



## **Generative Explainability for Next-Generation Networks: LLM-Augmented XAI with Mutual Feature Interactions**

Downloaded from: <https://research.chalmers.se>, 2026-04-18 13:47 UTC

Citation for the original published paper (version of record):

Rezaei, K., Ayoub, O., Troia, S. et al (2025). Generative Explainability for Next-Generation Networks: LLM-Augmented XAI with Mutual Feature Interactions. International Conference on Wireless and Mobile Computing, Networking and Communications. <http://dx.doi.org/10.1109/WIMOB66857.2025.11257542>

N.B. When citing this work, cite the original published paper.

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

# Generative Explainability for Next-Generation Networks: LLM-Augmented XAI with Mutual Feature Interactions

Kiarash Rezaei\*<sup>ORCID</sup>, Omran Ayoub†<sup>ORCID</sup>, Sebastian Troia‡<sup>ORCID</sup>, Francesco Lelli†§<sup>ORCID</sup>,  
Paolo Monti\*<sup>ORCID</sup>, Carlos Natalino\*<sup>ORCID</sup>

\* Department of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden  
{kiarashr, mpaolo, carlos.natalino}@chalmers.se

†University of Applied Sciences and Arts of Southern Switzerland, 6928 Lugano, Switzerland  
omran.ayoub@supsi.ch

‡Politecnico di Milano, 20133 Milan, Italy  
sebastian.troia@polimi.it

§Tilburg University, 5037 AB, Tilburg, Netherlands.  
f.elli@tilburguniversity.edu

**Abstract**—As artificial intelligence and machine learning (AI/ML) models become integral to network operations, their lack of transparency poses a significant barrier to operator trust. Existing explainable artificial intelligence (XAI) techniques often fail to bridge this gap for non-specialists, producing technical outputs that are difficult to translate into actionable insights. This paper presents a framework specifically designed to address this shortcoming. It leverages a moderately sized large language model (LLM) and extends beyond the standard use of SHapley Additive exPlanations (SHAP) feature influence values. The framework employs a structured prompt enriched with mutual feature interaction data to generate human-understandable natural language explanations. To validate our framework, we performed an empirical evaluation on an optical quality of transmission (QoT) estimation use case with human evaluators. We collected independent performance evaluations from specialists, which showed a high inter-evaluator agreement. Compared to a state-of-the-art baseline that uses only SHAP feature influence values in a straightforward prompt, our approach improves the explanation usefulness and scope by 12.2% and 6.2%, while achieving 97.5% correctness.

**Index Terms**—Large language model (LLM), Explainable AI (XAI), Interpretability, Explainability, Transparency.

## I. INTRODUCTION

Modern communication systems and artificial intelligence (AI) are increasingly intertwined. The rapid growth of data transport demands and widespread adoption of AI-driven applications have transformed communication networks into critical enablers of contemporary digital services [1]. Managing such systems is complex, posing significant challenges in network automation tasks related to resource allocation, fault detection, and maintaining optimal performance under

dynamic conditions [2]. Hence, AI/ML are increasingly employed to tackle these issues by providing data-driven solutions that aim at network efficiency and reliability. The broad applicability of these techniques is demonstrated by their use across diverse networking domains, from wireless mobility management and cellular resource allocation [3] to optical network resource allocation and management [4]–[6].

As AI/ML models become more integrated into network automation workflows, the need for transparency and interpretability of them has become critical [7]–[10]. Explainable artificial intelligence (XAI) techniques, such as SHAP [11], have been utilized to interpret model predictions and provide insights into model behavior. XAI is essential before model deployment (to inspect valid model behavior) and in scenarios where human supervision and/or intervention may be required to validate or override automated decisions [12]. In the context of network automation, an engineer may need to inspect specific decisions made by AI/ML-based systems, particularly when the system takes an action that appears counterintuitive, such as rerouting traffic through a seemingly congested path. In such cases, explainability mechanisms are critical for understanding the rationale behind the decision, verifying that it aligns with operational policies or inferred predictions, and deciding whether to trust the model or override its action manually. Nonetheless, the explanations extracted from XAI techniques are often presented in technical formats that are cognitively demanding. They also require domain-specific expertise to be interpreted [13], [14], posing a challenge for their broader adoption. The literature highlights this difficulty [15], [16], showing that while standard XAI methods can rank feature importance, they struggle to translate these rankings into contextually meaningful and actionable insights.

