Enhancing Reliability of Lightpath QoT Estimation Models using Bias Mitigation Techniques

Enhancing Reliability of Lightpath QoT Estimation Models using Bias Mitigation Techniques

Hussein Jammal[1], Omran Ayoub[2], Andrea Bianco[1], Michel Owayjan[3,4], Cristina Rottondi[1]

[1]Politecnico di Torino, Torino, Italy [2]University of Applied Sciences of Southern Switzerland, Switzerland
[3]Nantes University, France [4]American University of Science and Technology, Beirut, Lebanon

***Abstract*—Machine learning (ML) models for lightpath quality of transmission (QoT) estimation have shown high predictive performance but may exhibit unfair behavior by treating certain subgroups of lightpaths, such as those defined by specific modulation formats, unequally. This can lead to systematic underperformance on particular categories, raising concerns in practical deployment. In this work, we address this issue by developing a fair ML-based QoT estimation framework that explicitly minimizes disparities in prediction error across modulation format levels. Our approach integrates bias mitigation techniques at multiple stages of the ML pipeline, aiming to equalize performance across subgroups with minimal degradation of the overall accuracy. Experimental results demonstrate a substantial reduction in disparity, up to 95%, while maintaining ROC-AUC and accuracy reduction within 0.2% w.r.t. the performance of two baseline models that does not incorporate fairness considerations.**

***Index Terms*—Quality of Transmission, Optical Networks, Bias Mitigation, Fair Machine Learning.**

I. Introduction

Lightpath Quality of Transmission (QoT) estimation has been a core research focus in optical networks for more than a decade [1], [2]. Early efforts relied heavily on direct physical measurements, which provided real-world insights but were often limited by measurement overhead and scalability challenges [3]. As optical networks have grown in size and complexity, researchers increasingly turned to Machine Learning (ML) methods for QoT estimation [2], [4], where ML models are trained to predict the QoT of a prospective lightpath based on a set of descriptive attributes, such as the lightpath's path length and modulation format. These approaches leverage monitoring data and powerful classification algorithms to infer QoT with high accuracy, alleviating the need for continuous direct measurements.

The predominant focus in the development of ML-based approaches for lightpath QoT estimation has been on improving overall predictive performance, typically measured in terms of aggregate estimation quality or misclassification error [5], [6]. However, while the overall predictive performance of a model may appear satisfactory, it often masks discrepancies across different subgroups of lightpaths. For example, a model may demonstrate high accuracy when estimating the QoT of lightpaths using one modulation format (e.g., QPSK) but significantly underperform for lightpaths using another (e.g., 16-QAM). This is because the model optimization procedures may inherently favor one group over another. Figure 1 shows an example of such a scenario, where an ML model trained to predict lightpath QoT shows a large discrepancy in accuracy achieved across lightpaths of modulation formats 2-QPSK and 4-QAM w.r.t. the other modulation formats. Such inconsistencies in performance across subgroups (in this case, modulation format levels), may lead to unreliable estimations for certain categories of lightpaths, potentially exposing network operators to unexpected errors in QoT assessment and undermining network service guarantees.

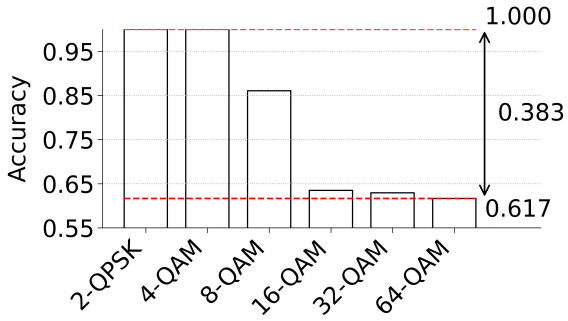

Fig. 1: Example of the discrepancies in terms of model's predictive accuracy across modulation formats.

To address this challenge, in this work we focus on developing a fair ML model for lightpath QoT estimation, where fairness is defined as minimizing disparities in prediction error between modulation format levels, ensuring the model does not systematically underperform on lightpaths using a specific modulation format. Our assumption is that an ML-based approach showing a consistent predictive performance across modulation formats is more desirable by network operators than approaches that exhibits disparities. Said so, it is worth noting that optimizing for fairness often introduces a trade-off: a model designed to be fair may sacrifice some degree of overall accuracy compared to a model optimized solely for minimum estimation error. The key challenge, therefore, is to develop a fairness-aware QoT estimation model that effectively balances this trade-off, i.e., that achieves equitable performance across modulation format levels with minimal degradation of the global prediction quality.

