Explanation-Based Runtime Verification for Trustworthy ML-driven Optical Networks

Omran Ayoub [*] Carlos Natalino [†], Ali Al Housseini [*], Felix Foschum[‡],
Philipp Morger[‡], Tiziano Leidi [*], David Hock [§], Paolo Monti [†]
[*] University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland
[†] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
[‡] SKOOR GmbH, Switzerland, [§] Infosim GmbH & Co. KG, Germany

***Abstract*—Machine learning (ML) models are increasingly integrated into optical network automation frameworks to support tasks such as failure management, performance monitoring and resource allocation. In these environments, ML-driven predictions may be directly coupled with control-plane actions where incorrect decisions can immediately impact service quality, resource efficiency, and network stability. As automation levels increase, ensuring the reliability of individual decisions at deployment time becomes a critical requirement. Explainable artificial intelligence (XAI) techniques have emerged to improve transparency by highlighting the factors influencing ML predictions. In addition to identifying influential features, they provide insights into the underlying reasoning process of the model, revealing how different input variables contribute to the final outcome and how feature interactions shape the decision boundary. In this work, we introduce explanation-based runtime verification, an approach that exploits model explanations to assess the soundness of individual ML decisions before they are executed in the network control loop. The proposed approach evaluates explanation coherence and physics grounding consistency at runtime, enabling the system to defer or reject decisions flagged as uncertain. We demonstrate the effectiveness of our approach on a representative use case of lightpath quality of transmission classification. Experimental results show that explanation-based verification can intercept a significant fraction of erroneous decisions while preserving high automation rate.**

I. Introduction

Machine learning (ML) models are increasingly deployed in optical networks to support operational decisions such as configuration selection, performance estimation, anomaly detection, and resource allocation [1]. As these models become embedded in control and orchestration systems, their outputs are no longer merely advisory but they directly influence network actions. In such environments, incorrect ML decisions can lead to service degradation, inefficient resource utilization, or instability in automated workflows [2].

Ensuring high aggregate predictive accuracy during offline validation is not sufficient to guarantee safe deployment as such models may still produce isolated erroneous decisions under specific network conditions. From an operator's perspective, the critical question is therefore not only *how accurate is the model overall?*, but rather *can a specific decision be trusted at runtime before it is enacted?* Avoiding wrong automated decisions, particularly false-positive ones that may trigger unsafe configurations, is therefore essential in operational optical networks.

Explainable artificial intelligence (XAI) techniques have been introduced to improve transparency by revealing which input features influence a model's prediction, with the ultimate goal of enabling model behavior verification and enhancing trust [3]. While such explanations enhance interpretability and facilitate offline analysis, they do not inherently provide mechanisms to assess the reliability of individual decisions at inference time. An explanation can describe a decision, but it does not verify whether the reasoning is coherent, plausible, or sufficiently supported to justify operational enforcement.

In this work, we introduce *explanation-based runtime verification* for ML-driven decision systems in optical networks. Runtime verification, in this context, refers to the continuous assessment of ML outputs against predefined acceptability criteria, enabling the system to selectively accept decisions that are well-supported and defer those that exhibit contradictory or insufficient explanatory evidence. In other words, we use explanations as runtime evidence to verify the decision of the underlying ML model.

We first define an uncertainty criterion to identify potentially unreliable or borderline predictions (e.g., low confidence or small margin between classes). For each such flagged instance, we extract feature attributions using an XAI technique, namely, Shapley Additive Explanations (SHAP) [4], and analyze the direction and magnitude of feature influence at the individual decision level. These attributions are then evaluated against predefined acceptability rules that assess both decision–explanation coherence (i.e., whether the dominant features genuinely support the predicted class) and consistency with optical-layer physics (i.e., whether the influential features align with known physical principles and operational knowledge). If the dominant explanatory factors consistently support the predicted outcome and correspond to meaningful network drivers, the decision, despite being uncertain, is accepted. Otherwise, it is deferred to conservative handling procedures or fallback mechanisms, such as human supervision. The objective is therefore to selectively identify and suppress unreliable uncertain decisions, rather than indiscriminately discarding all low-confidence predictions, so as to maintain

This work has been supported by the EUREKA CELTIC-NEXT SUSTAINET-Advance project, funded by the Swiss Innovation Agency Innosuisse No. 119.588 INT-ICT, and by Vinnova (Sweden's Innovation Agency) No. 2025-02987. Corresponding author: omran.ayoub@supsi.ch

978-3-903176-78-2

a high level of automation while significantly mitigating the risk of erroneous decisions.

