maximum permissible latency determined by deployment requirements.

This formulation establishes the IDS as a streaming, latency-constrained, time-evolving classification process.

Such modeling is essential for analyzing adversarial robustness under realistic operational conditions, where both predictive accuracy and timeliness jointly determine system effectiveness.

Adversarial Threat Model for Streaming IDS

To model adversarial behavior in real-time environments, this study expands traditional AML threat models by including temporal and adaptive characteristics.

Adversary Capabilities

The adversary is assumed to:

- Generate adversarial network traffic that maintains malicious functionality,
- Adapt attack strategies based on observed detection results,
- Operate under partial knowledge of the IDS model parameters.

Unlike static white-box or black-box models, the adversary interacts continuously with the IDS, learning from detection feedback over time.

Temporal Interaction Model

The adversarial process is modeled as a sequence of interactions:

$$x_t^{adv} = x_t + \delta_t$$

where $\delta_t$represents time-varying adversarial perturbations. The effectiveness of these perturbations evolves as the adversary refines its strategy, highlighting the need to consider robustness as a dynamic property.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Time-Dependent Robustness Modeling

In contrast to traditional static robustness definitions, this study conceptualizes robustness as a function of time:

Define expected adversarial risk:

$$R_{adv}(t) = \mathbb{E}_{(x_t,y_t)\sim D_t}\left[\max_{\|\delta_t\|\leq\epsilon} \mathcal{L}(f_{\theta_t}(x_t+\delta_t), y_t)\right]$$

Then define robustness:

$$\mathcal{R}(t) = 1 - R_{adv}(t)$$

Now robustness decay rate:

$$\frac{d\mathcal{R}(t)}{dt} < 0$$

where $\mathcal{L}$denotes a loss function reflecting misclassification or delayed detection.

This formulation illustrates how adversarial pressure and system adaptation together affect robustness over extended operations.

We expect robustness to decline when adversarial adaptation outpaces the IDS's learning or updating processes.

Latency-Aware Robustness Metrics

To evaluate adversarial robustness in real-time IDS, the study introduces three conceptual metrics:

Time-to-Evasion (TTE)

TTE measures the duration required for an adversary to achieve consistent evasion:

Define:

$$TTE = \inf\{t > 0 \mid R_{adv}(t) \geq \tau\}$$

where:

- τ= evasion success threshold

Current robustness metrics in adversarial machine learning mainly evaluate immediate classification behavior under fixed perturbation budgets. While suitable for offline analysis, these metrics fall short for real-time intrusion detection, where adversarial effectiveness unfolds over prolonged interactions.

The proposed metrics, Time-to-Evasion and detection stability do not replace traditional robustness measures but enhance them by capturing temporal endurance and consistency under adaptive adversarial pressure. As such, they fit the realities of real-time IDS and address limitations found in past evaluation methodologies.

Detection Stability

Detection stability quantifies the variance of detection outcomes under continuous adversarial perturbation:

Define prediction variance:

$$S(T) = \frac{1}{T}\sum_{t=1}^{T}\left(f_{\theta_t}(x_t) - \bar{f}\right)^2$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Low $S(T)$= high stability.

High variance indicates unstable detection and increased susceptibility to evasion.

Latency–Robustness Trade-off

This metric captures the inverse relationship between detection latency and adversarial resilience:

Where:

- C_t= computational complexity
- Show empirically plausible inverse proportionality:

$$\frac{\partial \mathcal{R}}{\partial \delta_t} < 0$$

indicating that increased processing time may degrade real-time robustness.

Concept Drift and Model Adaptation Considerations

Concept drift is modeled as a gradual shift in the underlying data distribution:

$$P_t(x, y) \neq P_{t+k}(x, y)$$

In real-time IDS, adversarial perturbations and concept drift may occur simultaneously, accelerating robustness degradation. The methodology accounts for this interaction by incorporating drift-aware assumptions into the robustness function $R(t)$.

Analytical Evaluation Approach

The proposed framework is assessed through logical consistency analysis, a synthesis of existing empirical findings, and theoretical reasoning based on adaptive machine learning (AML) literature. This study does not depend on experimental validation; instead, it seeks to establish principles that can guide future research and system design.

