A Unified Framework for Trustworthy Machine Learning for Intrusion Detection Systems

Hussein Fawaz[1,2,*], Omran Ayoub[1], Silvia Giordano[1]

[1]University of Applied Sciences and Arts of Southern Switzerland, Switzerland, [2]Università della Svizzera italiana, Switzerland
Corresponding author: hussein.fawaz@usi.ch

***Abstract*—The rapid growth of connected devices and Internet of Things (IoT) systems has significantly expanded the attack surface of modern networks, increasing the need for effective cybersecurity mechanisms. Intrusion Detection Systems (IDS) play a critical role by monitoring network traffic and identifying malicious activities. In recent years, machine learning (ML) techniques have been widely adopted in IDS to learn complex patterns of normal and malicious behavior from large-scale data. Despite strong performance on benchmark datasets, ML-based IDS often struggle to operate reliably in real-world environments. The dynamic and adversarial nature of cybersecurity introduces distribution shifts, evolving attack patterns, and concept drift, which can degrade model performance. In addition, many models exhibit overconfident predictions and limited transparency, hindering their trustworthiness in safety-critical applications.**

This research aims to develop a unified framework for trustworthy ML in intrusion detection. The goal is to design systems that are accurate, reliable, robust, adaptive, interpretable, and efficient in dynamic environments. In particular, this work investigates how reliability-related characteristics, such as predictive confidence, robustness to distribution shifts, and explanation validity, can be jointly modeled and evaluated, supporting the development of IDS that provide dependable and transparent decision-making.

***Index Terms*—Trustworthy AI, Cybersecurity, Explainable AI, Intrusion Detection**

I. Introduction

In an era of ubiquitous connectivity and rapidly expanding digital infrastructure, defending networks against malicious activity has become a paramount challenge [1]. Intrusion Detection Systems (IDS) are designed to monitor and analyze network traffic to detect suspicious or malicious activities. However, many traditional IDS rely on rule-based methods and signature matching, which have become increasingly inadequate as cyber threats grow in sophistication and scale, particularly against adaptive and previously unseen attacks that do not match known signatures [2].

To enable stronger and more adaptive defense mechanisms, the focus has shifted toward data-driven approaches based on machine learning (ML) [3]. These models, ranging from classical classifiers to deep learning approaches, can learn complex patterns of normal and malicious behavior directly from network data [4]. In this context, ML has been applied in both supervised and unsupervised learning settings, where classification models are used to distinguish between known attack types, while unsupervised methods such as autoencoders are employed to detect anomalies and previously unseen behaviors [5].

A fundamental peculiarity of IDS is its operation within a highly dynamic environment. Network traffic evolves continuously due to changes in user behavior, system configurations, and the emergence of new threats. Under such conditions, models trained on historical data often experience degraded performance when exposed to data that differs from those exploited during the training phase, a phenomenon known as distribution shift [6]. A particular example of distribution shift is concept drift, where the underlying data-generating process changes over time, requiring models to continuously adapt in order to remain effective [7]. In addition, IDS must handle the presence of novel and evolving attack patterns, for which no labeled training data may be available [8].

Another peculiarity of IDS is its safety-critical nature. Incorrect predictions may lead to severe consequences, such as undetected attacks or unnecessary system disruptions. In such settings, models must not only be accurate but also provide reliable measures of confidence [10]. In particular, calibration is a fundamental property of a reliable model, as it refers to the alignment between predicted probabilities and the true likelihood of correctness [9]. For example, predictions made with 80% confidence should be correct approximately 80% of the time. However, ML models are often prone to overconfidence, producing highly certain predictions even when incorrect or when encountering unfamiliar inputs [11]. This misalignment between predicted confidence and true correctness is commonly referred to as miscalibration. Furthermore, IDS also necessitates transparency, as analysts must be able to understand and validate the reasoning behind their predictions [12]. Ensuring that model decisions are based on meaningful and relevant features is therefore essential for establishing trust in IDS. In addition to these considerations, practical deployment of IDS requires models to be computationally efficient and scalable, particularly in resource-constrained environments such as edge and IoT networks.

Taken together, these challenges highlight the need for trustworthy ML approaches for IDS. In this context, trustworthiness encompasses multiple complementary dimensions, including reliability, robustness, and transparency [13]. While recent research has explored individual aspects such as uncer-

ISBN 978-3-903176-82-9

tainty estimation, calibration, explainability, and robustness, these components are typically addressed in isolation. As a result, there is limited understanding of how they interact, particularly in dynamic environments characterized by distribution shifts, concept drift, and emerging attack patterns. This gap motivates the need for a holistic approach to trustworthy ML for IDS, where multiple properties are jointly modeled, evaluated, and analyzed within a unified framework.