We hypothesize that predictions exhibiting incoherent, unstable, or physically inconsistent explanatory patterns are inherently more prone to being unreliable or spurious. By validating the reasoning structure underlying each prediction, the system can identify decisions whose internal logic conflicts with domain principles or operational knowledge, thereby reducing the risk of incorrect control actions. The overall goal is to strengthen the trustworthiness and operational safety of ML-driven optical network automation through explanation-aware consistency checks embedded directly in the deployment pipeline, without altering the underlying predictive model or imposing significant computational overhead.

We demonstrate the effectiveness of our approach on a representative use case of binary lightpath Quality of Transmission (QoT) classification. Experimental results on two open datasets show that the proposed approach supports the selective rejection of high-risk approvals, lowering the false positive rate while sustaining a high level of automation.

II. Preliminaries: Feature Attribution

This section introduces the fundamental concepts of XAI and feature attribution, as the latter constitutes a core building block of our approach.

XAI encompasses methods and frameworks designed to make the decision-making processes of ML and AI models transparent and understandable to humans. Its primary objective is to "open" black-box models by generating interpretable explanations that reveal how input features influence predictions. Beyond transparency, XAI allows practitioners to uncover relationships learned by the model, identify influential input features and verify alignment with domain knowledge. It further supports model debugging and human oversight in critical scenarios.

The most widely adopted XAI approaches are post-hoc explanation methods. These techniques are applied after model training, without modifying the internal structure or predictive behavior of the model, and aim to interpret how a given decision has been reached. Depending on their scope, post-hoc explanations can provide either global or local insights. Global explanations characterize the overall behavior of the model across the dataset, for example by revealing the influence of each feature and verifying consistency with known trends. Local explanations, in contrast, focus on a single prediction and identify the specific features that contributed most to that individual outcome.

In this work, we rely on local explanations and specifically on feature attribution methods, i.e., feature-based explanations that quantify the influence of each input variable on the model output. Figure 1 illustrates an example of a global explanation (left) and a local explanation (right)[1]. The local explanation plot quantifies the contribution of individual features to a specific prediction, distinguishing between features that support

[1] [5] provides a detailed interpretation of these plots in the context of lightpath QoT estimation

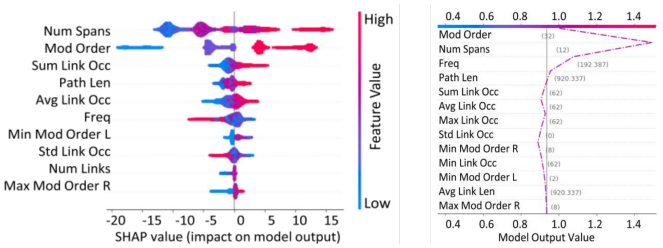

Fig. 1: Example of a global explanation (left) and a local explanation (right) generated using SHAP for a lightpath QoT classification model. The global explanation visualizes the overall feature impact (SHAP values) across the dataset for the unacceptable QoT class, characterizing the general behavior of the model. For example, it indicates that a higher number of spans, higher-order modulation formats, increased link occupation, and longer path length contribute to degraded lightpath QoT, in agreement with domain knowledge. The local explanation, in contrast, presents the feature attributions for a single prediction, quantifying how each input feature contributes to that specific decision. Further details on how to interpret these plots are provided in [5].

the model's decision (positive contribution) and those that counteract it (negative contribution). The magnitude of each contribution reflects the strength of its influence on the final output, thereby enabling a fine-grained understanding of the reasoning behind a single automated decision. In our work, we employ feature attribution to measure feature influence at the level of individual decisions. Methods such as SHAP [4] provide a principled way to decompose a prediction into additive feature contributions, supporting both local and aggregated analyses. This feature-level quantification of influence constitutes the foundation of our runtime verification framework, as it enables systematic assessment of whether the reasoning reflected in the attribution scores is operationally and physically consistent before a decision is enacted.