Figure 1(a) shows an illustrative example of SHAP feature influence values, where  $F_1..F_{12}$  represent arbitrary features.

This work has been partially supported by the Swedish Innovation Agency (VINNOVA 2025-01348), and by the EUREKA CELTIC-NEXT project SUSTAINET-Advance funded by the Swiss Innovation Agency.

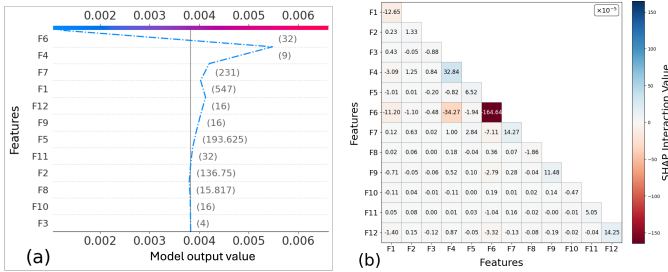


Fig. 1: Illustrative example of SHAP-based explanation for a hypothetical QoT estimation instance with 12 features: (a) Feature influence plot showing individual feature contributions. (b) Feature interaction heatmap illustrating pairwise effects.

They quantify the isolated, individual contribution of each one of them to the model’s output. Similarly, Fig. 1(b) shows an example of SHAP mutual feature interaction values. These values quantify the synergistic effect where the combined impact of two features differs from the sum of their individual influences. Both figures contain a high volume of raw, multidimensional data, including magnitude and directional influence, which makes it difficult to accurately distill the model’s core logic without significant effort and/or domain knowledge (from both XAI and the technical details of the specific use case).

In this work, we propose a framework designed to address the challenge of interpreting complex XAI outputs. Our framework leverages modern LLMs and the nuanced insights from mutual feature interaction values to generate human-understandable explanations. These explanations enhance the transparency (and perceived trustworthiness) of AI/ML models while significantly reducing the cognitive load, time, and expertise required by a human engineer to interpret XAI outputs correctly. As a case study, we focus on the quality of transmission (QoT) estimation problem in optical networks [17], [18]. This problem involves predicting whether a candidate lightpath will meet predefined signal quality thresholds (e.g., bit error rate) before it is provisioned (i.e., before it is allocated in the network). In practical scenarios, receiving clear explanations for QoT-related decisions is highly valuable for network engineers, especially when the model deems a particular lightpath as unacceptable [19], [20]. For instance, if a proposed route is rejected despite appearing viable based on traditional engineering rules, the operator must understand the model’s reasoning to assess whether the decision is based on subtle patterns (e.g., cumulative non-linear penalties or latent correlations with adjacent lightpaths) or is potentially the result of model overfitting or error. Human-readable explanations in such cases support trust, allow manual validation or override when needed, and ultimately lead to safer and more efficient network operation.

Empirical results validate the proposed framework. By guiding a reasoning LLM with a structured prompt that incorporates both SHAP feature influence and mutual feature

interaction values, we produce higher-quality explanations. To assess these explanations, we adopt human-centered evaluation metrics that capture how correct, comprehensive, and useful the generated explanations are from the perspective of network engineers. Compared to a state-of-the-art approach using only plain SHAP feature influences, our framework demonstrates significant improvements across these metrics: usefulness increased by 12.2%, scope by 6.2%, and correctness reached 97.5%. While the core of this work is an XAI-enhanced framework, its motivation lies in network automation, where timely and reliable decisions are essential [2], [4]. Our case study addresses QoT estimation, yet the framework applies broadly wherever AI/ML models require transparency. This includes tasks such as traffic engineering, mobility management, slicing, and fault recovery [1], [21], where explainability helps engineers interpret AI decisions, validate them against policies, and preserve human oversight.