To this end, we propose an approach that combines bias mit-

979-8-3315-9777-1/25/$31.00 ©2025 IEEE

igation techniques at different stages of the ML pipeline with the aim of minimizing the disparities in predictive performance across considered subgroups, while minimally impacting the overall predictive performance. We first propose the use of a set of fairness metrics to quantify disparities in estimation error across subgroups and then apply our approach, leveraging pre-processing [7], in-processing [8] and post-processing [9] bias mitigation techniques. We compare our proposed pipeline against two baseline approaches using a publicly available dataset. Results demonstrate that our method can significantly enhance fairness (i.e., minimize disparities) across modulation formats for lightpath QoT estimation, exhibiting improvements between 75% to 95% in three different fairness evaluation metrics) while minimally degrading the predictive performance w.r.t. to two baseline models not incorporating any fairness constraints.

Similar to our work, [10] addressed the presence of bias in lightpath QoT classification by focusing on reducing the accuracy disparities among different feature groups using a bias mitigation technique. Specifically, the work employs a sample-weighting strategy, assigning each sample a weight to correct feature imbalance across modulation formats, effectively reducing accuracy gaps among samples with different feature values by up to 45%, and the standard deviation by up to 36% from the original model. Our work differs as we introduce and formulate the problem of fair lightpath QoT estimation treating the modulation format as a protected attribute (i.e., an attribute whose influence on predictions should not lead to inequities in performance across different modulation formats), and leverage bias mitigation techniques, including preprocessing, in-processing, and post-processing strategies to reduce the impact on global accuracy of the model.

II. Preliminaries

Several fairness metrics are commonly used to evaluate the equitable behavior of classification models.

- *Demographic Parity Difference (DPD)*: Measures the difference in the selection rate, i.e., the probability of being predicted as *accepted*, across all modulation formats. Minimizing DPD ensures that all modulation formats are equally likely to be classified as acceptable.
- *Equalized Odds Difference (EOD)*: Assesses the disparity in error rates (e.g., false positive rates) across different modulation formats. A low EOD indicates that the likelihood of misclassifications is consistent across groups, preventing any modulation format from being disproportionately misclassified. In other words, minimizing EOD means no modulation format faces high misclassifications rates over the other formats.
- *Predictive Value Parity (PVP)* focuses on precision of the classifier, i.e., the proportion of *acceptable* predictions that are actually correct. Achieving predictive parity ensures that no specific modulation format is favored in terms of prediction reliability.

Bias mitigation techniques are generally divided into three main categories, as follows:

- *Pre-processing* techniques focus on data-level adjustments, often by re-balancing class distributions or transforming feature representations.
- *In-processing* techniques apply fairness constraints into the model while training, adapting the model's objective function to ensure equal treatment across modulation formats.
- *Post-processing* Techniques are applied once the classification model has been successfully trained. They adjust the model's final predictions by recalibrating it or applying decision thresholds for each modulation format to align with specific fairness metrics, without altering the model's internal parameters.

III. Problem Statement and Methodology

Problem Statement. The lightpath QoT estimation problem is formulated as a binary classification problem, where each candidate lightpath is classified as *acceptable (class 1)* or *unacceptable (class 0)* depending on whether the expected bit error rate at the receiver side falls above or below a predefined threshold [11]. Each prospective lightpath is characterized by a set of features, describing the ligthpath itself (e.g., in terms of total length, number of links and spans, modulation format in use) and its spectral proximities (e.g., characteristics of the spectrally adjacent left/right neighbors).

We formulate the *fair lightpath QoT estimation problem* as a fair binary classification task. Our goal is not only to accurately predict whether a candidate lightpath meets the required QoT but also to ensure that the predictive performance is equitably distributed across all modulation formats. This means that no particular modulation format should consistently benefit from or suffer due to systematic biases in the learning model. In this context, we aim to develop a fair QoT estimator that satisfies the following dual objectives: 1) maximizing the overall classification accuracy, maintaining high fidelity in the identification of acceptable and unacceptable lightpaths, and 2) minimizing disparities in model performance metrics across modulation formats.

