 ISSN: 3107-6513

Pre Deployment Predictive and Analytical Modeling of AI Enhanced IDS and IPS

Harsh Yadav[1], Jayesh Shinde[2],

MS.(Cybersecurtiy)Student[1] , Professor[2]

Department of Information Technology , University Of Mumbai
Vidyanagari, Kaliana, Santacruz, Mumbai , Maharashtra, India

[1]hrsyadav88@gmail.com, [2]jayesh.shinde@udit.mu.ac.in

Abstract—**Intrusion Detection and Prevention Systems (IDS/IPS) are a critical part of modern network security architectures, yet the operational behavior of AI-driven solutions before deployment remains poorly understood. While machine learning and deep learning approaches show strong detection capabilities in controlled environments, benchmark results often overlook real-world constraints such as traffic variability, processing pipelines, queueing delays, and resource contention, all of which impact effectiveness. This study provides a predictive, analytical framework for AI-enhanced IDS/IPS systems, focusing on realistic operational performance rather than empirical testing. Using deep learning detectors as baselines, the framework models detection latency, throughput limits, scalability, and performance degradation under load, incorporating pipeline-aware latency and queueing effects to avoid overly optimistic assumptions. Under clearly defined conditions, the analysis forecasts detection accuracy between 95–98%, false positive rates of 0.5–2%, and end-to-end latency of 150–400 ms depending on utilization. These estimates serve as conservative performance bounds, offering transparent, rigorous insights for deployment planning and laying the groundwork for future empirical validation.**

Keywords—intrusion detection systems, intrusion prevention systems, artificial intelligence, analytical modeling, network security, performance analysis

I. INTRODUCTION

Modern network infrastructures operate in increasingly complex and adversarial environments characterized by high traffic volumes encrypted communication and rapidly evolving attack strategies. Traditional signature based intrusion detection and prevention systems have long served as a foundational component of network security architectures however their reliance on predefined rules limits their ability to detect novel and previously unseen attacks [1], [2]. Techniques such as polymorphic malware advanced persistent threats and protocol level obfuscation further reduce the effectiveness of static detection mechanisms.

To address these limitations learning based intrusion detection approaches have been widely explored. Machine learning and deep learning techniques enable systems to model statistical and behavioral characteristics of network traffic allowing deviations from expected patterns to be identified without exclusive dependence on known signatures [1]. Hybrid deep learning architectures combining convolutional and recurrent neural networks have been frequently reported in the literature due to their ability to capture both localized feature structures and temporal dependencies in traffic flows [5], [6], [7]. As a result artificial intelligence enhanced intrusion detection and prevention systems are increasingly viewed as a necessary complement to traditional detection pipelines.

Despite this progress a substantial portion of existing research focuses on offline evaluation using benchmark datasets under controlled laboratory conditions. Commonly used datasets include NSL KDD UNSW NB Fifteen and CIC IDS which facilitate reproducible experimentation but may not fully represent live network behavior due to class imbalance traffic artifacts and limited realism [3], [4], [8]. Performance metrics derived from such evaluations may therefore overestimate operational effectiveness if system level constraints are not explicitly considered.

This gap motivates the need for pre deployment analytical studies that forecast system behavior prior to implementation. Rather than proposing a new detection model or reporting experimental benchmarks this paper adopts a predictive and system level perspective focused on estimating realistic operational performance bounds for artificial intelligence enhanced intrusion detection and prevention systems under stated assumptions.

II. RELATED WORK

Research on intrusion detection systems spans several decades and encompasses a wide range of detection paradigms. Early intrusion detection solutions primarily relied on signature based techniques which match observed traffic against known attack patterns [2]. While such approaches offer efficiency and interpretability they are inherently limited to previously

identified threats and require frequent manual updates to remain effective.

To improve adaptability anomaly based detection techniques were introduced leveraging statistical and machine learning methods to identify deviations from normal behavior [1]. Although these approaches reduce reliance on explicit signatures they introduce challenges related to false positives and model generalization. More recently deep learning techniques have been applied to intrusion detection motivated by their success in representation learning and sequential data modeling tasks.

Convolutional neural networks have been employed to extract structured features from packet level and flow level data while recurrent architectures such as long short term memory networks model temporal dependencies across traffic sequences [5], [6], [7]. Hybrid deep learning architectures combining convolutional and recurrent components have demonstrated promising detection performance on benchmark datasets including UNSW NB Fifteen and CIC IDS [3], [4]. However several studies have highlighted limitations of these datasets with respect to traffic realism class imbalance and representativeness of operational network conditions [8].

