Failure Prediction in Optical Transport Networks Through the Integration of Digital Twins and Deep Learning

Failure Prediction in Optical Transport Networks Through the Integration of Digital Twins and Deep Learning

Mashboob Cheruvakkadu Mohamed, Renato Ambrosone, Muhammad Umar Masood, Gulmina Malik, Stefano Straullu, Antonino Nespola, Sai Kishore Bhyri, Antonio Napoli, João Pedro, Gabriele Maria Galimberti, Walid Wakim and Vittorio Curri

***Abstract*—This work introduces an advanced comprehensive framework for predictive maintenance by integrating Digital Twin (DT) with multiple Deep Learning (DL) models with the aim of predicting amplifier failures in optical networks. Using GNPy (Gaussian Noise model in Python), an open source framework, a DT is created to emulate network behavior under both normal and failure conditions, enabling the generation of synthetic datasets representative of amplifier degradation fault scenarios. These datasets are used to train DL models based on Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Long- and Short-Term Time-Series Networks (LSTNet). A comparative analysis shows that all models exhibit strong performance, with LSTM achieving an accuracy of 99% and LSTNet, CNN models closely following at 98% and 96% respectively, demonstrating the ability of these DL models to identify complex temporal and statistical patterns in network telemetry data, facilitating accurate prediction of early failures. This integrated solution provides a scalable and data-driven approach for proactive fault management, improving operational capabilities in optical transport networks.**

***Index Terms*—Deep Learning, Digital Twin, Failure Prediction, Network Reliability, Optical Networks, Proactive Failure Management**

I. Introduction

In today's digital era, characterized by rapid developments in artificial intelligence and the widespread use of bandwidth-intensive applications, the need for extremely high data transmission rates has increased significantly. Optical fiber communication networks, which serve as the backbone of high-speed data infrastructure, play a crucial role in addressing this growing demand. As a result, service providers are increasingly focusing on enhancing the efficiency of existing systems to fully exploit available capacity. This push toward optimization reduces system margins. As current networks transition to lower operational margins while continuing to rely on traditional static failure management methods, the chances of critical failures increase [1]. The resulting extended network down-time can lead to substantial financial losses, motivating network operators to prioritize proactive management strategies over reactive approaches.

Mashboob Cheruvakkadu Mohamed (mashboob.cheruvakkadu@polito.it), Renato Ambrosone, Muhammed Umar Masood, Gulmina Malik and Vittorio Curri are with Politecnico di Torino, Italy. Stefano Straullu and Antonino Nespola are with LINKS Foundation, Italy and Sai Kishore Bhyri, Antonio Napoli, João Pedro, Gabriele Maria Galimberti and Walid Wakim are with Nokia - Optical Networks. João Pedro is also with Instituto de Telecomunicações, Instituto Superior Técnico.

Manuscript received xx, 2025; revised xx, 2025.

Conventional failure management approaches, predominantly utilize threshold-based techniques, in which fixed, predefined limits are used to identify anomalies or signs of performance degradation [2]. Although these methods are relatively simple to implement, they are inadequate in handling the complexities and dynamic behavior of modern network environments. To address these complexities with the exponential growth of traffic load, there is a growing need for smarter and more flexible failure management approaches that can understand the current state of the network and detect problems in advance. This work addresses the challenge of localizing soft-failures in optical amplifiers characterized by amplifier gain degradation, which leads to a decline in Optical Signal-to-Noise Ratio (OSNR), impacting the Quality of Transmission (QoT). At present, such failures can be difficult to detect and diagnose using the standard monitoring parameters exposed by network equipment and line cards. The persistence of such silent failures can eventually lead to hard failures, significantly impacting network performance and reliability. Predicting failures before they occur allows operators to take action in advance, which helps prevent service disruptions. This kind of early warning system improves network reliability, reduces downtime, and reduces costs associated with replacing or repairing equipments.