III. Related Work

XAI has been increasingly investigated in the context of optical networks, mainly to enhance transparency, extract operational insights, and understand the influence of input features on ML model decisions. However, existing works primarily exploit explainability for post-hoc interpretation and knowledge distillation, rather than for runtime verification or operational validation of ML-driven decisions.

Several studies focused on XAI for lightpath QoT estimation. In [5], XAI techniques were applied to a supervised QoT classification model to extract insights into feature relevance and analyze misclassifications. Similarly, [6], [7] leveraged SHAP-based explanations to quantify the influence of input features and identify combinations of feature values driving QoT predictions. In [8], XAI was used to identify the most impactful features, enabling feature set reduction to lower telemetry load and computational complexity while preserving predictive performance. Along the same line, [9] integrated XAI-driven feature selection with transfer learning techniques to enhance model adaptability and reliability across heterogeneous datasets. Beyond technical performance, [10] investigated interpretable AI models for optical network management tasks and assessed the impact of explainability on operator trust and fault localization.

Explainability has also been exploited for fault management. In [11], XAI was applied to interpret ML-based fault localization decisions using optical signal measurements. Similarly, [12] incorporated SHAP and Local Interpretable Model Agnostic Explanation frameworks to enhance transparency and operator trust in ML-based real-time fault detection using Optical Time Domain Reflectometer data. Moreover, in the context of traffic prediction and dynamic network optimization, [13] used XAI to identify which transmission and network state parameters most influence bandwidth blocking probability in elastic optical networks, while [14] extracted insights into traffic prediction models by analyzing feature relevance across traffic types and aggregation levels. More recently, explainability has been explored in reinforcement learning (RL)-based control frameworks. In [15], Shapley-based explanations were integrated with a deep RL agent for network slice admission control to analyze and validate rejection decisions prior to deployment. Similarly, [16] enhanced Shapley-based explainability for RL-driven routing and spectrum assignment to interpret proactive lightpath rejections.

Apart from explainability, several works have focused on enhancing the reliability of ML models for optical networks through uncertainty quantification (UQ) and probabilistic modeling. These approaches aim to provide calibrated confidence intervals, prediction margins, or distributional estimates to support risk-aware decision making. For instance, calibrated regression and probabilistic QoT modeling techniques were proposed in [17], [18], while deep quantile regression and Monte Carlo dropout were leveraged in [19], [20] to explicitly represent prediction uncertainty and reduce overly conservative design margins. More recently, conformal prediction frameworks have been introduced to provide statistically guaranteed QoT estimates under domain shift [21].

While these approaches quantify predictive uncertainty at the output level, they do not explicitly validate the internal reasoning structure of individual decisions. In contrast, our work leverages model explanations as runtime evidence to assess decision–explanation coherence and physics consistency. Importantly, explanation-based verification is orthogonal to uncertainty-aware modeling and can be naturally combined with UQ or conformal prediction techniques to further enhance decision reliability.

Finally, we note that a few works have explored the use of explanations for verification-oriented purposes during operation. In [22], SHAP explanations were used as secondary signals to distinguish between previously seen and unseen network attacks, effectively leveraging feature attributions for robustness enhancement. Similarly, [23] proposed lightweight interpretable verifiers to perform near-real-time consistency checks of DRL agents in Open RAN environments. While these approaches use interpretability as an auxiliary verification mechanism, they do not explicitly incorporate domain knowledge or physics-based constraints into the validation process. Our work follows a similar direction but incorporates operational and physical consistency considerations into explanation-based verification for optical networks.

