Synthetic Data Generation using Diffusion Models for ML-based Lightpath Quality of Transmission Estimation Under Extreme Data Scarcity

Synthetic Data Generation using Diffusion Models for ML-based Lightpath Quality of Transmission Estimation Under Extreme Data Scarcity

Davide Andreoletti[1], Cristina Rottondi[2], Omran Ayoub[1], Andrea Bianco[2]
[1]University of Applied Sciences and Arts of Southern Switzerland {name.surname@supsi.ch}
[2]Politecnico di Torino {name.surname@polito.it}

ABSTRACT
Generative diffusion models are gaining attention as a promising solution for synthetic data generation, offering a distinct advantage over traditional statistical methods and basic generative models. This work focuses on evaluating the effectiveness of such models in the context of estimating Lightpath Quality of Transmission (QoT) in optical networks, especially when real data availability is strongly limited. Numerical results demonstrate that leveraging diffusion models for data augmentation can significantly improve QoT classification accuracy and F1-score when available data are limited to a few dozens of samples. These findings highlight the potential of generative diffusion models in improving data-driven tasks for optical network management under sparse data conditions.
Keywords: Quality of Transmission Estimation; Machine Learning; Diffusion Models; Dataset Augmentation;

1. INTRODUCTION

Assessing the Quality of Transmission (QoT) for prospective lightpaths is crucial for the strategic planning of optical network infrastructures. Traditionally, QoT predictions were derived from precise mathematical models, such as the split-step Fourier method [1], or through the use of margined formulas [2]. While the former offers high accuracy, it suffers from scalability challenges in practical network environments. On the other hand, margined formulas are computationally efficient but can lead to underutilization of network resources due to their highly conservative assumptions.

Recently, Machine Learning (ML) algorithms have emerged as a promising solution for QoT estimation, addressing the scalability issues while achieving high prediction accuracy [3]. However, these ML models require substantial training datasets to achieve acceptable predictive capabilities. In the domain of optical networking gathering large datasets can be difficult due to a variety of factors. For example, the telemetry equipment already available in the network may be insufficient to monitor the totality of the deployed lightpaths and the installation of additional equipment may be deemed too costly. Another possible motivation for data shortage is that, in the early life stage of a network, the limited number of established lightpaths in greenfield network deployments restricts the volume of data available for collection. To overcome the issues of data scarcity, synthetic data generation methodologies are exploited to augment the training datasets. These techniques learn how to model the probability distribution of a dataset, in turn allowing for the sampling of synthetic data points, possibly conditioned to a given event (e.g., the probability distribution of lightpath's QoT conditioned on adopting a given modulation format for transmission).

Diffusion models [4] have recently emerged as a promising methodology for augmenting training datasets from a smaller original set across various data formats, including tabular data [5, 6]. In this paper, we explore the advantages of enhancing ML training sets with synthetic data generated by diffusion models under extreme data scarcity scenarios. We focus on a QoT binary classification problem, i.e., determining if the Signal to Noise Ratio (SNR) of a candidate lightpath falls below or exceeds a set threshold. We experiment with varying proportions of real and synthetic data within the training sets, as well as with different training set sizes, and assess the impact of training data augmentation on prediction performance in terms of classification accuracy and F1 score. In particular, we focus on reducing the number of incorrect predictions where candidate lightpaths that would yield unacceptable QoT are wrongly classified as acceptable. This type of prediction error has the most severe consequences, as it would lead to the deployment of a lightpath with insufficient QoT, potentially causing a violation of service level agreements by the network operator.

Results show that, when the amount of available field measurements is in the order of a few tens, dataset augmentation via diffusion models improves QoT classification accuracy by up to 5% and F1-score by up to 54%. Such improvement is mainly due to the major reduction of false negatives, which drop on average by two thirds thanks to the inclusion of synthetically-generated data in the training set.

The remainder of the manuscript is organised as follows: Sec.2 details the considered scenarios and defines the QoT estimation problem under study, whereas Sec.3 describes the methodology we adopted for dataset augmentation. Sec.4 reports the effectiveness of our proposed data augmentation procedure and Sec.5 draws some conclusive remarks.