In recent years, the use of Machine Learning (ML) techniques in the field of optical network failure management has attracted growing interest from both academia and industry [3]–[10].This surge in attention stems from the ability of ML to analyze vast amounts of network data, uncover hidden patterns, and make intelligent inferences without relying on predefined rules. A wide range of research efforts have been dedicated to exploring how ML can be applied for fault detection, isolation, and localization, particularly soft-failures [11]–[15]. These studies show that ML can provide reliable and adaptable solutions for recognizing and diagnosing subtle network faults that often go unnoticed by traditional monitoring techniques. These soft-failures, which may not trigger fixed threshold-based methods, can be addressed more effectively through data-driven models capable of identifying complex and evolving patterns in network behavior. However, to ensure the practical deployment and reliability of an ML model in the context of failure management, it is imperative that the model is trained on a comprehensive and representative data set [16]. This implies that training data should cover various types of failure conditions, including common and rare cases, as

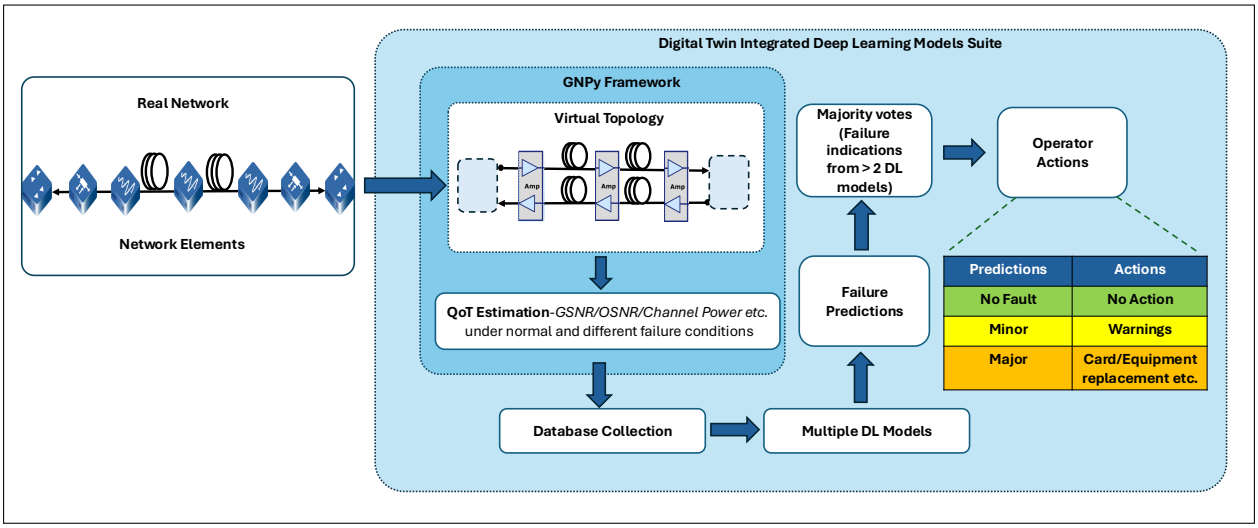

Fig. 1: DT integrated DL models suite for failure prediction in optical transport networks- an illustration.

well as varying operating environments and system behaviors. Learning from such a broad dataset improves the model's capability to recognize and predict failures, making it more reliable and effective when applied in real systems, where unexpected failures can occur. Therefore, building a diverse and comprehensive dataset is a key step in developing an ML model that performs well in practical failure management applications.