RESULTS

The results here come from the formal framework in the Methodology. Since this study is theoretical, the findings are logical and mathematical outcomes of the proposed time-dependent adversarial robustness model rather than empirical benchmarks.

The results focus on (i) the changes in robustness over time, (ii) adversarial endurance measured by Time-to-Evasion (TTE), (iii) the interaction between latency and robustness, and (iv) the compounded degradation from concept drift.

Time-Dependent Robustness Dynamics

From methodology, adversarial risk is defined as:

$$R_{adv}(t) = \mathbb{E}_{(x_t, y_t) \sim D_t}\left[\max_{\|\delta_t^{adv}\| \leq \epsilon} \mathcal{L}(f_{\theta_t}(x_t + \delta_t^{adv}), y_t)\right]$$

and time-dependent robustness is defined as:

$$\mathcal{R}(t) = 1 - R_{adv}(t)$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Under adaptive adversarial interaction, the perturbation strategy evolves as a function of prior detection outcomes. If adversarial refinement increases misclassification likelihood over time, then:

$$\frac{dR_{adv}(t)}{dt} > 0$$

which directly implies:

$$\frac{d\mathcal{R}(t)}{dt} < 0$$

This conclusion formalizes why robustness guarantees obtained at deployment time cannot be assumed to persist indefinitely in operational settings.

Synthetic Numerical Illustration of Robustness Decay

To provide quantitative intuition, consider a simplified streaming IDS scenario.

Assume:

- Initial adversarial risk:

$$R_{adv}(0) = 0.05$$

(i.e., 5% adversarial misclassification rate)

- Adversarial adaptation increases risk linearly at rate $\alpha = 0.02$per time unit:

$$R_{adv}(t) = 0.05 + 0.02t$$

Then robustness becomes:

$$\mathcal{R}(t) = 1 - (0.05 + 0.02t)$$

Let the evasion threshold be defined as:

$$\tau = 0.5$$

Time-to-Evasion (TTE) is obtained by solving:

$$0.05 + 0.02t = 0.5$$
$$t = 22.5$$

Thus:

$$TTE = 23 \text{ time units}$$

This minimal synthetic example demonstrates that even moderate adversarial adaptation rates can produce predictable robustness collapse within a finite time horizon. Importantly, the model shows that endurance, not only instantaneous accuracy, determines operational resilience.

The example also illustrates how TTE provides a more informative robustness measure than static accuracy metrics. Two IDS models with identical initial performance may exhibit substantially different TTE values depending on their ability to resist adaptive refinement

Sensitivity to Adaptation Rate

The synthetic model can be generalized as:

$$R_{adv}(t) = R_{adv}(0) + \alpha t$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Solving for TTE:

$$TTE = \frac{\tau - R_{adv}(0)}{\alpha}$$

This reveals:

- TTE is inversely proportional to adversarial adaptation rate α.
- Small increases in adversarial learning efficiency dramatically reduce system endurance

Latency–Robustness Interaction

From Section 3, robustness is influenced by detection latency:

$$\mathcal{R}(t) = g(\delta_t, C_t)$$

Operationally:

$$\frac{\partial \mathcal{R}}{\partial \delta_t} < 0$$

To illustrate, suppose detection latency increases from 100ms to 300ms due to added computational defense layers. Even if classification accuracy improves slightly offline, the delayed response window allows adversarial actions to propagate further before mitigation.

Thus, robustness must be evaluated jointly with timeliness.

Detection Stability as Early Warning Indicator

Detection stability was defined as:

$$S(T) = \frac{1}{T}\sum_{t=1}^{T}\left(f_{\theta_t}(x_t) - \bar{f}\right)^2$$

Under increasing adversarial adaptation:

- Variance increases before complete evasion occurs.
- Instability precedes total robustness collapse.

This suggests practical deployment value in monitoring prediction consistency rather than solely aggregate accuracy.