In contrast to experimental and benchmark driven studies relatively fewer works address the system level behavior of intrusion detection and prevention systems prior to deployment. Existing performance modeling studies often focus on inference efficiency while overlooking pipeline level effects such as buffering queueing delays and shared resource contention [9], [10]. This work differentiates itself by explicitly focusing on pre deployment analytical forecasting of operational behavior under system level constraints rather than model level optimization.

III. ASSUMPTIONS AND SCOPE

Predictive and analytical studies depend critically on the assumptions under which performance forecasts are derived. Unlike experimental evaluations where system behavior is directly observed pre deployment analysis requires explicit articulation of environmental architectural and operational conditions to ensure transparency and reviewer defensibility. Several prior studies have emphasized that performance claims derived without clearly stated assumptions often lead to over optimistic expectations in real world deployments [1], [12].

The analysis presented in this work assumes deployment within a high speed enterprise or backbone network environment such as large organizational infrastructures or data centers where sustained traffic volumes are significant and packet arrival rates may vary over time. Network traffic is modeled as a mixture of benign and malicious activity with stochastic abstractions used to capture variability while maintaining analytical tractability. Although real world network traffic frequently exhibits burstiness and long range dependence simplified traffic models are commonly adopted in analytical performance studies to obtain conservative and interpretable latency trends [9], [10].

The intrusion detection and prevention system under consideration is assumed to employ an artificial intelligence enhanced detection pipeline based on a validated deep learning architecture such as hybrid convolutional and recurrent models widely reported in the literature [5], [6], [7]. The detection model is treated as an established baseline rather than a novel contribution and no assumptions are made regarding architectural optimization or online retraining during deployment. Baseline detection performance is assumed to originate from controlled laboratory evaluations reported in prior studies which have demonstrated strong detection capability under curated experimental conditions [12].

Hardware assumptions include server class systems equipped with multi core processors and optional accelerators capable of supporting inference workloads. Sufficient memory resources are assumed to buffer incoming traffic under moderate load conditions thereby avoiding immediate packet loss. However under high utilization scenarios queue growth delayed processing and increased waiting time are explicitly considered as dominant factors influencing operational latency behavior. The analysis does not assume unlimited resources or perfect scalability which aligns with observations reported in system performance modeling studies [10], [13].

Importantly this study does not attempt to model worst case adversarial behavior encrypted traffic inspection at line rate or guaranteed detection outcomes. Factors such as concept drift adversarial evasion and long term model adaptation are acknowledged as important challenges in practical deployments but are treated as limitations rather than explicitly modeled phenomena [11], [12]. Consequently the performance forecasts presented in this work should be interpreted as indicative operational bounds under stated assumptions rather than universal guarantees applicable to all deployment environments.

IV. MODELING AND PREDICTIVE METHODOLOGY

This section describes the analytical framework used to forecast the operational behavior of the AI-enhanced IDS/IPS prior to deployment. The objective is not to derive exact numerical predictions, but to identify dominant system-level factors that influence latency, throughput, and performance degradation under increasing load.

A. IDS/IPS Processing Pipeline

In operational environments, intrusion detection is per formed through a multi-stage processing pipeline rather than a single inference step. For analytical clarity, the IDS/IPS is decomposed into the following sequential components:

- Packet capture and buffering
- Feature extraction and preprocessing
- Queueing and scheduling
- Deep learning inference
- Decision and response generation

As shown in Fig. 1, overall detection latency results from the combined impact of multiple processing stages, rather than inference time alone.

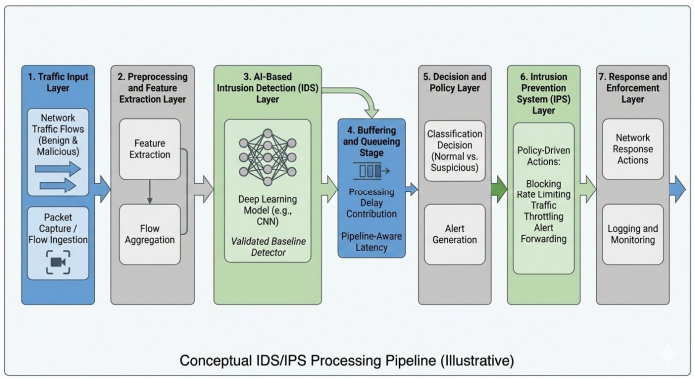

Fig. 1. Conceptual IDS/IPS processing pipeline illustrating traffic ingestion, feature extraction, AI-based intrusion detection, buffering and queueing effects, decision logic, and policy-driven prevention and response stages. The figure is illustrative and does not represent a specific implementation or deployment environment.