However, collecting labeled failure condition datasets from deployed optical networks for training ML models remains challenging due to the uneven occurrence of different fault types. Some failures particularly rare or transient ones occur infrequently or only under specific conditions. As a result, the datasets collected are often unbalanced, with certain failure types significantly underrepresented. Additionally, in experimental laboratory test beds, the generation of failure datasets is limited by the cost, time, and risks associated with inducing faults. This general imbalance in the training dataset hinders the ability of the models to learn effectively, reducing its accuracy in detecting various types of fault. Recent research emphasizes the problem of data scarcity when training ML models for failure management in optical networks [17], [18]. To address this challenge, various data augmentation techniques have been proposed in these works with the aim of enriching the training datasets by generating additional synthetic or transformed data samples. These methods help improve model performance by mitigating class imbalance and improving the representation of rare failure scenarios. Although data augmentation methods can partially alleviate the problem of data scarcity, they fail to capture the full complexity and variability of real-world failure scenarios. Augmented data may lack nuanced patterns and contextual factors present in actual failure events, limiting its effectiveness. Consequently, ML models trained solely on augmented data may face difficulties in generalizing to real-world conditions, potentially reducing their accuracy and reliability in practical implementations. Another method of enriching datasets is using Digital Twins (DTs), which are virtual copies that replicate the structure and behavior of physical systems [19]. In contrast to data augmentation, DTs can leverage physics-based modeling, domain information, and past data to more accurately generate failure events. Using DTs, it becomes possible to generate varied and realistic fault scenarios that might be rare or unpredictable in live networks. DTs and network streaming telemetry provide a flexible and scalable framework for developing and testing machine learning models aimed at managing failures in optical networks [20]–[22]. DTs allow us to replicate normal operations, as well as a broad spectrum of failure events in a controlled and repeatable environment [23], [24].

This research work presents a novel framework that combines DT with a suite of multiple Deep Learning (DL) models to anticipate failures of optical amplifiers in optical networks. Leveraging GNPy (Gaussian Noise model in Python) [25], an open source framework, the DT is designed to mirror the operational characteristics of the network in both nominal and degraded amplifier states. DT enables the systematic generation of synthetic datasets that capture the progression of amplifier deterioration and fault scenarios. The DL models are trained using an extensive dataset produced through the DT. Failure predictions are derived from the results of individual models, enabling a majority vote-based decision mechanism that improves the reliability of the prognostics provided to network operators. The general architecture of the DT-integrated multiple DL suite framework for the prediction of failures in optical networks is illustrated in Fig. 1.

The remainder of this paper is organized as follows. Section II presents the GNPy-based DT framework and explains the methodology used to generate the training dataset used in this study. Section III outlines the suite of DL models utilized for failure prediction, which are trained and tested using the dataset generated as described in the previous section. Section IV evaluates the performance of these models in detail, providing both quantitative metrics and comparative analysis to assess their effectiveness. Section V concludes the paper by summarizing the key findings and highlighting the broader

implications of this work, as well as outlining possible avenues for future research.

II. GNPy – The DT Open Source Framework

A. Digital Twin creation with GNPy

GNPy is an open-source physical-layer engine that estimates end-to-end QoT for coherent Wavelength Division Multiplexing (WDM) systems by explicitly modeling linear and nonlinear impairments along an optical path. From human-readable descriptors, it builds a DT of the network and propagates channel spectra span-by-span, accumulating noise and filtering effects to produce per-channel QoT metrics, e.g. OSNR, GSNR (Generalized Signal to Noise Ratio) at the receiver [25]. GNPy adopts a Gaussian Noise (GN) [26] family model to approximate nonlinear Interference (NLI) generated by Kerr nonlinearity, while accounting for: (i) fiber attenuation and chromatic dispersion, (ii) amplifier Amplified Spontaneous Emission (ASE) noise via Noise Figure (NF) and gain, (iii) cascaded filtering due to Wavelength Selective Switch (WSS)/ Reconfigurable Optical Add-Drop Multiplexers (ROADMs) (including passband narrowing across multiple hops), (iv) per-channel launch power, symbol rate, and spectral shape.

The DT is defined by two primary JSON descriptor files in the GNPy framework:

- **Topology**: This file captures the physical and logical structure of the network. It defines all Network Elements, such as ROADMs, In-Line Amplifiers (ILAs), and transceivers. It also specifies the interconnections between these elements, highlighting how signals cross over the network. The topology descriptor essentially maps the entire infrastructure and serves as the blueprint for the network layout.
- **Equipment library**: This file contains the detailed specifications and performance parameters of all hardware components used in the network. It includes: Fiber characteristics, such as attenuation, chromatic dispersion, and nonlinear coefficients. Amplifier properties, including gain range, NF, and optional attributes such as gain ripple or tilt. WSS/ROADM filter profiles, describing bandwidth, shape, and other filtering behaviors. Transceiver parameters, such as symbol rate, roll-off factor, and Forward Error Correction (FEC) or Bit Error Rate (BER) thresholds.