Methodology. We propose an approach combining three bias mitigation techniques, including preprocessing, in-processing and post-processing techniques, throughout the ML model development pipeline.

As a pre-processing bias mitigation technique, we employ *Oversampling* to balance the class distribution within each modulation format. Specifically, we apply the *Synthetic Minority Over-sampling Technique (SMOTE)* separately to each modulation format. For every format, synthetic samples of the minority class (which corresponds to class 0 in our case) are generated by identifying the k-nearest neighbors of each minority sample and interpolating new instances along the line segments connecting the sample to its neighbors. The interpolation is performed using the following formula:

$$x_{\text{new}} = x + \delta \cdot (x_{\text{nn}} - x),$$

where x_{nn} is one of the k-nearest neighbors of the original minority sample x, and $\delta \sim U(0, 1)$ is a random scalar drawn

from a uniform distribution. In our implementation, we set $k = 5$. This approach effectively reduces class imbalance within each modulation format, thereby lowering the likelihood of misclassifications in QoT predictions.

As an In-processing bias mitigation technique, we implement *Exponentiated Gradient (EG)* to optimize a composite loss function that balances the XGBoost binary logistic loss function and a fairness constraint based on Equalized Odds. Formally, the loss can be defined as

$$L = L_{\text{pred}} + \lambda\, L_{\text{fair}},$$

where L_{pred} is the classifier's loss function, L_{fair} quantifies disparities in false positive and false negative rates across modulation formats and λ is a hyperparameter that controls the trade-off between accuracy and fairness. During training, EG iteratively updates a weight distribution over the training examples using an exponentiated update rule:

$$w_i(t+1) = \frac{w_i(t)\, \exp\Big(-\eta\, \ell_i(t)\Big)}{\sum_j w_j(t)\, \exp\Big(-\eta\, \ell_j(t)\Big)},$$

where $\ell_i(t)$ is the loss for the ith sample at iteration t and η is a learning rate parameter. In practical terms, applying EG with Equalized Odds ensures that the model adjusts its training focus toward reducing error rate disparities among different modulation formats, ultimately leading to a more balanced and fair QoT prediction across all groups [12].

As a post-processing bias mitigation technique, we apply *Threshold Optimizer (TO)*, which works by calibrating the model's outputs once the model has been trained. TO is built to satisfy Equalized Odds as specified fairness constraint with no remaining disparity. For each modulation format group, TO generates a set of candidate thresholds

$$T = \{t_1, t_2, \ldots, t_k\}$$

over the range of the classifier's prediction scores. The fairness constraint (Equalized Odds) is evaluated by computing the difference in False Positive Rate (FPR) and False Negative Rate (FNR) between each modulation format and the overall results. In addition, the objective function, which is the balanced accuracy, is defined as

$$BA(t) = \frac{1}{2}\left(TPR(t) + TNR(t)\right)$$

and is computed for each candidate threshold t. Thus, the optimal threshold t^* for each modulation format is selected according to the following equation:

$$t^* = \arg\max_{t \in T} BA(t) \quad \text{subject to } EOD(t) \leq \varepsilon.$$

where ε is a predefined tolerance ($\varepsilon = 0.005$ in our case) that defines the maximum allowed difference in the error rates between a certain modulation format and the overall set. In our implementation, candidate thresholds are evaluated in steps of 0.01 within the observed score range, and the combination of optimized computed thresholds.

TABLE I: Comparison among models for bias mitigation: performance and fairness metrics.

Model	Performance Metrics			Fairness Metrics		
	Acc.	AUC	F1-score	DPD	EOD	PVP
No-Fairness	0.991	0.985	0.995	0.269	0.028	0.016
FTU	0.903	0.757	0.942	0.138	0.966	0.389
Fair QoT (ours)	0.991	0.982	0.991	0.012	0.007	0.005

Baseline Approaches. We implemented two baseline approaches:

1) *No Fairness*: A standard XGBoost classifier is implemented with no fairness constraints applied.
2) *Fairness Through Unawareness (FTU)*: In this approach, the model is trained without the protected attribute (i.e., modulation format) to eliminate its direct impact on predictions. This is a starting baseline method for bias mitigation; however, it has some drawbacks, since removing the attribute does not guarantee eliminating the bias due to feature correlations, while it can negatively impact classification performance.