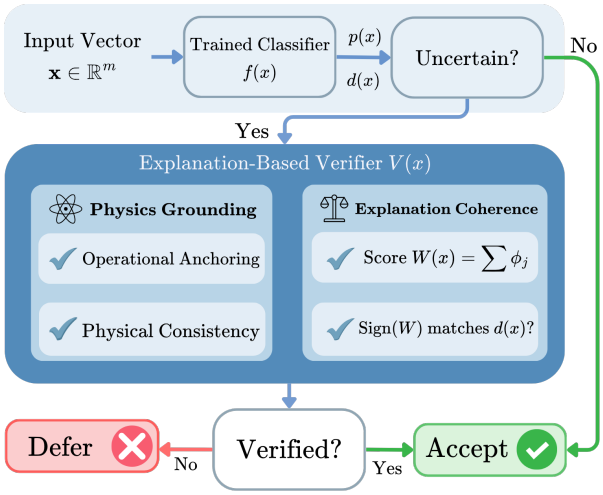

Fig. 2: Schematic representation of the proposed approach.

IV. Explanation-based Runtime Verification

The objective of our work is to introduce a runtime mechanism that evaluates, on a per-decision basis, i.e., at inference time, whether a given prediction should be trusted for execution, focusing specifically on predictions flagged as uncertain. We formalize this mechanism as a runtime verifier

$$V(\mathbf{x}) \in \{\texttt{ACCEPT}, \texttt{DEFER}\}, \tag{1}$$

which operates immediately after inference and prior to execution.

The verifier relies exclusively on information available at runtime and derived from model's explanations. The focus is on identifying weakly supported decisions before they propagate into automated network control. For a number of identified uncertain decisions, the verifier will do specific checks and then decide whether to accept the decision or to defer it. The goal is to preserve high automation rate (i.e., reduce human intervention) while reducing the error rate among enforced decisions.

As an illustrative case study, we consider binary lightpath QoT classification, where $d(\mathbf{x}) = 1$ denotes a predicted feasible lightpath.

Figure 2 provides an overview of the proposed approach, referred to as *XAI-based*, and its main components. Consider a trained binary classifier $f : \mathbb{R}^m \to [0, 1]$ deployed within an optical network control system. Given an input vector $\mathbf{x} \in \mathbb{R}^m$, the model outputs a probability $p(\mathbf{x})$ and a decision

$$d(\mathbf{x}) = \mathbb{1}[p(\mathbf{x}) \geq \theta] \tag{2}$$

where θ is a predefined threshold.

To identify predictions requiring additional scrutiny, we compute predictive entropy from the classifier's output probabilities[2]. Samples are ranked by entropy, and the top $q\%$ most uncertain predictions are flagged and forwarded to the runtime verifier.

[2]Entropy is higher when the model assigns similar probabilities to both classes, indicating uncertainty

For uncertain decisions, we compute a feature attribution vector

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})], \tag{3}$$

where $\phi_j(\mathbf{x})$ quantifies the contribution of feature j toward the positive class[3]. However, it is important to note that the framework can be applied to a wide range of ML tasks such as binary classification and multi-class classification.

A. Verification Checks

We consider two verification checks.

Physics Grounding: The first verification aims to ensure that decisions are supported by operationally meaningful and physically consistent evidence. We therefore introduce a *physics grounding* check.

Let $\mathcal{P}$ denote the set of core operational features known to influence ML model's decision, such as path length, number of spans, modulation order, and which correspond to physically interpretable parameters directly related to signal degradation mechanisms and feasibility conditions [5].

At runtime, we restrict attention to the dominant explanatory factors $\mathcal{T}_K(\mathbf{x})$, defined as the top-$K$ features ranked by absolute attribution magnitude. The grounding check enforces two complementary requirements:

- **Operational anchoring:** At least one dominant explanatory feature must belong to $\mathcal{P}$. This aims to ensure that the decision is supported by core impairment-related features rather than weakly-correlated attributes.
- **Physics consistency:** For a subset of features that show a monotonic impact on model's decisions during training, the direction of their contribution at inference must align with basic physical intuition. This property was verified by inspecting the global explanations of the model (e.g., feature importance rankings and aggregated SHAP value distributions), as reported in [5].

Concretely, a decision is deferred if a dominant impairment-related feature appears among the top-K attributions but its contribution contradicts expected monotonic behavior. For example, if a high modulation format or an unusually long path length provides strong positive attribution toward the feasible class, the decision is flagged. The intuition is that enforcing monotonic consistency constraint can reduces the likelihood that the model relies on spurious correlations or unstable statistical artifacts that contradict general domain knowledge.