II. Research Objectives

Based on these challenges, the central research problem is therefore:

> *How can we identify, quantify, and evaluate the key dimensions of trustworthiness in ML-based IDS, and understand the interplay between them?*

Building on the identified challenges and requirements, the objectives of my research are as follows:

O1: Identify, quantify, and analyze the key dimensions of trustworthiness in IDS, including reliability, robustness, and transparency, and investigate how their interactions influence overall system effectiveness.

O2: Develop and evaluate ML approaches that maintain reliable performance under distribution shifts, concept drift, and novel attack scenarios, enabling robust generalization in dynamic and evolving environments.

O3: Design and assess mechanisms for trustworthy decision-making, focusing on reliable confidence estimation and explainability, to ensure that predictions are well-calibrated and based on meaningful and interpretable features.

III. Dimensions of Trust for IDS

Building upon the challenges identified in the previous section, this section outlines the key dimensions of trustworthy ML for IDS. While the dimensions of trust in ML systems can be broadly applicable across ML systems in diverse domains, our work focuses on those that are particularly critical for IDS, where dynamic environments, adversarial behavior, and safety-critical decision-making impose unique requirements. In this context, these dimensions are not independent, but interact and jointly influence the overall reliability of the system.

Reliability: The ability of an ML model to produce consistent and dependable predictions with well-calibrated confidence estimates that reflect the true likelihood of correctness. It also involves quantifying and handling different sources of uncertainty, including epistemic uncertainty arising from limited knowledge and aleatoric uncertainty inherent in the data [11], [13].

Robustness: The ability to maintain stable performance when exposed to distribution shifts, noise, or adversarial perturbations. Robust models remain effective even when operating under conditions that differ from the training data, ensuring dependable behavior in dynamic environments [14].

Adaptability: The capability of a model to adjust to evolving data distributions and changing attack patterns over time. Adaptive systems maintain performance without requiring complete retraining, enabling sustained operation in continuously changing environments [15].

Generalizability: The ability to perform well on unseen data, environments, and previously unknown attack types. Generalizable models extend learned knowledge beyond the training distribution, which is essential for detecting zero-day and emerging threats [16].

Interpretability: The ability to provide explanations that make model predictions understandable to human analysts. It facilitates insight into model reasoning and supports human oversight, which is critical for building trust in security applications [17], [18].

Verifiability: The capability to assess whether model decisions are based on meaningful and relevant features. It enables validation against domain knowledge, ensuring that decisions are grounded in legitimate patterns rather than spurious correlations [19].

Security: The ability of the model to resist adversarial manipulation and evasion attempts. A secure model maintains its effectiveness under targeted adversarial conditions [20].

Efficiency and Sustainability: The ability to operate with minimal computational and energy resources while maintaining performance. Such models support real-time detection and scalable deployment, particularly in resource-constrained environments [21], [22].

IV. Proposed Research Methodology

This research adopts a modular and model-agnostic methodology to develop a unified framework for trustworthy ML in IDS. The approach integrates multiple complementary components, including explainability, uncertainty quantification, calibration, and concept-based reasoning, to jointly model and evaluate different dimensions of trustworthiness.

Explanation-based representations are leveraged to capture the underlying reasoning of model predictions. Feature attribution methods such as SHAP are used to transform raw inputs into explanation vectors, enabling the analysis of whether predictions are based on meaningful and consistent patterns. These representations are further utilized to support verification and to detect deviations from expected behavior.

Uncertainty-aware learning is investigated to improve the reliability of model predictions. Both epistemic and aleatoric uncertainty are considered through techniques such as ensemble modeling and uncertainty-aware loss functions. These uncertainty estimates are used to quantify predictive confidence and also support the detection of unfamiliar or out-of-distribution instances.

Calibration-aware optimization is applied to ensure that predicted probabilities accurately reflect true likelihoods. This includes both intrinsic approaches, where calibration is integrated into the training objective, and post-hoc calibration techniques. The impact of calibration is evaluated not only in terms of confidence reliability but also in relation to computational efficiency and resource constraints.

Representation learning techniques such as autoencoders and alternative density estimation models are employed to

model normal and known attack behaviors. By analyzing reconstruction error or likelihood estimates in combination with uncertainty and explanation signals, the framework enables the identification of novel or zero-day attacks.

In addition, concept-based learning is explored to enhance interpretability and verifiability. Intermediate human-interpretable concepts are incorporated into the learning process, allowing model decisions to be grounded in semantically meaningful attributes and facilitating validation by human analysts. This also enables a more transparent understanding of how high-level concepts influence the model's predictions and behavior.