## II. RELATED WORK

Recent research addressed the gap between technical model outputs and user comprehension by proposing the integration of LLMs and AI/ML to help translate complex model explanations into human-interpretable language, Zeng et al. [22] demonstrated that LLMs can effectively translate SHAP value outputs into natural language explanations. This approach has shown particular promise across diverse application domains, with studies such as [23] examining how LLMs can generate near-human-level explanations for healthcare applications, enhancing both interpretability and clinical decision-making processes. Building on these results, research in communication networks has specifically investigated the integration of LLMs with SHAP methodologies to enhance the interpretability of AI/ML models deployed in network automation systems. For instance, a pipeline combining anomaly detection, SHAP-based root cause analysis, and LLM-generated corrective actions was proposed for managing 6G microservices environments [24]. However, this evaluation relied primarily on automatic metrics (e.g., BERT, ROUGE) that assess surface-level properties like text similarity. they achieved a BERT score around 0.74 out of 1.0 using Llama2 model with 70 billion parameters. In the context of QoT estimation, ChatGPT 3.5 with 175 billion parameters has been used to enhance SHAP explanations through direct prompting and self-reflection [25], although without a structured prompt design. It achieved, with self-reflection, a correctness of 65%. Both works show a relatively low score for the explanations generated by the LLMs. Moreover, both works utilize traditional LLMs without reasoning capabilities and rely solely on isolated SHAP values as a measure of feature importance. No importance is given to mutual feature interactions that can potentially show how features jointly affect the ML model predictions.

Based on the above considerations, this paper considers two research questions: (i) do reasoning LLMs improve the performance of explanations?; and (ii) can the performance of LLMs in generating interpretations be enhanced by incorporating SHAP feature interaction values within a structured

prompt? To answer these questions, the paper presents a framework with two key advancements over the state-of-the-art. First, to explore the impact of LLM capability, we harness the power of a reasoning LLM with a moderate number of parameters. Second, to test the value of deeper contextual data, we enrich our structured prompt with SHAP mutual feature interaction values. While these values are often disregarded due to their complex characteristics (as illustrated in Fig. 1(b)), our assumption is that they provide critical context for the LLM’s interpretation process, positioning our work beyond prior studies such as [25] by combining reasoning LLMs with structured feature-interaction-aware context.

### III. LLM-AUGMENTED XAI FRAMEWORK WITH FEATURE INTERACTIONS

The proposed framework comprises four parts: (1) AI/ML model, (2) XAI method, (3) explanation augmentation module, and (4) dashboard (Fig. 2). The goal is to enhance the interpretability of AI/ML model decisions through an LLM by processing the feature contribution values extracted via SHAP.

*AI/ML Model:* The first component is the AI/ML model to be explained. The use case assumed for this work is the QoT estimation of optical lightpaths. It was selected primarily for its practical importance in ensuring service reliability and building operator trust. The technically challenging nature of the task, characterized by complex non-linear feature interactions, serves as an additional benefit, providing a robust testbed for our framework. To this end, we employ an XGBoost (XGB) model, given its proven effectiveness in prior works [25]. The model takes as input a set of features describing a candidate lightpath and outputs the estimated bit error rate (BER). It is important to note that the proposed framework remains model-agnostic, i.e., it is applicable to various AI/ML models and tasks.

*XAI Method:* To explain the model’s decisions, we apply SHAP [11], a method for post-hoc explainability quantifying local feature contributions using Shapley values. This method allows for quantifying both the individual importance of features—capturing their mutual influence on the model prediction (Fig. 1(a))—and their mutual interactions, which reflect the joint impact of feature pairs (Fig. 1(b)). Together, these outputs serve as the core inputs for the explanation augmentation module. We focus on local explanations, which provide insight into an individual model’s prediction. This is particularly relevant in an AI/ML-aided scenario where an expert reviews the model’s outputs before they are adopted in production.