Optionally, a service/demand file enumerates end-to-end requests (source/destination, bit rate, symbol rate, spacing, spectral slot). The experimental testbed and its corresponding GNPy based virtual topology are shown in Fig. 2 and Fig. 3, respectively. The optically amplified link comprises nine Erbium Doped Fiber Amplifiers (EDFAs) including a booster stage, a pre-amplifier, and seven ILAs separated by eight standard Single Mode Fiber (SMF) spans, each with a length of 65 km. The complete C-band WDM spectrum is launched at the booster input, consisting of 64 carriers spaced at 75 GHz and modulated at 64 GBd. The channels are generated using programmable coherent transceivers equipped with remote configuration and telemetry capabilities through a NETCONF interface. The routing configuration is fixed

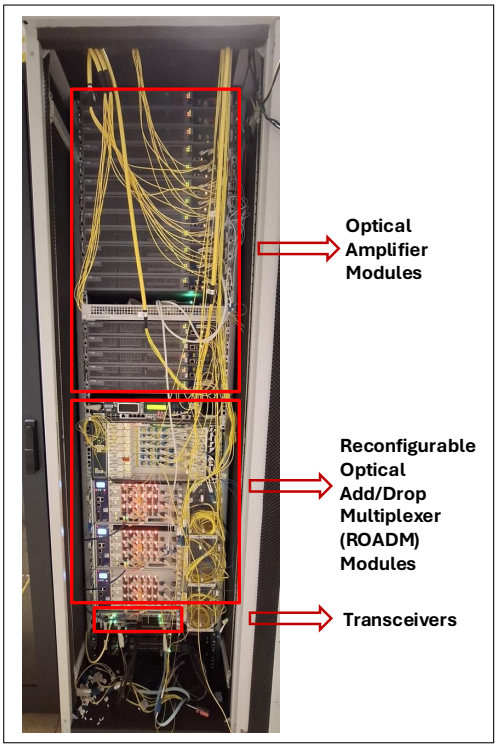

Fig. 2: Experimental testbed network consisting of transceivers, ROADMs, optical amplifiers, fiber spools etc.

throughout the analysis, with the primary focus placed on the evolution of the physical layer parameters in response to amplifier degradations. The experimental testbed is replicated by constructing an equivalent network topology and assigning identical parameters to all network elements, thereby generating a DT of the real transmission setup. The interactions between the real network and the digital twin occur primarily during the parameter mapping and validation stages. The configuration of the real network such as EDFA operating points, span lengths, fiber characteristics, and spectrum features are mapped to ensure that the digital twin accurately represents real operational conditions. Given a routed light path and its channel plan, evaluation proceeds element-by-element:

- **Transmitter**: Instantiate per-channel spectra from symbol rate and roll-off; set launch powers; initialize noise budget.
- **Fiber span**: Apply deterministic attenuation and dispersion; compute the NLI span (self- and cross-channel) via the GN kernel using the span parameters and channel grid.
- **ILA**: Restore power budget with configured target gain (typically compensating span loss) and add ASE noise per NF; optional gain tilt/ripple may be applied.
- **ROADM/WSS**: Enforce add/drop and pass-through filtering; track cumulative filter narrowing across nodes; and apply to spectra.
- **Receiver**: Integrate total noise in the reference/matched bandwidth and report OSNR/GSNR (and margins) per channel.