2. PROBLEM STATEMENT

We consider the problem of predicting the QoT achievable along a candidate lightpath before its actual deployment in the network. Each lightpath is characterized by a set of features, such as total length, number of spans, transmitted power, modulation format in use, etc. We assume the availability of a limited amount of field measurements gathered from a paucity of already deployed lightpaths, where each lightpath is associated with a binary label indicating whether the lightpath QoT is acceptable or not. Such a label is obtained by comparing the SNR measured at the receiver to a given acceptability threshold: SNR values higher than the threshold indicate that the lightpath QoT is acceptable, whereas SNR values below the threshold indicate unacceptability. We aim at training a ML classification model capable of predicting the QoT binary class, providing as input the sets of lightpath features and the associated QoT binary label. To enhance the predictive capabilities of the model, we create a training dataset consisting of: i) the available field-measured samples; and ii) several synthetically generated samples, produced by means of diffusion models according to the procedure described in Sec.3.

3. METHODOLOGY

Diffusion models [4] learn how to generate data by transforming a Gaussian distribution into the probability distribution of a target dataset. They operate in two main phases. In the first phase, known as *forward process*, the model incrementally adds noise to the data from the dataset, gradually making it indistinguishable from a Gaussian distribution. In the second phase, namely the *backward process*, the model is trained to estimate the score function, which is the gradient of the log probability distribution of the data at various noise levels. Once trained, the diffusion model samples a data point from a Gaussian distribution and, using the estimated score function, iteratively removes the noise, eventually obtaining a data point as if it was sampled from the target distribution. In this work, we train a diffusion model to learn the probability distribution of lightpaths, conditioned to the quality of the signal, classified as either acceptable/not acceptable (class 0/1, when the SNR measured at the receiver is above/below a reference threshold value). The diffusion model is based on a neural network architecture, trained to estimate the score function, that comprises an input layer, an intermediate layer with 20 neurons employing sigmoid activation functions, and a final output layer with a single neuron using linear activation.

4. EXPERIMENTAL SETTINGS

4.1 Data Description

We consider the dataset available in [7], consisting of 585 samples collected using a combination of laboratory experiments and field measurements in real-world optical communication networks. The set of 12 lightpath features considered in our experiments includes: signal launch power, modulation format and depth, lightpath and span length, fiber attenuation coefficient, splice losses, optical amplifier gain, Polarization Mode Dispersion (PMD) coefficient and compensation technique, external temperature and humidity levels. Each sample is associated with a label indicating the QoT class. In total, the dataset includes 538 samples with label 0 and 47 samples with label 1, This reflects realistic conditions faced by network operators, who rarely maintain active lightpaths with insufficient QoT, thus leading to shortage of samples in such class. In the following, samples drawn from this dataset will be referred to as “real samples”.

4.2 ML Model Training and Testing

We train an XGB model for the classification task at hand. We perform our evaluations following an 80-20 split of the dataset, where 80% of the available samples are used for training (note that, as explained in the next subsection, the actual amount of real samples used depends on the considered scenario) and the remaining 20% for testing. We repeat each experiment 100 times, randomly selecting the train/test split.

4.3 Scenarios

We consider two benchmark scenarios. In the first one, named *All Real*, the entire original dataset is assumed to be available. This scenario represents an upper bound of the ML model’s performance for the classification task at hand. In the second one, named *Fraction Real*, only a subset of the samples contained in the original dataset is assumed to be available, thus simulating a condition of data shortage. We perform our experiments varying the size of such subset from 5% to 30% of the original dataset size. Moreover, to validate our data augmentation approach, we consider two additional scenarios, namely AUG_50% and AUG_80%, where we complement the training dataset, which already includes the considered fraction of available real data as in the *Fraction Real* scenario, by including synthetically-generated data samples starting from the available real ones. In the AUG_50% scenario, the generation process considers an equal sampling of data points across the two classes, whereas in the AUG_80% scenario a sampling skewed with a probability of 80% towards the less represented class (i.e., class 1) is used. We perform a sensitivity analysis by varying the amount of synthetic data samples generated, as detailed in the next Section.