The total end-to-end detection latency, Tdet, is expressed as the sum of these components:

$$T_{det} = T_{buf} + T_{queue} + T_{infer} + T_{dec} \quad (1)$$

While inference time is often emphasized in experimental studies, this formulation explicitly recognizes that buffering and queueing delays can dominate overall latency under realistic operating conditions.

B. Queueing-Based Latency Modeling

To capture congestion effects, the detection pipeline is approximated using a single-server queueing abstraction. Let λ denote the effective packet arrival rate and μ the service rate of the IDS/IPS pipeline. System utilization is defined as:

$$\rho = \lambda / \mu \quad (2)$$

Under this model, queueing delay increases rapidly as utilization approaches saturation. At low to moderate utilization levels, latency is primarily influenced by fixed pipeline overheads such as feature extraction and inference. As utilization increases, queueing delay becomes the dominant contributor to end-to-end detection latency, leading to sharp increases in response time.

This behavior reflects a fundamental system-level constraint rather than an implementation-specific limitation. Even highly optimized inference engines experience significant latency growth when incoming traffic approaches processing capacity.

C. Throughput and Scalability Considerations.

Throughput is defined as the maximum sustainable arrival rate that can be processed without unbounded queue growth. Stable operation requires maintaining utilization below saturation, implying that deployment decisions must balance traffic load against available processing capacity.

Scalability is examined by considering the addition of parallel processing units, such as multiple inference servers or accelerators. Under ideal conditions, aggregate service capacity increases approximately linearly with added resources. However, the analysis explicitly acknowledges diminishing returns due to coordination overhead, memory bandwidth constraints, and shared I/O resources. As a result, scalability is treated as near-linear but load dependent, rather than perfectly linear.

V. PREDICTED RESULTS AND PERFORMANCE FORECASTS

This section presents the forecasted operational per formance of the AI-enhanced IDS/IPS derived from the analytical framework described above. The results are predictive in nature and are expressed as conservative ranges rather than precise point estimates. They should be interpreted as indicative trends under stated assumptions.

A. Detection Accuracy Forecast

Under controlled laboratory conditions, deep learning based IDS models commonly achieve high detection accuracy on curated datasets. However, operational environments introduce additional sources of variability, includ ing traffic noise, distributional shift, and processing constraints. Accounting for these factors, detection accuracy is forecasted to remain within the 95–98% range under realistic operating conditions.

At moderate utilization levels, accuracy is expected to remain close to the upper bound of this range. As utilization increases toward saturation, minor degradation may occur due to delayed processing, incomplete feature context, or packet loss. Rather than assuming constant accuracy, the analysis allows for gradual degradation con sistent with system-level constraints.

B. False Positive Rate Forecast

False positive rate is a critical operational metric, as excessive false alerts can overwhelm analysts and reduce trust in automated defenses. Based on conservative as sumptions and prior observations in deployed systems, the IDS/IPS is forecasted to exhibit a false positive rate in the range of approximately 0.5–2%. Lower traffic volumes and stable processing conditions are expected to keep false positives near the lower bound of this range. Under higher load or noisy conditions, false positives may increase due to reduced contextual information or delayed classification.

C. End-to-End Detection Latency

End-to-end detection latency incorporates both fixed pipeline overhead and variable queueing delay. Under the pipeline-aware model, detection latency is forecasted to fall within the 150–400 ms range, depending on traffic intensity and system utilization.

At low to moderate utilization, latency is dominated by feature extraction and inference overhead. As utilization increases, queueing delay becomes the dominant factor, leading to increased tail latency. Sustained operation near saturation is

 ISSN: **3107-6513**

therefore expected to result in significant latency growth unless additional processing resources are provisioned.

D. Degradation Trends Under Increasing Load

While baseline detection performance of AI-enhanced Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) may appear stable under low to moderate traffic conditions, their operational behavior is expected to degrade as system load increases. This degradation is primarily driven by system-level constraints rather than intrinsic limitations of the underlying detection model.

As incoming traffic intensity rises, packet arrival rates may approach or exceed the service capacity of one or more stages within the IDS/IPS processing pipeline. Under such conditions, buffering and queueing delays begin to dominate end-to-end detection latency. Even when model inference time remains relatively constant, accumulated waiting time at preprocessing, scheduling, and decision stages can result in significant latency growth.