The loop runs hop-by-hop for all spans/ILAs/ROADMs;

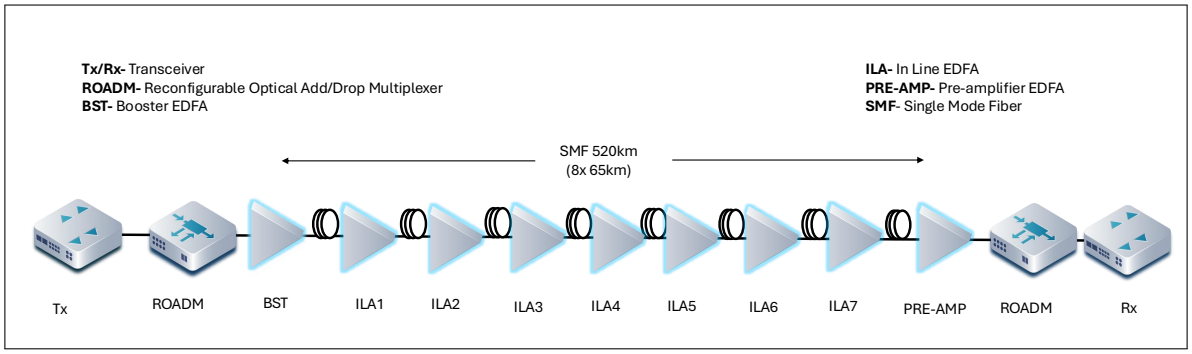

Fig. 3: Virtual topology in DT representing the experimental testbed for optical amplifier failure prediction.

multi-channel WDM interactions are handled consistently when computing GN contributions. The main outputs are: (i) OSNR in a standard reference bandwidth (e.g., 0.1 nm), (ii) GSNR, aggregating ASE and NLI in the receiver bandwidth, and (iii) margins versus transceiver thresholds, optionally mapped to outage. These enable the ranking of operating points and early detection of QoT degradation as the physical parameters drift.

B. Optical Amplifier Failures and Database Generation

Optical amplifiers are integral to enabling high-capacity, long-haul optical communication systems by enhancing signal strength directly in the optical domain, thereby eliminating the need for Optical-Electrical-Optical (OEO) conversion. However, the reliability and performance of these amplifiers are critical as any deterioration can have a direct impact on the OSNR, ultimately compromising the overall QoT across the network. Several degradation mechanisms and failure modes can affect the operational efficiency of optical amplifiers, which are discussed below [1]:

- **Amplifier Gain Degradation:** Over time, the amplification capability of optical amplifiers can degrade due to various physical and operational factors. A primary cause is the gradual decline in the output of the pump lasers, a result of component aging. Additionally, excessive input power can lead to gain medium saturation, causing the amplifier to no longer be able to maintain a linear gain response. These conditions collectively reduce effective amplification, leading to weak signal transmission over long distances.
- **Increased Noise Contribution:** A significant challenge in the operation of optical amplifiers is the accumulation of optical noise, particularly noise originating from spontaneous emission processes. Over time, this noise, known as ASE [27], can intensify due to several underlying causes. These include thermal instability, deterioration or aging of the pump laser source, and inefficiencies in the energy transfer process within the gain medium. Additionally, spectral drift or wavelength misalignment of the pump laser can lead to suboptimal excitation of the active medium, further exacerbating the noise output. As the level of noise increases relative to the signal, the OSNR degrades, impairing the clarity and reliability of the transmitted data across the optical link.
- **Irregular Output Signal Levels:** Another common failure scenario involves inconsistency in the output power of the amplifier. Such anomalies manifesting as sudden power surges or unexpected drops may stem from erratic pump laser current levels or abrupt pump diode malfunctions. These instabilities can disrupt the optical power budget, impair the functioning of downstream network elements, and lead to transient or persistent quality issues within the transmission path.

The persistent occurrence of such impairments in optical amplifiers can lead to severe system-level disruptions, potentially causing service outages or significant degradation in network performance. Given the critical role of amplifiers in maintaining signal integrity and ensuring transmission quality over extended distances in an Optical Line System (OLS), it becomes imperative to adopt robust monitoring frameworks. By enabling detailed modeling of physical-layer impairments and signal degradation across the network, the GNPy-based DT framework facilitates a more accurate estimation of system limitations and helps identify optimal design configurations.