5. NUMERICAL RESULTS

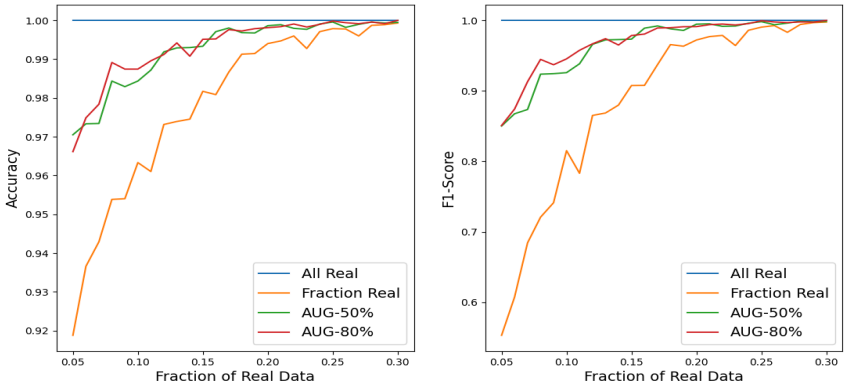

Fig. 1. Accuracy (left) and F1-score (right) achieved in the AUG-50% and AUG-80% scenarios depending on the amount of available real samples, benchmarked against the *All Real* and *Fraction Real* scenarios.

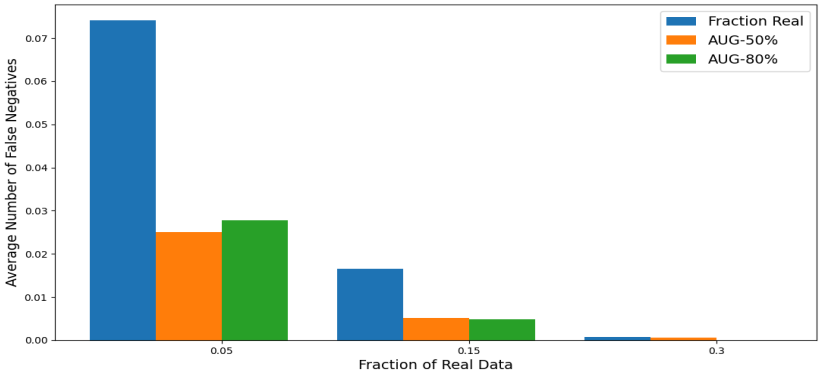

Fig. 2. False Negative rate achieved in the AUG-50% and AUG-80% scenarios depending on the amount of available real samples, benchmarked against the *Fraction Real* scenario (In the omitted *All Real* scenario the False Negative rate is 0).

We start by analysing the impact of the probability of sampling (i.e., the probability of augmenting the data of a given class). In this experiment, we generate enough synthetic data points to complement the available real samples to have as many data points as in the original dataset (i.e., 585). Fig.1 reports the accuracy (left) and the F1-score (right) achieved by the ML model in the AUG-50% and AUG-80% scenarios. First, we highlight that the ML model achieves an accuracy and an F1-score of 1.00 in the *All Real* benchmark, when the whole original dataset is used. Conversely, in the *Fraction Real* benchmark, where only a fraction of the real data is used (without any data augmentation), the ML model achieves an accuracy ranging between 0.92 (when using only 5% of the original dataset) and 1.00 (when 30% of the original dataset is used), and an F1-score ranging between 0.55 (with 5% of real data) and almost 1.00 (with 30% of real data). While the high accuracy indicates that the model generally produces correct predictions, the trend of the F1-score shows that it may be biased towards the majority class, thus lacking robustness in classifying samples of the minority class (unacceptable QoT). When data augmentation is adopted, in both AUG-50% and AUG-80% scenarios, the ML model achieves an accuracy ranging between 0.97 and 1.00, with a slight advantage for AUG-80%. Compared to the *Fraction Real* scenario, employing data augmentation shows a larger advantage under higher data scarcity (accuracy of 0.97 instead of 0.92 when only 5% of real data is available). In terms of F1-score, the advantage of data augmentation is more evident, as the ML model can achieve an F1-score of 0.85 in both data augmentation scenarios, when only 5% of the original dataset is available. This shows that data augmentation via synthetic samples can be exploited at the very early stages of the network deployment, when only a very limited number of lightpaths are installed (and, consequently, monitorable).

In Fig. 2 we report the average number of false negatives achieved in three scenarios (AUG-50%, AUG-80% and *Fraction Real*) for different fractions of real data utilized (0.05, 0.15 and 0.3). Results show that augmenting the data, considering either AUG-50% or AUG-80%, can significantly reduce the number of false negatives (by 50% to 70% for a fraction of real data of 0.05 and 0.15, respectively). This highlights the benefit for the network operator in reducing the amount of wrong predictions that would lead to the establishment of lightpaths with unacceptable QoT.