*Explanation augmentation module:* The key element of this module is the structured prompt, detailed in Table A-I in the Appendix, which instructs the LLM in generating augmented explanations. This prompt incorporates the model prediction and the individual feature influence values as primary content. It is then further enriched with mutual feature interaction scores to capture the more subtle aspects of the model’s behavior.

*Dashboard:* Experts need to inspect the output of AI/ML models before these outputs, or decisions based on them, are deployed in the network. Fig. 3 illustrates a dashboard example where information about features, SHAP influence, interaction matrix, and AI/ML model output values is augmented with the textual interpretations provided by our framework.

### IV. EXPERIMENTAL AND EMPIRICAL RESULTS

To evaluate the performance of the proposed framework, we began by training an XGB regressor model using the dataset available in [26] from which we selected a subset of 12 features that are the most indicative of signal quality and topological structure. The subset includes metrics related to path and link lengths, modulation formats, lightpath and connection line rates, source and destination node degrees, number of spans, and detailed link occupation statistics. The dataset was split into training and test sets using a 90:10 ratio. The model’s performance was evaluated using the test set, yielding a mean squared error (MSE) of 0.0000 and a mean absolute error (MAE) of 0.0001, both of which indicate high accuracy. We randomly extracted 40 local SHAP explanations for our experiments<sup>1</sup>.

For the explanation augmentation module, we utilized a 4-bit quantized version of DeepSeek-R1 LLM, consisting of 32.8 billion parameters, following the work of [28]. We selected DeepSeek-R1 for its strong reasoning capabilities, which are essential for generating high-quality explanations. To enable efficient local inference with relatively fast response times, we adopted quantization instead of distillation, a choice supported by findings in [29], which show that quantization more effectively preserves reasoning performance in LLMs.

As a baseline, we employed a straightforward prompt asking the LLM to explain the rationale behind the predictions made by the XGB model based only on the SHAP feature influence values [24], [25]. An example comparing the prompt designs and generated explanations for the baseline approach and proposed frameworks is shown in Table A-II in the Appendix. The baseline and the proposed prompting approaches require fewer than 1,500 input tokens. The LLM output was also below 1,500 tokens, well below any token limit of modern LLMs.

In the final evaluation step, two human experts analyzed and assessed the generated interpretations. To ensure objectivity, the interpretations were presented in a randomized order, preventing the experts from identifying the prompting strategy that had generated each one. Additionally, the experts conducted their evaluations independently, without access to each other’s assessments, thereby minimizing potential bias. Both experts are senior researchers with extensive experience in applying XAI to optical network automation. We adopt the human-centered evaluation metrics proposed in [25], which assess the LLM-generated interpretations across three dimensions: (i) *correctness*, measuring how faithfully the

<sup>1</sup>The GitHub repository [27] contains the prompt templates, the dataset including SHAP values and SHAP interaction values, as well as the corresponding LLM explanations.

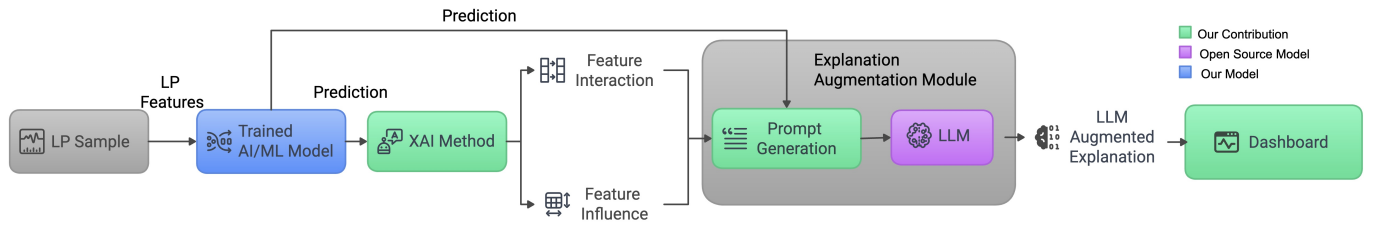


Fig. 2: LLM-Augmented XAI framework pipeline: from input sample processing (left) to the generation of LLM augmented explanations (right). Key components and contributions are color-coded.