Coherence: The second verifiability check relates to the coherence between the model's decision and the dominant part of the explanation. It aims to enforce logical consistency between the model's decision and its own dominant explanatory evidence.

Let $\mathcal{T}_K(\mathbf{x})$ denote the indices of the top-K features ranked by absolute attribution magnitude $|\phi_j(\mathbf{x})|$. We define a witness score

$$W(\mathbf{x}) = \sum_{j \in \mathcal{T}_K(\mathbf{x})} \phi_j(\mathbf{x}). \tag{4}$$

[3]We obtain attributions using SHAP but any XAI technique that provides local explanations can be used.

Also here, K is a design parameter that determines how many dominant explanatory features are considered when assessing the decision. The attribution value $\phi_j(\mathbf{x})$ represents the contribution of feature j to the prediction of the positive class. Positive values indicate that the feature pushes the decision toward class 1, while negative values indicate that it pushes the decision toward class 0.

Because $\phi_j(\mathbf{x})$ quantifies directional support toward the positive class, the sign of $W(\mathbf{x})$ captures the *net support* provided by the dominant explanatory factors. In other words, $W(\mathbf{x})$ summarizes whether the strongest drivers (strongest features) of the prediction collectively argue in favor of feasibility or infeasibility (and consequently, in favor of the decision or not).

The coherence condition is defined as:

$$\begin{cases} W(\mathbf{x}) > 0, & \text{if } d(\mathbf{x}) = 1, \\ W(\mathbf{x}) < 0, & \text{if } d(\mathbf{x}) = 0. \end{cases} \tag{5}$$

If the classifier predicts feasibility ($d(\mathbf{x}) = 1$), then the strongest contributing features should collectively provide positive support toward feasibility. Conversely, if the classifier predicts infeasibility ($d(\mathbf{x}) = 0$), the dominant features should collectively provide negative support. If the explanation contradicts the decision, i.e., the most influential features argue in the opposite direction of the predicted class, this condition is violated and hence flagged by the verifier. For example, in the lightpath QoT estimation use case, if the classifier predicts that a configuration is feasible ($d(\mathbf{x}) = 1$), the dominant explanatory features should indicate conditions consistent with feasibility. On the contrary, it may happen that the top contributors show strongly negative attributions, factors typically associated with signal degradation, while the final decision is feasible. Here, the strongest evidence argues against the predicted outcome, indicating incoherent reasoning. Similarly, if a configuration is predicted infeasible ($d(\mathbf{x}) = 0$), but the dominant features collectively provide positive contributions, the explanation does not support the decision. The intuition is that such inconsistency suggests that the decision may rely on weaker secondary effects, fragile feature interactions, or borderline threshold behavior.

B. Decision Policy

The runtime verifier combines both checks to determine whether a model decision should be enacted:

$$V(\mathbf{x}) = \begin{cases} \texttt{ACCEPT}, & \text{if both conditions hold}, \\ \texttt{DEFER}, & \text{otherwise}. \end{cases} \tag{6}$$

If the constraints are satisfied, the decision is accepted; otherwise, it is labeled as `DEFER`. Deferred decisions are not directly executed but may trigger fallback strategies, conservative configurations, additional validation, or operator review.

V. Experimental Evaluation

A. Experimental Setup

To evaluate the proposed approach, we conduct experiments on two lightpath QoT estimation datasets [24]. We employ an Extreme Gradient Boosting (XGBoost) classifier as the

TABLE I: Percentage improvement over the baseline classifier (positive values indicate improvement) over Dataset 1.

Method	q	ΔAuto	ΔAcc	ΔTPR	ΔTNR	ΔFPR	ΔFNR
Baseline	–	0	0	0	0	0	0
Entropy	0.005	-0.51	0.23	0.13	1.61	44.32	55.35
	0.010	-1.01	0.31	0.18	2.25	61.85	74.17
	0.020	-2.06	0.41	0.23	2.87	79.07	97.00
	0.050	-6.50	0.42	0.24	0.25	86.90	100.00
XAI-based	0.005	-0.20	0.10	-0.00	1.67	46.08	-0.12
	0.010	-0.33	0.15	-0.00	2.36	65.11	-0.20
	0.020	-0.55	0.19	-0.00	3.03	83.49	-0.37
	0.050	-0.86	0.20	-0.00	3.25	89.67	-0.74

TABLE II: Percentage improvement over the baseline classifier (positive values indicate improvement) over Dataset 2.