We now focus on the AUG-80% scenario and investigate the effect of varying the amount of synthetic data samples on the performance of the ML model. This experiment differs from our previous experiments, where we constrained the generation of synthetic samples to match the size of the original (real) dataset. Instead, we now

consider three settings, namely AUG-50, AUG-100, AUG-150, where we add to the real data 50, 100, and 150 synthetic samples, respectively. Fig. 3 reports the accuracy (Fig. 3(a)) and the F1-score (Fig. 3(b)) of the ML model across the three above-mentioned settings, compared to the *All Real* and *Fraction Real* benchmarks. Results show that, in terms of accuracy, increasing the amount synthetic samples from 50 to 150 does not yield significant improvement, i.e., for a given fraction of real data, the accuracy achieved by the ML model in the various settings varies minimally (0.96 accuracy for AUG-50 and 0.965 for AUG-150). Differently, in terms of F1-score, the advantage brought by increasing the amount of synthetic samples is significant only under extreme cases of data scarcity (e.g., for 5% of real data, the ML model achieves an F1-score of 0.8 in AUG-50, whereas in AUG-150 the achieved F1-score is 0.86), However, the advantage is reduced when a larger fraction of real data is used, i.e., for a fraction of real data of 15% or higher, the F1-score across the three settings differs only slightly (variations within 1-2%).

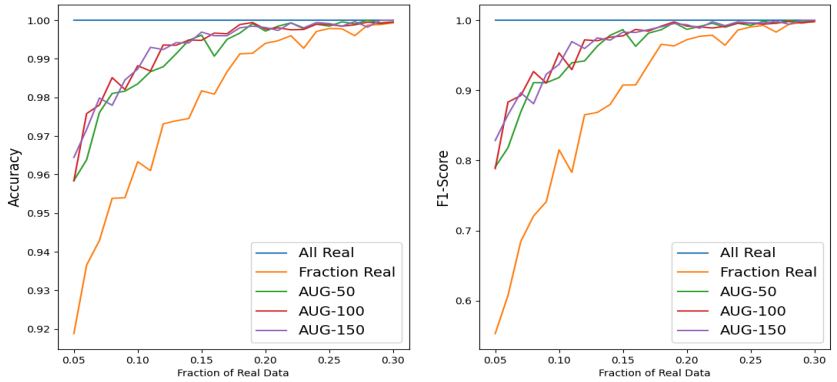

Fig. 3. Accuracy (left) and F1-score (right) achieved in the AUG-50, AUG-100 and AUG-150 scenarios depending on the amount of available real samples, benchmarked against the *All Real* and *Fraction Real* scenarios.

6. CONCLUSION

We addressed the issue of data scarcity for the problem of machine learning-based lightpath QoT estimation. We leveraged diffusion models to generate synthetic samples and train machine learning with the aim of improving lightpath QoT estimation accuracy. By effectively augmenting limited real training datasets with synthetic samples, these models yielded a notable enhancement in classification accuracy and F1-score, obtained by drastically reducing the amount of false negatives, when only a few dozens of training samples are available.

ACKNOWLEDGEMENTS

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").

REFERENCES

[1] Sinkin, O. V., Holzlöhner, R., Zweck, J., & Menyuk, C. R. (2003). *Optimization of the split-step Fourier method in modeling optical-fiber communications systems*. Journal of lightwave technology, 21(1), 61.

[2] Sartzetakis, I., Christodoulopoulos, K., Tsekrekos, C. P., Syvridis, D., & Varvarigos, E. (2016). *Quality of transmission estimation in WDM and elastic optical networks accounting for space–spectrum dependencies*. Journal of Optical Communications and Networking, 8(9), 676-688.

[3] Ayassi, R., Triki, A., Crespi, N., Minerva, R., & Laye, M. (2022). *Survey on the use of machine learning for quality of transmission estimation in optical transport networks*. Journal of Lightwave Technology, 40(17), 5803-5815.

[4] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., ... & Yang, M. H. (2023). *Diffusion models: A comprehensive survey of methods and applications*. ACM Computing Surveys, 56(4), 1-39.

[5] Nguyen, Quang, et al. *Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation*. Advances in Neural Information Processing Systems 36 (2024).

[6] Kotelnikov, Akim, et al. "Tabddpm: Modelling tabular data with diffusion models." International Conference on Machine Learning. PMLR, 2023.

[7] *OptiCom Signal Quality Dataset*, Accessed on April 1, 2024. Url: https://www.kaggle.com/datasets/tinnyrobot/opticom-signal-quality-dataset