Compounded Degradation Under Concept Drift

Let streaming distribution evolve:

$$D_t \neq D_{t+1}$$

If both adversarial adaptation (α_{adv}) and concept drift (α_{drift}) influence degradation:

$$\frac{d\mathcal{R}(t)}{dt} = -(\alpha_{adv} + \alpha_{drift})$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

This analytical finding explains empirical reports where IDS degrades faster in live deployments than under controlled adversarial benchmarks

Consolidated Analytical Insights

The proposed framework leads to these theoretical outcomes:

- Robustness changes over time instead of being a fixed property.
- Time-to-Evasion measures adversarial endurance and adds to static accuracy metrics.
- The rate at which adversaries adapt affects the lifespan of robustness.
- Detection latency negatively impacts operational resilience.
- Detection instability appears before full evasion and may act as an early warning signal.
- Concept drift increases adversarial degradation in streaming environments.

Together, these results explain the performance gap between offline adversarial robustness evaluations and real-world IDS deployments.

Future Research Directions

Building on the theoretical results, several avenues for future research are identified:

1. **Empirical Validation of Time-Aware Metrics:**

   Experimental studies should confirm TTE and detection stability using real-time network traffic datasets.

2. **Adaptive Learning Mechanisms:**

   Explore online and continual learning methods that can counter adversarial changes without excessive latency.

3. **Hybrid Human-AI Defense Models:**

   Explore collaborative frameworks where human analysts intervene when detection instability is detected.

4. **Benchmarking Standards:**

   Develop standardized benchmarks and evaluation protocols for adversarial robustness in streaming IDS environments.

DISCUSSION

The findings of this study provide significant theoretical insights into how machine learning-based intrusion detection systems (IDS) behave under adversarial conditions in real-time settings. By framing adversarial robustness as a time-dependent and interaction-driven property, this work enhances understanding of why many existing IDS defenses fail in real-world cases despite strong performance in controlled tests. This discussion places the results within the context of existing literature and highlights their implications for research and practice.

Revisiting Adversarial Robustness in IDS

Most prior research on adversarial machine learning in intrusion detection evaluates robustness using static performance metrics such as classification accuracy, precision, and recall [1–3]. While these metrics are informative in controlled settings, they fail to capture the temporal dynamics of real-time detection systems. The present study's findings challenge the implicit assumption in much of the literature that robustness is a fixed attribute of a trained model.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

By demonstrating that robustness degrades over time under sustained adversarial interaction, this work aligns with recent observations that attackers adapt their strategies in response to IDS behavior [4,5]. However, unlike existing studies that primarily report empirical performance drops, this research provides a theoretical explanation for such degradation. This contribution fills a critical gap by linking adversarial learning dynamics with real-time system constraints.

Distinction from Existing AML and Online IDS Models

Unlike traditional adversarial machine learning research, which focuses on bounded perturbation robustness under static datasets, the present framework introduces temporal robustness decay as a first-class property.

Classical AML robustness:

$$\min_{\|\delta\| \leq \epsilon} f_\theta(x + \delta)$$

Proposed framework:

$$\mathcal{R}(t) = 1 - R_{adv}(t)$$

capturing robustness evolution over streaming interactions.

Compared to online IDS models that emphasize incremental learning and concept drift handling, this framework explicitly integrates adversarial adaptation dynamics and latency constraints into a unified formal model.

Therefore, the novelty lies in:

- Temporal adversarial modeling
- Latency-aware robustness theory
- Formal definition of robustness decay rate
- Introduction of endurance-based metrics (TTE)

Latency as a Central Determinant of Robustness

The analysis points out that detection latency is a key factor affecting adversarial resilience in real-time IDS. Prior studies have proposed increasingly complex defense mechanisms such as ensemble learning and adversarial training to improve robustness [6,7]. While effective in offline evaluations, these approaches often introduce additional computational overhead that is incompatible with real-time detection requirements.

The latency–robustness trade-off identified in this study offers a conceptual framework for understanding this phenomenon. It suggests that robustness improvements achieved through computationally intensive defenses may be offset by increased detection delays, ultimately reducing operational effectiveness. This insight provides a plausible explanation for the gap between experimental results and real-world deployment outcomes reported in recent IDS studies [8,9].