Method	q	ΔAuto	ΔAcc	ΔTPR	ΔTNR	ΔFPR	ΔFNR
Baseline	–	0	0	0	0	0	0
Entropy	0.005	-0.54	0.27	0.22	0.38	23.68	27.41
	0.010	-1.06	0.45	0.40	0.59	36.68	49.56
	0.020	-2.04	0.68	0.65	0.72	45.12	81.37
	0.050	-5.06	0.91	0.77	1.26	78.92	95.59
XAI-based	0.005	-0.26	0.11	-0.00	0.38	23.99	-0.21
	0.010	-0.52	0.17	-0.00	0.60	37.27	-0.46
	0.020	-0.78	0.21	-0.01	0.75	46.65	-0.78
	0.050	-1.81	0.36	-0.02	1.30	80.94	-2.03

underlying ML model. Model hyperparameters are selected using standard validation procedures on the training folds and remain fixed across experiments.

For each dataset, we randomly sample $N = 30{,}000$ instances and perform stratified repeated holdout evaluation using five independent train–test splits. In each split, 70% of the data are used for training and 30% for testing, while preserving the original class distribution. All reported results correspond to averages across the five splits.

We compute feature attributions using SHAP. For *coherence*, we consider the top-K most influential features ranked by absolute SHAP value, with $k = 4$. For *Physics Grounding*, we focus on three domain-relevant features: *Path Length*, *Number of Spans*, and *Modulation Order*. Physics-related thresholds are derived from the 80th percentile of the corresponding feature distributions computed on the training folds. In these cases, if the sign of the SHAP contributions contradict expected physical behavior (e.g., high path length contributing positively to a feasible prediction beyond acceptable limits), the decision is rejected.

We compare XAI-based to two other approaches:

- **Baseline:** the XGBoost classifier without any rejection mechanism.
- **Entropy Rejection:** a selective rejection baseline that discards the top-q fraction of test samples with highest predictive entropy. This approach serves as a purely uncertainty-driven rejection strategy.

Selective rejection is controlled via an inspection budget parameter $q \in \{0.005, 0.01, 0.02, 0.05\}$, corresponding to inspection rates of 0.5%, 1%, 2%, and 5% of the test set.

We report both operational and predictive performance metrics computed on the accepted subset of samples (i.e., those not rejected by the verifier), following a selective classification setting. All results are reported as percentage improvements relative to the baseline classifier, where positive values indicate improvement (for FPR and FNR, improvement corresponds to a reduction). This evaluation protocol enables analysis of the risk–automation trade-off under different rejection strategies.

B. Numerical Results

Tab. I reports the percentage improvement over the baseline classifier for two deferring strategies: entropy-based rejection and the proposed XAI-based verifier, on dataset 1. Recall that class 1 corresponds to the "good" lightpath QoT; therefore, the most critical error is the false positive (i.e., approving a bad sample, $0 \to 1$).

Entropy rejection, which removes the most uncertain predictions according to predictive entropy, achieves substantial reductions in false positive rate (FPR), reaching up to 79.07% at $q = 0.02$. This confirms that predictive entropy effectively concentrates a large fraction of harmful errors within a small inspection budget. However, this reduction comes at a noticeable automation cost (e.g., -2.06% at $q = 0.02$), with increasingly aggressive rejection as q grows.

In contrast, the proposed XAI-based verifier achieves even stronger FPR reductions, up to 89.67%, while sacrificing significantly less automation (e.g., only 0.55% at $q = 0.02$). This indicates that explanation- and physics-aware verification provides more selective and targeted filtering than pure uncertainty rejection. Rather than deferring all uncertain samples, the verifier defers predictions whose explanations are incoherent, weakly anchored in operational features, or inconsistent with physical conditions (e.g., long paths or high span counts contributing positively to a "good" prediction).