IV. Experimental Settings

Dataset Description. We used a modified version of the dataset [11], consisting of approximately 1.3 million samples, where each sample is described by 32 features and an associated binary label indicating acceptable/unacceptable QoT (*class 1* and *class 0*). The original dataset includes modulation formats 2-QAM, 4-QAM, 8-QAM, 16-QAM, 32-QAM, and 64-QAM. However, we dropped 2-QAM and 4-QAM as they consist solely of positive (class 1) samples, making them unsuitable for the computation of the thresholds of these classes in the post-processing technique. As a result, the dataset consists of around 430,000 samples. We used a subset of 13 features identified as the most influential features in [13].

Train-test Split. We split the above described dataset into training (60%), validation (20%), and testing (20%) sets and further employ a 5-fold cross-validation.

Evaluation Metrics. We evaluate the predictive performance of our model using Accuracy, ROC-AUC, and F1-Score, and assess its fairness using the metrics introduce in Sec. II, namely, DPD, EOD and PVP. Additionally, we use the False Positive Rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR) and True Negative Rate (TNR) to compare the model's performance across modulation formats.

V. Results

Predictive Performance vs. Fairness. Table I reports the averaged results of accuracy, AUC, F1-score, along with DPD, EOD, and PVP (lower is desired) for the three approaches, namely, *No-Fairness*, *FTU*, and our proposed approach, *Fair QoT*. Results show that *Fair QoT* can achieve an accuracy equal to that of the *No-Fairness* model (0.991) and falls only a little short in terms of AUC and F1-score (*Fair QoT* achieves an AUC of 0.982 and an F1-score of 0.991, compared to *No-Fairness* model of 0.985 and 0.995, respectively). Note that,

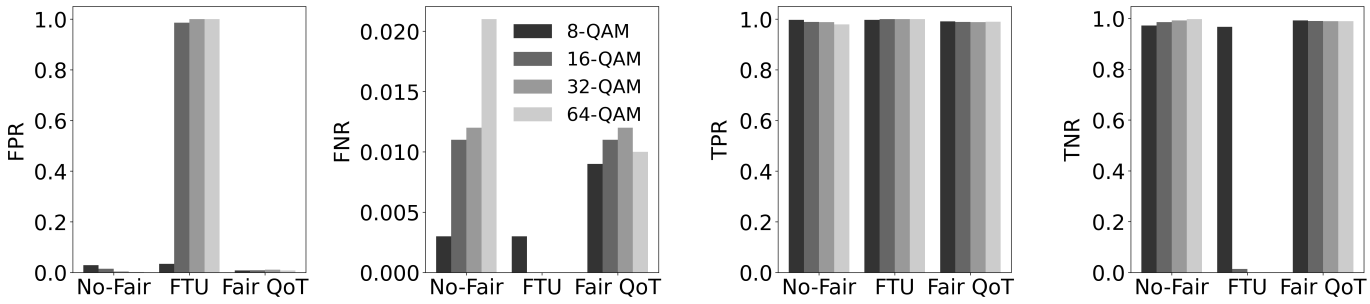

Fig. 2: Models Comparison by Modulation Format: False Positive and Negative Ratios, True Positive and Negative Ratios

as expected, *FTU* shows a degraded performance in terms of accuracy (0.903), AUC (0.757) and F1-score (0.942) with respect to the *No-Fairness* model and *Fair QoT*. Concerning the three fairness metrics, results show that *Fair QoT* achieves a near-optimal performance in terms of DPD, EOD and PVP, achieving 0.012, 0.007 and 0.005 respectively, thus significantly enhancing fairness w.r.t. that of *No-Fairness* model (0.267, 0.028, and 0.016, respectively) and that of *FTU* (0.138, 0.966, and 0.389, respectively). This shows that our proposed approach can significantly enhance fairness (i.e., eliminate bias and reduce disparity in performance across modulation formats) while minimally impacting the overall predictive performance (AUC and F1-score are marginally impacted).