The proposed framework is evaluated under dynamic conditions, including distribution shifts and concept drift. Continual learning strategies are investigated to enable models to adapt to evolving data distributions while mitigating catastrophic forgetting, particularly under limited or unavailable labeled data, reflecting realistic intrusion detection scenarios.

The evaluation relies on publicly available datasets that provide labeled network traffic with diverse attack types. To simulate realistic deployment conditions, the data will be partitioned into known and unknown attack scenarios, where models are trained on a subset of attack types and evaluated on previously unseen attacks. In addition, temporal or distribution-based splits will be considered to emulate concept drift and evolving network conditions.

Given the resource-constrained nature of IoT environments, the computational efficiency of the proposed methods is also considered. In particular, trade-offs between model complexity, inference latency, and detection performance are analyzed to ensure feasible deployment in practical network settings.

Through this integrated methodology, the research investigates the interaction between reliability, robustness, and interpretability, supporting the development of trustworthy IDS for real-world environments.

V. Conclusion

This work presented a research direction toward developing trustworthy machine learning approaches for intrusion detection systems, with a focus on reliability under dynamic and evolving conditions. The initial phase of this research has primarily addressed reliability in static and controlled settings, providing insights into confidence estimation, uncertainty modeling, and explainability. Building on these foundations, the next phase will focus on extending these capabilities to dynamic environments, where distribution shifts, concept drift, and emerging attack patterns are inherent. Addressing this is expected to contribute toward the development of IDS that maintain dependable performance and trustworthy decision-making in real-world deployments.

Acknowledgment

This work has partially been supported by the Swiss Government Excellence Scholarship (ESKAS) No. 2024.0474 and by Innosuisse, the Swiss Innovation Agency, through the innovation project SUSTAINET (No. 119.588 INT-ICT), carried out within the EUREKA Cluster CELTIC-NEXT under the project SUSTAINET-Advance.

References

[1] L. Diana et al., "Overview on intrusion detection systems for computers networking security," *Computers*, vol. 14, no. 3, p. 87, 2025.

[2] R. Chinnasamy et al., "Deep learning-driven methods for network-based intrusion detection systems: A systematic review," *ICT Express*, vol. 11, no. 1, pp. 181–215, 2025.

[3] Z. Ahmad et al., "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, 2021.

[4] A. S. Dina and D. Manivannan, "Intrusion detection based on machine learning techniques in computer networks," *Internet of Things*, vol. 16, p. 100462, 2021.

[5] Z. Yang et al., "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Computers & Security*, vol. 116, p. 102675, 2022.

[6] M. Dragoi et al., "AnoShift: A distribution shift benchmark for unsupervised anomaly detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32854–32867, 2022.

[7] L. Yang et al., "CADE: Detecting and explaining concept drift samples for security applications," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2327–2344.

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

[9] B. Pal et al., "Calibrated uncertainty estimation for trustworthy deep IoT attack detection," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[10] Fawaz, H, et al. "Towards Better-Calibrated ML Models for Reliable Network Intrusion Detection via Calibration-Aware SHAP-Based Feature Selection." 21th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2025.

[11] J. Talpini et al., "Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification," *Journal of Reliable Intelligent Environments*, vol. 10, no. 4, pp. 501–520, 2024.

[12] I. H. Sarker et al., "Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects," *ICT Express*, vol. 10, no. 4, pp. 935–958, 2024.

[13] A. Nascita et al., "Improving performance, reliability, and feasibility in multimodal multitask traffic classification with XAI," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1267–1289, 2023.

[14] B. Chander et al., "Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–49, 2025.

[15] M. Al Rawajbeh et al., "Trustworthy adaptive AI for real-time intrusion detection in industrial IoT security," *IoT*, vol. 6, no. 3, p. 53, 2025.

[16] Z. Zhang et al., "Trustworthy generative few-shot learning-based intrusion detection method in Internet of Things," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 1, pp. 1992–2002, 2024.

[17] A. Nascita et al., "A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 5, pp. 3165–3198, 2024.

[18] M. T. Islam et al., "Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 11, pp. 5337–5360, 2024.

[19] W. Wei and L. Liu, "Trustworthy distributed AI systems: Robustness, privacy, and governance," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–42, 2025.

[20] S. Ennaji et al., "Adversarial challenges in network intrusion detection systems: Research insights and future prospects," *IEEE Access*, 2025.

[21] A. Jaddoa et al., "Toward scalable and sustainable detection systems: A behavioural taxonomy and utility-based framework for security detection in IoT and IIoT," *IoT*, vol. 6, no. 4, p. 62, 2025.

[22] Fawaz, H, et al. "Energy Cost of Enhancing Reliability of ML Models for Edge IoT Security." (2026). 2026 21st Wireless On-Demand Network Systems and Services Conference (WONS). IEEE 2026.