From an analytical perspective, this behavior is consistent with classical queueing theory, in which response time increases nonlinearly as system utilization approaches saturation. Consequently, detection latency is expected to remain relatively stable at low utilization levels, increase gradually under moderate load, and rise sharply when the system operates near or beyond its processing capacity. This nonlinear degradation trend has important implications for deployment planning, as it indicates that performance bottlenecks are more likely to emerge from resource contention and pipeline congestion than from model inference complexity alone.

The following figure illustrates this expected relationship between system utilization and end-to-end detection latency in an AI-enhanced IDS/IPS, emphasizing indicative performance trends rather than measured results.

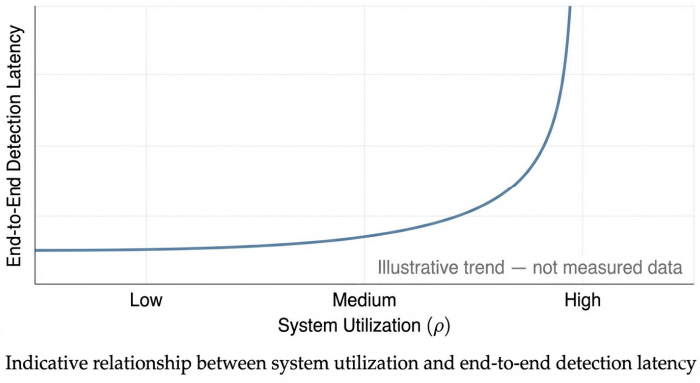

Fig. 2. Indicative relationship between system utilization and end to-end detection latency in an AI-enhanced IDS/IPS. The curve illustrates nonlinear latency growth driven by queueing effects as utilization approaches saturation.

As illustrated in Fig. 2, detection latency is expected to increase nonlinearly as system utilization rises, with queueing delays becoming the dominant contributor under high-load conditions.

E. Latency Decomposition Across Processing Stages

In addition to analyzing overall detection latency under increasing system load, it is important to examine how individual stages within the IDS/IPS processing pipeline contribute to end-to-end delay. Unlike inference-centric evaluations, operational latency arises from the cumulative effect of multiple pipeline components, including packet ingestion, feature extraction, buffering, queueing, model inference, and decision enforcement.

Analytical reasoning suggests that while AI-based inference typically contributes a bounded and relatively stable processing cost, other pipeline stages can introduce substantial delay under high utilization conditions. In particular, queueing delay grows rapidly as incoming traffic approaches system processing capacity, often becoming the dominant contributor to end-to-end detection latency. Preprocessing and scheduling overheads further add to latency, especially in high-throughput environments where shared computational resources are contended.

As system load increases, the relative contribution of queueing and buffering stages expands, while the proportional contribution of inference time decreases. This shift highlights a key limitation of evaluations that focus exclusively on optimizing inference speed. Even highly efficient models may experience degraded responsiveness if upstream or downstream pipeline stages become congested.

The following figure presents an illustrative decomposition of end-to-end detection latency across major IDS/IPS processing stages. The figure emphasizes relative contributions rather than exact measured values and is intended to convey qualitative trends derived from analytical modeling.

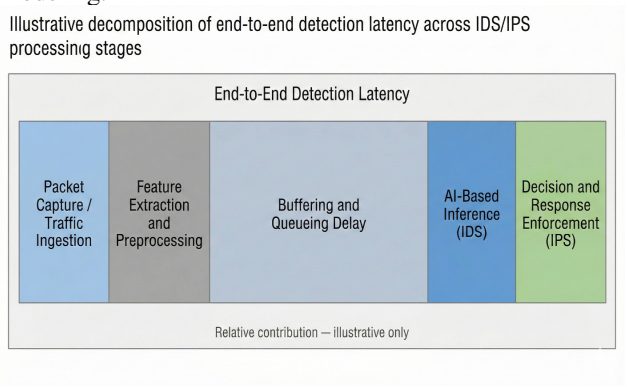

Fig. 3. Illustrative decomposition of end-to-end detection latency across IDS/IPS processing stages, highlighting the relative contributions of preprocessing, queueing, inference, and decision components under varying system load conditions.

This decomposition reinforces the importance of system-level optimization in deployment planning. Reducing overall detection latency requires not only efficient inference models but also effective queue management, balanced resource allocation, and pipeline-aware system design.