Implications of Time-to-Evasion and Detection Stability

Introducing Time-to-Evasion (TTE) and detection stability as measures of robustness expands existing evaluation methods in adversarial IDS research. Traditional metrics focus on immediate performance but ignore how quickly an adversary can change tactics to avoid detection. TTE addresses this gap by focusing on endurance against adversarial adaptation over time.

Detection stability, as identified in the results, serves as an early indicator of robustness degradation. Fluctuations in detection outcomes under adversarial pressure may signal impending evasion before complete failure occurs.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

This observation complements existing work on uncertainty estimation and model confidence in ML systems [10], suggesting new directions for real-time monitoring and alert prioritization in security operations centers.

Interaction Between Concept Drift and Adversarial Behavior

Concept drift has long been recognized as a challenge for ML-based IDS, even in non-adversarial settings [11,12]. The present study extends this understanding by showing that concept drift amplifies the effectiveness of adversarial attacks in streaming environments. When normal network behavior evolves concurrently with adversarial perturbations, the IDS faces compounded uncertainty that accelerates robustness degradation.

Previous research in adversarial machine learning often assumes stable data distributions and has largely neglected this interaction. By including concept drift in the theoretical framework, this study offers a better depiction of real-world conditions, highlighting the limitations of static defense strategies.

Practical Implications for IDS Deployment

From a practical standpoint, the findings carry important implications for designing and deploying machine learning-based IDS. Security professionals should be cautious when interpreting claims of robustness based only on offline tests. Instead, evaluation protocols should incorporate time-related metrics and consider how IDS perform over time under adaptive adversarial stress.

The results indicate that adaptive and incremental learning approaches may be necessary to maintain robustness in real-time conditions. However, such strategies need careful design to avoid excessive delays and instability. Collaboration between humans and AI, where analysts can act based on stability indicators, may provide a balanced approach between automation and flexibility.

Limitations of the Study

While this research provides useful theoretical insights, it has limitations. The framework is conceptual and does not include empirical testing with real-world data. Additionally, the mathematical formulations are intentionally abstract to remain applicable across different IDS architectures. Future research should validate the proposed metrics and refine the models in specific deployment scenarios.

Contribution to the Field

Overall, this study contributes by redefining adversarial robustness in IDS as a dynamic, time-sensitive issue influenced by real-time challenges and adversarial adaptation. Linking adversarial machine learning theory with operational realities, it creates a foundation for future research and more realistic evaluations of IDS defenses.

Applicability Across IDS Architectures

The proposed framework is designed to be independent of any specific architecture and applies to a variety of machine learning-based intrusion detection systems. This includes signature-augmented classifiers, deep learning models, ensemble-based IDS, and hybrid systems used in network, host-based, and cloud contexts. Since the framework emphasizes temporal interaction, adversarial adaptation, and latency issues instead of model-specific behavior, its principles remain relevant regardless of feature representation or learning method. This flexibility broadens its relevance in different operational settings and supports its use as a reference for future IDS research.

CONCLUSION AND RECOMMENDATIONS

Conclusion

This study aimed to address a crucial gap in research on adversarial machine learning as it relates to intrusion detection systems (IDS). Specifically, it tackled the lack of a solid theoretical framework for understanding

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

adversarial robustness in real-time, machine learning-based IDS. While previous research has thoroughly examined adversarial attacks and defenses, much of it relies on static datasets and offline evaluations that do not capture the time-sensitive, adaptive, and latency-focused nature of real-world intrusion detection environments.

By viewing adversarial robustness as dependent on time and driven by interactions, this paper enhances our understanding of why many ML-based IDS defenses struggle in real situations, even when performing well in offline tests. The proposed theoretical framework brings together real-time constraints, adaptive adversarial behavior, and concept drift into a clear analytical model. It shifts the view of robustness from being a static model feature to a dynamic process influenced by ongoing adversarial interactions and system responses.

Key contributions of this study include outlining real-time IDS constraints, creating a streaming-aware adversarial threat model, and introducing latency-aware robustness metrics like Time-to-Evasion (TTE) and detection stability. These contributions build a conceptual base for assessing IDS performance beyond traditional accuracy metrics and help explain why existing adversarial defenses often fall short in practice.