To construct the training dataset, GNPy was employed to perform a series of iterations in which the input network topology and relevant operational parameters were systematically varied. This methodology allowed the generation of diverse network scenarios that covered a wide range of configurations and performance conditions. By altering parameters such as amplifier gain levels, transponder configurations, and fault conditions, GNPy facilitated the emulation of a broad spectrum of network behaviors and corresponding signal quality outcomes. Each iteration yielded detailed QoT metrics, including GSNR and OSNR, which were systematically incorporated into the training dataset. To generate progressive, undetectable gain degradation with power monitors, a case of soft-failure, a controlled reduction in amplifier gain was introduced incrementally over time. For the amplifier i, gain at time t be denoted by $G_i(t)$. The gain loss is coupled with an output power hold via a Variable Optical Attenuator (VOA) control logic in the DT, so that the span output power remains nominal while QoT (OSNR/GSNR) silently degrades. Initially, internal VOA attenuation of the amplifier $\alpha_{\mathrm{VOA},i}(t)$ is set to a high attenuation value $\alpha_{\mathrm{VOA},i}(0)$; as the imposed gain reduction

$\Delta G_i(t) = G_{i,\text{nominal}} - G_i(t)$ increases, for the compensation of gain loss, the VOA attenuation is gradually reduced :

$$\alpha_{\text{VOA},i}(t) \approx \alpha_{\text{VOA},i}(0) - \Delta G_i(t), \tag{1}$$

The ASE noise generated in amplifier i at time t is given by

$$P_{\text{ASE},i}(t) = 2n_{\text{sp},i}(t)\left(G_i(t) - 1\right) h\nu\,\Delta f, \tag{2}$$

where $n_{\text{sp},i}(t)$ is the spontaneous emission factor, h is Planck's constant, ν is the optical carrier frequency, $G_i(t)$ is the gain in linear scale and Δf is the ASE bandwidth. After the VOA, the output ASE power becomes

$$P_{\text{ASE,out},i}(t) = P_{\text{ASE},i}(t)\, 10^{-\alpha_{\text{VOA},i}(t)/10}. \tag{3}$$

The open up of VOA (i.e, decrease in $\alpha_{\text{VOA},i}(t)$) increases the propagation of ASE noise. In addition to ASE, NLI accumulates along the link. The total noise at the receiver is therefore

$$P_{\text{noise}}(t) = P_{\text{ASE,total}}(t) + P_{\text{NLI}}(t), \tag{4}$$

and the GSNR evolves as

$$\text{GSNR}(t) = \frac{P_{\text{signal}}}{P_{\text{ASE,total}}(t) + P_{\text{NLI}}(t)}. \tag{5}$$

The table I provides a detailed description of the fault levels, categorizing them according to the extent of performance reduction observed in the amplifier. The soft-failures of the amplifier were classified into three fault classes based on the extent of gain reduction: deviations up to 2 dB were considered normal (no fault), reductions between 2 and 3 dB were labeled as minor faults, and any degradation exceeding 3 dB was classified as a major fault. The fault thresholds were chosen to represent soft failures up to 3 dB, following link budget engineering practices in which each span includes over 3 dB of VOA applied attenuation during testing to guarantee operational margin [28]. To ensure continuity of service and accurately model soft-failure scenarios, the extent of degradation is limited by the VOA compensation range. For this reason, degradation in ILA6 and ILA7 is limited up to 3 dB. Surpassing this compensation threshold would risk triggering a hard failure, thereby falling outside the intended scope of soft-failure prediction and localization.

TABLE I: Amplifier Fault Levels Based on Gain Degradation

Fault level	Gain degradation(dB)	Severity
0	< 2	No fault
1	2-3	Minor fault
2	> 3	Major fault

The effect of faults on transmission performance is evaluated by monitoring the evolution of key signal parameters as degradation progresses. To realistically emulate gradual deterioration observed in operational networks, the gain reduction is applied in randomly selected incremental steps, allowing for non-uniform accumulation of impairment over time. This stochastic variation in the size of the steps over different time intervals captures the inherent variability present in live network conditions. This yield a multivariate time-series dataset (GSNR, OSNR, effective gain, etc.) for nominal and degraded states.