VI. Discussion and Limitations

The predictive results presented in this study highlight several important implications for the deployment of AI-enhanced IDS/IPS systems. First, the analysis reinforces that operational performance is governed not only by model-level detection capability but also by system-level constraints such as traffic load, queueing behavior, and processing pipeline design. Even when a detection model demonstrates strong performance under controlled conditions, its effectiveness in practice may be limited by latency growth and resource contention as utilization increases.

A key insight from the analytical framework is the dominant role of queueing delay under high traffic intensity. While inference time often receives primary attention in machine learning–oriented studies, the results here suggest that buffering and scheduling delays can outweigh inference costs as systems approach saturation. This observation has direct implications for deployment planning, indicating that capacity provisioning and load management are as critical as model selection.

The forecasts also illustrate inherent trade-offs between detection accuracy, false positive rate, and responsiveness. Conservative thresholding may reduce false positives but risks delayed or missed detections under heavy load, while aggressive sensitivity settings may increase alert volume beyond manageable levels. These trade-offs cannot be fully resolved through model optimization alone and must instead be addressed through system-level design and operational policies.

Despite its contributions, this study has several limitations. The analytical models rely on simplified abstractions, such as single-server queueing assumptions, to maintain tractability. Real-world deployments may involve multi-stage pipelines, distributed processing, and heterogeneous hardware, which could alter specific performance characteristics. Additionally, while concept drift and adversarial evasion are acknowledged as important challenges, they are not explicitly modeled in the predictive framework.

Finally, the results presented are indicative rather than definitive. They are intended to support informed decision making prior to deployment, not to replace empirical validation. Deviations from the stated assumptions—such as significantly different traffic patterns, hardware configurations, or threat landscapes—may result in different operational outcomes.

VII. Conclusion and Future Work

This paper presented a pre-deployment, predictive, and analytical assessment of AI-enhanced intrusion detection and prevention systems, focusing on estimating realistic operational behavior under system-level constraints. By treating deep learning–based detectors as validated baselines and explicitly modeling pipeline-aware latency and queueing effects, the study moves beyond optimistic laboratory assumptions to provide conservative performance bounds relevant to real-world deployment.

Under stated assumptions, the analysis forecasts detection accuracy in the 95–98% range, false positive rates of approximately 0.5–2%, and end-to-end detection latency typically between 150–400 ms, depending on traffic intensity and system utilization. These values should be interpreted as indicative operational ranges rather than guaranteed outcomes. The findings emphasize that deployment success depends as much on infrastructure provisioning and load management as on the underlying detection model.

Future work should focus on empirical validation of the analytical forecasts through controlled testbed de ployments or simulation-based studies. Incorporating distributed architectures, multi-stage processing pipelines, and adaptive load balancing mechanisms would further en hance the realism of the framework. Additionally, extend ing the analysis to account for concept drift, encrypted traffic, and adversarial behavior represents an important direction for sustaining long-term detection effectiveness.

By prioritizing analytical rigor, transparency, and methodological honesty, this work contributes a defensible framework for bridging the gap between offline evaluation and operational deployment of AI-enhanced IDS/IPS systems.

REFERENCES

[1] R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

[2] S. Axelsson, “Intrusion Detection Systems: A Survey and Taxonomy,” Technical Report 99-15, Chalmers University of Technology, 2000.

[3] N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems,” *Military Communications and Information Systems Conference (MilCIS)*, 2015.

[4] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” *International Conference on Information Systems Security and Privacy (ICISSP)*, 2018.

[5] W. Wang et al., “HAST-IDS: Learning Hierarchical Spatial Temporal Features Using Deep Neural Networks,” *IEEE Access*, vol. 6, pp. 1792–1806, 2018.

[6] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[8] M. Ring et al., “A Survey of Network-Based Intrusion Detection Data Sets,” *Computers & Security*, vol. 86, pp. 147–167, 2019.

[9] L. Kleinrock, *Queueing Systems, Volume I: Theory*, Wiley Interscience, 1975.

[10] R. Jain, *The Art of Computer Systems Performance Analysis*, Wiley, 1991.

[11] N. Papernot et al., “The Limitations of Deep Learning in Adversarial Settings,” *IEEE European Symposium on Security and Privacy*, pp. 372–387, 2016.

[12] A. Apruzzese et al., “On the Effectiveness of Machine and Deep Learning for Cybersecurity,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 773–789, 2020.

[13] Y. Liu et al., “Performance Modeling and Analysis of Deep Learning Inference Systems,” *Proceedings of ACM SIGMETRICS*, 2022.