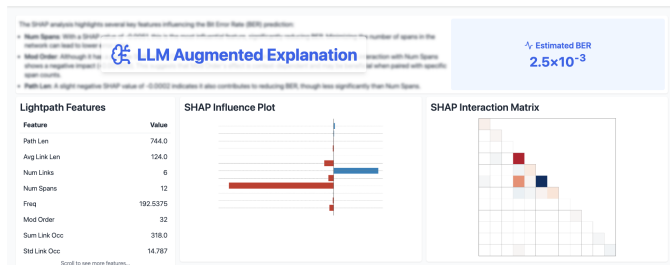


Fig. 3: Dashboard example: QoT estimation (top right) and LLM augmented explanation (top left) based on mutual feature interactions and influences (bottom part).

interpretation reflects the underlying explanation; (ii) *scope*, evaluating whether the interpretation, when correct, highlights the most critical aspects of the explanation<sup>2</sup>; and (iii) *usefulness*, gauging, when correct, how helpful the interpretation is in supporting human understanding. Correctness and scope are rated using binary labels (e.g., correct/incorrect, in/out of scope). Usefulness is rated on a scale of 0 to 5, reflecting the extent to which the explanation helps the evaluator understand the model’s behavior. Table I reports the empirical results. Results for *scope* and *usefulness* are only considered when the expert evaluates the interpretation as correct. We also report the agreement among experts across *correctness* and *scope*, as well as the standard deviation of *usefulness*.

Results indicate that the baseline strategy achieves an average correctness of 96.3%, while the proposed framework yields a slightly higher correctness of 97.5%. Evaluator agreement on correctness is high for both strategies, at 93% and 95%, respectively. These results confirm that the integration of a reasoning LLM, mutual feature interaction data, and a structured prompt yields a significant performance improvement over methods available in the literature [25]. This is particularly evident in the scope of the explanations, where the proposed framework exhibits a notably higher score (95.0%) compared to the baseline (88.8%). This substantial improvement suggests that the use of mutual feature interactions helps the LLM generate more comprehensive and contextually relevant explanations. Regarding usefulness, the evaluations of experts further support the benefits of the proposed framework.

<sup>2</sup>An interpretation may be accurate but still fail to emphasize the most relevant features.

TABLE I: Results Showing the Average (Avg), Agreement (Agr), and Standard Deviation (Std).

Metric	Baseline		Proposed	
	Avg	Agr/Std	Avg	Agr/Std
Correctness	96.3%	93% / -	97.5%	95% / -
Scope	88.8%	83% / -	95.0%	90% / -
Usefulness	3.77	- / 1.5	4.38	- / 1.4

Avg = Average, Agr = Agreement, Std = Standard Deviation.

The baseline approach received an average usefulness rating of 3.77 (std 1.4), while the proposed framework achieved a higher rating of 4.38 (std 1.5). This further supports the finding that the structured design and the explicit inclusion of SHAP mutual feature interaction values significantly improve the clarity and interpretability of the explanations.

## V. CONCLUSION

This paper presents a framework designed to bridge the gap between complex explanations from XAI methods and human-readable interpretations. It leverages a moderately-sized LLM with advanced reasoning capabilities, and a structured prompt that incorporates both individual SHAP feature influence values and mutual feature interaction data. We evaluated our proposed framework against a state-of-the-art prompting approach that utilizes SHAP influence values. Both used a reasoning LLM and were found to be effective, with the proposed framework consistently outperforming the basic approach. These results highlight the importance of considering structured prompts with feature-interaction-aware context. While SHAP offers detailed insights, its cost in complex AI/ML models invites exploring lighter methods like LIME or gradient-based approaches within this framework [30]. Ultimately, the adoption of this approach in operational settings can enable scalable interpretability of AI/ML model outputs, improve transparency when using AI/ML models in real-world deployments, and increase the overall trustworthiness of decisions made based on AI/ML outputs.