Performance Across Modulation Format Levels. Fig. 2 presents the FPR, FNR, TPR and TNR achieved by the different approaches across various modulation formats. Overall, the results demonstrate that *Fair-QoT* delivers consistently strong and balanced performance across all four metrics and modulation formats. In particular, *Fair-QoT* maintains low FPR and FNR values while achieving high TPRs (exceeding 0.98) and TNRs (close to 0.99). This indicates that the model provides accurate and equitable predictions across all modulation formats, aligning with our goal. In contrast, both *No-Fair* and *FTU* exhibit notable variability in performance across modulation formats. For example, while *No-Fair* attains a very low FNR of 0.002 for 8-QAM, it increases to 0.02 for 64-QAM, reflecting inconsistency in classification accuracy. Furthermore, *FTU* consistently underperforms, as it is characterized by elevated FPRs and reduced TNRs for several formats. This suggests that FTU struggles particularly with correctly identifying *unacceptable (class 0)* lightpaths, leading to frequent misclassifications and degraded reliability.

VI. Conclusion

We introduced the problem of fair lightpath QoT estimation, where fairness is defined as minimizing disparities in prediction error across modulation format levels. In contrast to state-of-the-art approaches that primarily focus on maximizing the overall predictive performance, our proposed method incorporates bias mitigation techniques to reduce disparities in model performance across modulation format subgroups, while minimally affecting overall accuracy, as demonstrated by our experimental results.

Acknowledgments

This work has been partially supported by the Italian Ministry for University and Research under the PRIN program (grant n2022YA59ZJ - ZeTON) and by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") and by the EUREKA cluster CELTIC-NEXT project SUSTAINET-Advance funded by the Swiss Innovation Agency.

References

[1] S. Aladin, A. V. S. Tran, S. Allogba, and C. Tremblay, "Quality of transmission estimation and short-term performance forecast of lightpaths," *Journal of Lightwave Technology*, vol. 38, no. 10, pp. 2807–2814, 2020.

[2] I. Sartzetakis *et al.*, "Accurate quality of transmission estimation with machine learning," *J. Opt. Commun. Netw.*, vol. 11, no. 3, pp. 140–150, 2019.

[3] T. Panayiotou *et al.*, "Performance analysis of a data-driven quality-of-transmission decision approach on a dynamic multicast-capable metro optical network," *J. Opt. Commun. Netw.*, vol. 9, no. 1, pp. 98–108, 2016.

[4] L. Zhang, X. Li, Y. Tang, J. Xin, and S. Huang, "A survey on qot prediction using machine learning in optical networks," *Optical Fiber Technology*, vol. 68, p. 102804, 2022.

[5] T. Panayiotou, G. Savva, B. Shariati, I. Tomkos, and G. Ellinas, "Machine learning for qot estimation of unseen optical network states," in *Optical Fiber Communication Conference*. Optica Publishing Group, 2019, pp. Tu2E–2.

[6] L. Barletta, A. Giusti, C. Rottondi, and M. Tornatore, "Qot estimation for unestablished lighpaths using machine learning," in *Optical Fiber Communication Conference*. Optica Publishing Group, 2017, pp. Th1J–1.

[7] Z. Chen *et al.*, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 4, pp. 1–30, 2023.

[8] M. Hort *et al.*, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM J. Responsible Comput.*, vol. 1, no. 2, pp. 1–52, 2024.

[9] M. Hardt *et al.*, "Equality of opportunity in supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[10] C. Natalino, B. Shariati, P. Safari, J. K. Fischer, and P. Monti, "Analysis and mitigation of unwanted biases in ml-based qot classification tasks," in *2024 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE, 2024, pp. 1–3.

[11] G. Bergk *et al.*, "Ml-assisted qot estimation: a dataset collection and data visualization for dataset quality evaluation," *J. Opt. Commun. Netw.*, vol. 14, no. 3, pp. 43–55, 2021.

[12] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International conference on machine learning*. PMLR, 2018, pp. 60–69.

[13] O. Ayoub *et al.*, "Towards explainable artificial intelligence in optical networks: the use case of lightpath qot estimation," *J. Opt. Commun. Netw.*, vol. 15, no. 1, pp. A26–A38, 2022.