Importantly, while entropy rejection slightly increases true positive rate (TPR), this effect largely stems from indiscriminate filtering of uncertain samples. In contrast, the proposed verifier preserves TPR almost unchanged while achieving stronger reductions in FPR. Such behavior is particularly desirable in safety-critical settings, where reducing dubious approvals must not compromise acceptance of genuinely good samples.

The results in Tab. II, corresponding to dataset 2, further confirm the structured advantage of the XAI-based verifier. Entropy rejection again achieves strong FPR reductions (up to 78.92% at $q = 0.05$), but at substantially higher automation loss (-5.06%). In contrast, the proposed verifier achieves comparable or superior FPR reductions with markedly smaller reductions in automation (e.g., -1.81% at $q = 0.05$). This consistent pattern indicates that entropy operates as a broad uncertainty filter, whereas the proposed method performs structured risk discrimination. The distinction becomes clearer when examining TPR. Entropy rejection produces small positive shifts in TPR (up to 0.77%), reflecting non-specific filtering effects. Conversely, the coherence+physics verifier leaves TPR essentially unchanged (variations below 0.02%), indicating that it selectively targets risky approvals without degrading acceptance of truly good samples. Although minor increases in false negative rate (FNR) are observed under the

proposed verifier, their magnitude remains negligible relative to the achieved FPR reductions. This confirms that the verifier primarily suppresses the critical error mode ($0 \rightarrow 1$) while preserving overall predictive stability.

These results suggest that integrating explanation coherence with domain-specific physical constraints may provide a structured mechanism to improve the risk–automation trade-off beyond uncertainty-based rejection alone. Unlike entropy, which treats uncertainty as a generic signal, the proposed verifier leverages domain-specific properties of the explanation, such as semantic consistency and physical plausibility, to differentiate between potentially risky approvals and merely uncertain predictions. The observed reductions in the critical $0 \rightarrow 1$ error mode, together with limited impact on automation levels, indicate that explanation-based signals carry actionable information that can be exploited for selective deferring. While the approach should be interpreted as a proof-of-concept rather than a finalized methodology, the findings highlight the potential of explanation-aware verification as a complementary safety layer alongside conventional uncertainty measures.

VI. Conclusion

In this work, we introduced an explanation-based runtime verification framework that leverages model explanations by assessing their coherence together with physics-aware consistency checks. Experimental results demonstrate that the proposed approach enables selective rejection of high-risk approvals, substantially reducing the critical false positive rate while preserving a high automation level and maintaining stable true positive performance. The proposed methodology is complementary to uncertainty quantification and conformal prediction techniques, and can be naturally integrated with them to further strengthen decision reliability. More broadly, the findings point toward the relevance of structured, concept-aware learning paradigms such as concept bottleneck models for optical networks, where intermediate, physically meaningful representations could facilitate both interpretability and verifiable control.

References

[1] P. Castoldi, F. Cugini, M. Gharbaoui, A. Giorgetti, F. Paolucci, A. L. Ruscelli, N. Sambo, A. Sgambelluri, and L. Valcarenghi, "Shaping the future of optical networks by integrating sdn, telemetry, and ai," *Journal of Optical Communications and Networking*, vol. 17, no. 7, pp. C51–C61, 2025.

[2] C. Natalino, A. Panahi, N. Mohammadiha, and P. Monti, "Ai/ml-as-a-service for optical network automation: use cases and challenges," *Journal of Optical Communications and Networking*, vol. 16, no. 2, pp. A169–A179, 2024.

[3] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM computing surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[5] O. Ayoub, S. Troia, D. Andreoletti, A. Bianco, M. Tornatore, S. Giordano, and C. Rottondi, "Towards explainable artificial intelligence in optical networks: the use case of lightpath qot estimation," *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. A26–A38, 2022.

[6] O. Ayoub, A. Bianco, D. Andreoletti, S. Troia, S. Giordano, and C. Rottondi, "On the application of explainable artificial intelligence to lightpath qot estimation," in *Optical Fiber Communication Conference*. Optica Publishing Group, 2022, pp. M3F–5.

[7] O. Ayoub, D. Andreoletti, S. Troia, S. Giordano, A. Bianco, and C. Rottondi, "Quantifying features' contribution for ml-based quality-of-transmission estimation using explainable ai," in *2022 European Conference on Optical Communication (ECOC)*. IEEE, 2022, pp. 1–4.