## APPENDIX

This appendix contains supplementary tables referenced in the main text. Table A-I specifies the complete, multi-part prompt structure designed for our proposed framework, and Table A-II shows a qualitative comparison of the generated explanations to highlight the practical differences in output quality between our approach and the state-of-the-art baseline.

TABLE A-I: Prompt Structure Design for Explanation Augmentation Module

Component	Description	Example Snippet from Prompt
[Context]	<p>Defines the LLM’s persona and the problem domain.</p> <ul style="list-style-type: none"> <li>- <b>Role:</b> Interpreter of a SHAP explanation.</li> <li>- <b>Target:</b> Bit Error Rate (BER) prediction for optical lightpaths.</li> <li>- <b>Rules:</b> Specifies how to interpret positive/negative SHAP values.</li> </ul>	<p>You are an interpreter of a SHAP explanation... predicting Bit Error Rate (BER)... Positive SHAP → Increases BER (undesired)</p>
[Task]	<p>Outlines the primary instructions for the LLM.</p> <ol style="list-style-type: none"> <li>1. Identify the 2-3 most impactful features from SHAP values.</li> <li>2. Describe each feature’s value and its influence on BER.</li> <li>3. Summarize findings with actionable insights to reduce BER.</li> </ol>	<ol style="list-style-type: none"> <li>1. Identify the top 2 or 3 features...</li> <li>3. Summarize with a concise explanation covering... Practical insights to reduce BER</li> </ol>
[Interaction Handling]	<p>Provides conditional logic for complex or unexpected results.</p> <ul style="list-style-type: none"> <li>- <b>Trigger:</b> If a feature’s effect contradicts domain knowledge.</li> <li>- <b>Action:</b> Analyze SHAP interaction values. to find the root cause.</li> <li>- <b>Output:</b> Revise the conclusion to reflect interaction-driven effects.</li> </ul>	<p>If a feature seems to reduce BER in a way that contradicts domain expectations: 1. Refer to the SHAP interaction values...</p>
[Input Explanation]	<p>The raw data provided to the LLM for a single instance.</p> <ul style="list-style-type: none"> <li>- Includes the model’s prediction (ML Model Prediction).</li> <li>- Includes individual feature data (Feature-wise breakdown).</li> <li>- Includes pairwise interaction data (Feature Interaction Breakdown).</li> </ul>	<p>Input Explanation: ML Model Prediction: 0.00024... Feature-wise breakdown: - Num Spans: Min Value = 2.0, ...,Value= 13, SHAP Value= 0.0059 ... Feature Interaction Breakdown: - Num Spans &amp; Mod Order: SHAP Interaction Value = -0.000862 ...</p>
[Response Structure]	<p>Defines the required format for the LLM’s output.</p> <ol style="list-style-type: none"> <li>1. <b>Interpretation:</b> A detailed breakdown of key features and their effects.</li> <li>2. <b>Conclusion:</b> A concise summary with actionable insights.</li> </ol>	<p>Response Structure 1. Interpretation ... 2. Conclusion ...</p>

TABLE A-II: Qualitative Comparison of Baseline and Proposed Frameworks for an Example Instance