The resulting time-series dataset, used to train the ML model, reflects temporal fluctuations in GSNR as a function of amplifier-specific degradation. Each amplifier instance in the dataset is characterized by unique gain and tilt profiles, enabling the model to learn and generalize from a diverse range of degradation patterns and signal quality impacts across different operating conditions. This approach ensures a more comprehensive and representative dataset for the training of ML models and overcomes the limitation of obtaining a larger failure dataset from the laboratory experimental test-beds. The dataset obtained provides the basis for the training and evaluation of ML models aimed at predicting optical amplifier failures. Leveraging the flexibility of GNPy, diverse network conditions including various fault levels, amplifier behaviors, and configuration settings, are modeled to closely approximate real-world dynamics.

III. Deep Learning Architectures for Amplifier Fault Prediction

The dataset generated through the DT in the GNPy framework as elaborated in Section II provides a time-series record of the transmission parameters under normal and different fault conditions of the amplifiers. The fault levels were determined on the basis of the degree of amplifier gain degradation, as outlined in the previous section. The models are trained to predict and localize soft-failures by identifying both the faulty amplifier and the corresponding fault severity level. The proposed models estimate the current fault state for instance fault level 0, indicating no fault and can detect elevated fault levels (level 1, 2) representing higher gain degradations, which serve as early indicators of an impending failure. Although the model effectively characterizes fault conditions, it does not yet provide absolute long term forecasting of when a failure will occur. Each input data instance corresponds to a fixed-length telemetry window containing multiple features such as GSNR, OSNR, and amplifier gain. The objective is to learn a mapping

$$f : \mathbf{X} \in \mathbb{R}^{L\times F} \quad \longrightarrow \quad y \in \{0, 1, 2\}, \tag{6}$$

where $\mathbf{X}$ represents the input sequence of length L with feature dimension F, and y denotes the amplifier fault class. To address this classification task, three deep learning architectures, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and Long- and Short-Term Time-Series Network (LSTNet) are employed. All models receive the same feature streams as inputs within each telemetry segment and output the predicted fault class through a softmax layer. Each architecture employs model and training hyperparameters tailored to its structural characteristics for example, progressively decreasing recurrent units in the LSTM, multi-stage convolutional filter banks in the CNN, and a combined convolution–recurrent structure in the LSTNet model. These configurations, which specify the layer dimensions, kernel sizes, learning rate, batch size, dropout rate, and the maximum

number of training epochs (ranging from 100 to 200 depending on the architecture), are kept fixed across all experiments to ensure methodological consistency and reproducibility while allowing each model to operate under an appropriately tuned setup. Model training follows a standardized procedure using the sparse categorical cross-entropy loss function and the Adam optimizer, with the dataset partitioned into 70% training, 10% validation, and 20% testing. To mitigate overfitting, we have used early stopping with restoration of the best validation weights, dropout regularization in all hidden layers, and a learning rate scheduler that progressively reduces the learning rate after a predefined epoch threshold.

A. Convolutional Neural Network (CNN)

Convolutional neural networks are designed to capture local temporal dependencies in sequential data [29]. For an amplifier telemetry window, a convolutional filter of size k slides along the time axis to compute

$$h_t = \sigma\left(\sum_{i=0}^{k-1} W_i \cdot \mathbf{X}_{t+i} + b\right), \tag{7}$$

where W_i are learnable filter weights, b is a bias term, and $\sigma(\cdot)$ is a nonlinear activation function. This operation enables the CNN to extract short-term patterns such as sudden fluctuations in channel power or abrupt gain changes. Pooling layers further condense the learned features, and dense layers project them into the three defined fault categories. Owing to its parallel structure, CNN provides low inference latency and is suitable for real-time monitoring.