[8] H. Fawaz, F. Arpanaei, D. Andreoletti, I. Sbeity, J. A. Hernández, D. Larrabeiti, and O. Ayoub, "Reducing complexity and enhancing predictive power of ml-based lightpath qot estimation via shap-assisted feature selection," in *2024 International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2024, pp. 1–6.

[9] S. Aladin, L. Wosinska, and C. Tremblay, "Automated, interpretable and efficient ml models for real-world lightpaths' quality of transmission estimation," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 9785–9801, 2025.

[10] S. Kumar, P. Sukhroop, V. Sharma, A. Kumar, and R. Rani, "Building trust through explainable ai in adaptive optical transport networks," *Journal of Optical Communications*, no. 0, 2025.

[11] O. Karandin, O. Ayoub, F. Musumeci, Y. Hirota, Y. Awaji, and M. Tornatore, "If not here, there. explaining machine learning models for fault localization in optical networks," in *2022 International Conference on Optical Network Design and Modeling (ONDM)*. IEEE, 2022, pp. 1–3.

[12] C. Jenila, C. Gowtham, T. Dani Akash, M. Hemanth *et al.*, "Performance-enhanced explainable ai-assisted fault detection in optical fiber networks using otdr analysis," in *2025 7th International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2025, pp. 853–858.

[13] R. Goścień, "Traffic prediction-and explainable artificial intelligence-based dynamic routing in software-defined elastic optical networks," in *2024 IFIP Networking Conference (IFIP Networking)*. IEEE, 2024, pp. 750–756.

[14] A. Knapińska, O. Ayoub, C. Rottondi, P. Lechowicz, and K. Walkowiak, "Explainable artificial intelligence-guided optimization of ml-based traffic prediction," in *International Conference on Optical Network Design and Modelling (ONDM)*, 2024.

[15] J. P. Asdikian, A. Amro, L. Mehyeddine, C. Natalino, I. Sbeity, G. Maier, P. Monti, S. Troia, and O. Ayoub, "Verifying behavior of reinforcement learning agents for network slice admission control," in *2025 21st International Conference on Network and Service Management (CNSM)*. IEEE, 2025, pp. 1–6.

[16] L. Mehyeddine, C. Natalino, A. Amro, J. P. Asdikian, I. Sbeity, G. Maier, P. Monti, S. Troia, and O. Ayoub, "Beyond performance: Explaining non-intuitive deep reinforcement learning actions in elastic optical networks," in *2025 European Conference on Optical Communications (ECOC)*. IEEE, 2025, pp. 1–4.

[17] N. Di Cicco, M. Ibrahimi, C. Rottondi, and M. Tornatore, "Calibrated probabilistic qot regression for unestablished lightpaths in optical networks," in *2022 international Balkan conference on communications and networking (BalkanCom)*. IEEE, 2022, pp. 21–25.

[18] N. Di Cicco, J. Talpini, M. Ibrahimi, M. Savi, and M. Tornatore, "Uncertainty-aware qot forecasting in optical networks with bayesian recurrent neural networks," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 441–446.

[19] H. Maryam, T. Panayiotou, and G. Ellinas, "Learning quantile qot models to address uncertainty over unseen lightpaths," *Computer Networks*, vol. 212, p. 108992, 2022.

[20] ——, "Representing uncertainty in deep qot models," in *2022 20th Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE, 2022, pp. 113–121.

[21] K. Rezaei, O. Ayoub, P. Monti, and C. Natalino, "Policy-driven conformal prediction for trustworthy QoT estimation," in *2026 Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2026, p. M4A.3.

[22] P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust network intrusion detection through explainable artificial intelligence (xai)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.

[23] R. Soundrarajan, C. Fiandrino, M. Polese, S. D'Oro, L. Bonati, and T. Melodia, "On ai verification in open ran," *IEEE Communications Magazine*, 2026.

[24] G. Bergk, B. Shariati, P. Safari, and J. K. Fischer, "Ml-assisted qot estimation: a dataset collection and data visualization for dataset quality evaluation," *Journal of Optical Communications and Networking*, vol. 14, no. 3, pp. 43–55, 2021.