Component	Baseline Approach	Proposed Framework
Prompt	<p>Below is the output of a SHAP explanation from a machine learning model that predicts Bit Error Rate (BER) in optical networks. Please write a short explanation of what this means and which features seem most important... Feature-wise breakdown: - Path Len: Min Value = 84.0, ..., SHAP Value = 0.0001 ...</p>	<pre>## [Context] You are an interpreter of a SHAP explanation... [Task] 1. Identify the top 2{3 features... ## [Interaction Handling]... (see Table A-I for full details).</pre>
Generated Explanation	<p><b>Interpretation</b> The SHAP (SHapley Additive exPlanations) output explains how different features contribute to the model’s prediction of Bit Error Rate (BER). In this case, the model predicts a BER of approximately 0.00045.</p> <p><b>Key Insights:</b></p> <ol style="list-style-type: none"> <li>1. <b>Most Important Features:</b> <ul style="list-style-type: none"> <li>- <b>Num Spans:</b> This feature has the largest negative contribution (-0.0033), meaning that increasing the number of spans in the network reduces the predicted BER.</li> <li>- <b>Mod Order (Modulation Order):</b> This feature has the largest positive contribution (+0.0028), indicating that higher modulation orders increase the predicted BER.</li> </ul> </li> <li>2. <b>Other Notable Features:</b> <ul style="list-style-type: none"> <li>- <b>Path Len:</b> Slightly increases the predicted BER (+0.0001).</li> <li>- <b>Freq, Sum Link Occ, and others:</b> These features have negligible contributions to the prediction, as their SHAP values are close to zero.</li> </ul> </li> </ol> <p><b>Summary:</b> The model’s prediction of a low BER (0.00045) is primarily influenced by the number of spans in the network and the modulation order used. More spans tend to reduce errors, while higher modulation orders increase them.</p>	<p><b>Interpretation</b> The SHAP analysis reveals that the Bit Error Rate (BER) prediction of approximately 0.00045 is significantly influenced by two key features:</p> <ol style="list-style-type: none"> <li>1. <b>Modulation Order (Mod Order):</b> With a SHAP value of +0.0028, Mod Order has the most substantial positive impact on increasing BER. The current value is set to its maximum at 64, which contributes to higher error rates.</li> <li>2. <b>Number of Spans:</b> This feature has the largest negative SHAP value (-0.0033), indicating that fewer spans reduce BER. The current span count is 6.0. Additionally, a notable interaction exists between Mod Order and Number of Spans (SHAP interaction value -0.001591). This suggests that when both factors are present, their combined effect on reducing BER is more pronounced than the sum of their individual effects.</li> </ol> <p><b>Conclusion</b> To optimize BER reduction:</p> <ol style="list-style-type: none"> <li>1. <b>Reduce Modulation Order:</b> Lowering Mod Order from 64 to a feasible minimum (e.g., closer to 16) could significantly decrease BER.</li> <li>2. <b>Optimize Number of Spans:</b> Further reducing spans below 6, if technically viable, would enhance BER improvement, especially when combined with lower Mod Order.</li> </ol> <p>By addressing these factors, particularly focusing on lowering Mod Order and maintaining fewer spans, the system can achieve a more substantial reduction in BER.</p>