Overall, this work addresses an important gap between adversarial machine learning theory and real-world cybersecurity applications, providing insights relevant to various IDS architectures, network situations, and threat environments.

Recommendations

Based on the theoretical findings and insights derived from this study, the following recommendations are proposed:

Recommendations for Researchers

1. **Adopt Time-Aware Evaluation Metrics:**

   Future IDS research should use temporal robustness metrics like Time-to-Evasion and detection stability, along with standard performance measures.

2. **Move Beyond Static Evaluation Settings:**

   Researchers should create evaluation protocols that reflect real-time streaming conditions, including adaptive adversarial behavior and extended interaction.

3. **Integrate Concept Drift into AML Studies:**

   Adversarial robustness research must consider concept drift, as its interplay with adversarial changes importantly impacts IDS performance.

4. **Develop Formal Mathematical Models:**

   Further research should aim to mathematically describe robustness decline rates and adversarial adaptation dynamics under real-time conditions.

Recommendations for Practitioners and Security Operations Centers

1. **Exercise Caution with Offline Robustness Claims:**

   IDS deployments should not depend only on offline evaluations of robustness when judging adversarial strength.

2. **Balance Robustness and Latency:**

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue I January 2026

Defense mechanisms should be chosen and adjusted while considering detection latency to ensure that improved robustness does not compromise operational efficiency.

3. **Monitor Detection Stability:**

   Security teams should track the consistency of detections over time to detect early signs of adversarial adaptation and potential evasion.

4. **Consider Hybrid Human–AI Approaches:**

   Involving human judgment in IDS decision-making can help address weaknesses exposed by adaptive adversaries in real-time situations.

Final Remarks

As machine learning continues to influence the future of intrusion detection, ensuring resilience against adversarial threats in real-time settings remains a major challenge. This study offers a foundational theoretical approach to tackling that challenge. By shifting the focus from static guarantees of robustness to dynamic, time-aware resilience, it sets the stage for more realistic, reliable, and deployable intrusion detection systems in adversarial environments.

REFERENCES

1. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305–316.
2. Scarfone, K., & Mell, P. (2007). Guide to intrusion detection and prevention systems (IDPS). *NIST Special Publication 800-94*.
3. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
4. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.
5. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*, 86, 147–167.
6. Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50.
7. Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. *International Conference on Platform Technology and Service (PlatCon)*.
8. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *International Conference on Machine Learning (ICML)*.
9. Apruzzese, G., Colajanni, M., Ferretti, L., Marchetti, M., & Guido, A. (2020). On the effectiveness of machine and deep learning for cyber security. *IEEE International Conference on Cyber Conflict*.
10. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. (2011). Adversarial machine learning. *ACM Workshop on Security and Artificial Intelligence*.
11. Lin, Z., Shi, Y., & Xue, Z. (2018). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. *arXiv preprint arXiv:1809.02077*.
12. Rigaki, M., & Garcia, S. (2018). Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. *IEEE Security and Privacy Workshops*.
13. Cheng, L., Liu, F., & Yao, D. (2019). Enterprise data breach: Causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5).
14. Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.
15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

16. Apruzzese, G., Marchetti, M., & Colajanni, M. (2017). Deep reinforcement adversarial learning against botnet evasion attacks. *IEEE Transactions on Network and Service Management*, 14(4), 864–876.
17. Papernot, N., McDaniel, P., Goodfellow, I., et al. (2016). Practical black-box attacks against machine learning. *ACM Asia Conference on Computer and Communications Security*.
18. Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. *International Conference on Machine Learning (ICML)*.
19. Corona, I., Giacinto, G., & Roli, F. (2013). Adversarial attacks against intrusion detection systems. *ACM Transactions on Information and System Security*, 16(2).
20. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4).
21. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis. *Information Fusion*, 37, 132–156.
22. Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. *Computer Science Department, Trinity College Dublin*.
23. Jordaney, R., Sharad, K., Dash, S. K., et al. (2017). Transcend: Detecting concept drift in malware classification models. *USENIX Security Symposium*.
24. Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems (NeurIPS)*.
25. Xu, W., Evans, D., & Qi, Y. (2016). Feature squeezing: Detecting adversarial examples in deep neural networks. *NDSS Symposium*.