## REFERENCES

- [1] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 8845070, 2024.
- [2] New York State Wireless Association, "The future of telecom infrastructure: The network on which AI will be built," <https://nyswa.org/the-future-of-telecom-infrastructure-the-network-on-which-ai-will-be-built/>, Mar. 2025, accessed: 2025-04-30.
- [3] M. El-Hajji, "Enhancing communication networks in the new era with artificial intelligence: Techniques, applications, and future directions," *Network*, vol. 5, no. 1, 2025.
- [4] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An overview on application of machine learning techniques in optical networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383–1408, 2019.
- [5] C. Natalino, A. Panahi, N. Mohammadiha, and P. Monti, "AI/ML-as-a-Service for optical network automation: use cases and challenges Invited," *J. Opt. Commun. Netw.*, vol. 16, no. 2, pp. A169–A179, Feb 2024.
- [6] E. Etezadi, C. Natalino, R. Diaz, A. Lindgren, S. Melin, L. Wosinska, P. Monti, and M. Furdek, "Deep reinforcement learning for proactive spectrum defragmentation in elastic optical networks," *J. Opt. Commun. Netw.*, vol. 15, no. 10, pp. E86–E96, Oct 2023.
- [7] Y. Wu, G. Lin, and J. Ge, "Knowledge-powered explainable artificial intelligence for network automation toward 6G," *IEEE network*, vol. 36, no. 3, pp. 16–23, 2022.
- [8] S. Wang, M. A. Qureshi, L. Miralles-Pechuan, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Applications of explainable AI for 6G: Technical aspects, use cases, and research challenges," *arXiv preprint arXiv:2112.04698*, 2021.
- [9] 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.
- [10] O. Ayoub, C. Natalino, and P. Monti, "Towards explainable reinforcement learning in optical networks: The RMSA use case," in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2024, p. W4I.6.
- [11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.
- [12] B. Dutta, A. Krichel, and M.-P. Odini, "The challenge of zero touch and explainable AI," *Journal of ICT Standardization*, vol. 9, no. 2, pp. 147–158, 2021.
- [13] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal, "Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence," in *Information Systems and Neuroscience: NeuroIS Retreat*. Springer, 2021, pp. 237–246.
- [14] L. Weber, S. Lopuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, 2023.
- [15] U. Ahmed, Z. Jiangbin, A. Almogren *et al.*, "Explainable AI-based innovative hybrid ensemble model for intrusion detection," *Journal of Cloud Computing*, vol. 13, no. 1, p. 150, 2024.
- [16] D. C. Bikkasani and M. R. Yerabolu, "AI-driven 5G network optimization: A comprehensive review of resource allocation, traffic management, and dynamic network slicing," *American Journal of Artificial Intelligence*, vol. 8, no. 2, pp. 55–62, 2024.
- [17] C. Rottondi, L. Barletta, A. Giusti, and M. Tornatore, "Machine-learning method for quality of transmission prediction of unestablished lightpaths," *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A286–A297, 2018.
- [18] S. Allogba, S. Aladin, and C. Tremblay, "Machine-learning-based light-path QoT estimation and forecasting," *Journal of Lightwave Technology*, vol. 40, no. 10, pp. 3115–3127, 2022.
- [19] 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 *European Conference on Optical Communication (ECOC)*, 2022, p. We3B.4.
- [20] 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 *International Conference on Optical Network Design and Modeling (ONDM)*, 2024.
- [21] D. Adanza, C. Natalino, L. Gifre, R. Muñoz, P. Alemany, P. Monti, and R. Vilalta, "IntentLLM: An AI chatbot to create, find, and explain slice intents in TeraFlowSDN," in *IEEE 10th International Conference on Network Softwarization (NetSoft)*, 2024, pp. 307–309.
- [22] X. Zeng, "Enhancing the interpretability of SHAP values using large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2409.00079>
- [23] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *Informatics*, vol. 11, no. 3, 2024.
- [24] A. Mekrache, M. Mekki, A. Ksentini, B. Briki, and C. Verikoukis, "On combining XAI and LLMs for trustworthy zero-touch network and service management in 6G," *IEEE Communications Magazine*, vol. 63, no. 4, pp. 154–160, 2025.
- [25] O. Ayoub, S. Troia, C. Natalino, C. Rottondi, D. Andreoletti, F. Lelli, S. Giordano, and P. Monti, "Natural language interpretability for ML-based QoT estimation via large language models," *International Conference on Transparent Optical Networks (ICTON)*, p. Tu.C2.4, 2025.
- [26] 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.
- [27] K. Rezaei, "LLM-augmented XAI with mutual feature interactions: The QoT estimation use case," <https://github.com/kiarashRezaei/llm-for-xai-qotEstimation>.
- [28] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.
- [29] N. Zhang, Y. Zhang, P. Mitra, and R. Zhang, "When reasoning meets compression: Benchmarking compressed large reasoning models on complex reasoning tasks," 2025. [Online]. Available: <https://arxiv.org/abs/2504.02010>
- [30] D. Vale, A. El-Sharif, and M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law," *AI and Ethics*, vol. 2, no. 4, pp. 815–826, 2022.