B. Long Short-Term Memory (LSTM)

Long short-term memory networks extend recurrent neural networks by introducing gated memory cells to capture long-term dependencies. At each time step t, the hidden state h_t and cell state c_t are updated according to

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{8}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{9}$$
$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c), \tag{10}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{11}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad h_t = o_t \odot \tanh(c_t). \tag{12}$$

where f_t, i_t, o_t denote the forget, input, and output gates, respectively, and c_t represents the memory cell state, following the standard LSTM formulation. The equations have been further restated and formalized in literature [30]–[32], where their application to time-series modeling and fault prediction is widely discussed. This makes the model effective for identifying gradual amplifier degradations and long-term variations in telemetry sequences that span multiple time steps.

C. Long- and Short-Term Time-Series Network (LSTNet)

The long and short term time series network combines convolutional, recurrent, and linear components to jointly capture multiple temporal scales. The convolutional block extracts short-term features, while the recurrent block, typically implemented using Gated Recurrent Units (GRUs), models longer dependencies. In addition, LSTNet includes a temporal highway component that preserves linear trends through an autoregressive mapping:

$$\hat{y}_t = \mathbf{w}^\top \mathbf{X}_{t-p:t} + b, \tag{13}$$

where $\mathbf{X}_{t-p:t}$ is the most recent input segment of length p, and $\hat{y}_t$ represents the autoregressive output, following the formulation of the LSTNet model [33]. This AR component has been widely adopted in subsequent works as a temporal highway mechanism to capture linear trends in multivariate time series [34]. This hybrid design enables LSTNet to detect sudden anomalies, track sequential dependencies, and capture slow, trend-like degradations simultaneously. Such capability makes it particularly suitable for amplifier telemetry, where faults may manifest as either abrupt events or gradual drifts.

Together, these architectures provide complementary perspectives on amplifier telemetry, and their effectiveness in classifying the defined fault categories is quantitatively evaluated in Section IV.

IV. Deep Learning Models: Performance Analysis

As outlined in the preceding section, this study employs three distinct DL architectures, CNN, LSTM, and LSTNet to predict faults in optical amplifiers. These models are trained to detect and classify incoming faults by analyzing the gain characteristics of the amplifiers over time. Specifically, the classification is based on the level of gain degradation, which serves as a direct indicator of the operational health of the amplifier. To enable fault categorization, gain degradation is segmented into three distinct classes, as summarized in Table I. DL models are trained using labeled datasets representing three fault levels, with the objective of learning temporal and spectral features relevant to fault prediction and classification. The CNN model is utilized for its strength in extracting spatial or local features from time-series representations of the signal. The LSTM network is used to capture both short and long term temporal dependencies in the sequential data, making it particularly suitable for tracking gradual degradation patterns over time. LSTNet, which combines convolutional layers with recurrent structures and a temporal attention mechanism, is leveraged to integrate both local and global temporal dynamics. By utilizing these architectures, the study aims to achieve a robust and accurate prediction of amplifier faults across a wide range of degradation scenarios.

The classification performance of the DL models is assessed using standard evaluation metrics such as accuracy, precision, recall, and the F1-score. In classification problems, each prediction falls into one of four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These outcomes form the basis for computing the performance metrics summarized in Table II.

Accuracy measures the overall proportion of correctly classified instances and serves as a general indicator of model performance. Precision quantifies the proportion of correct positive predictions, indicating how well the model avoids false alarms. Recall, or sensitivity, measures the ability of the

TABLE II: Classification Performance Metrics Used for Model Evaluation

Metric	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
F1-Score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$

model to detect actual positive cases, reflecting its effectiveness in capturing faults. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced metric that accounts for both types of misclassification and is particularly useful when the costs of false positives and false negatives are significant.

A bar graph comparing the accuracy of CNN, LSTM, and LSTNet models across the training, validation, and test sets is presented in Fig. 4. The results highlight the strong performance of all three models, with each achieving perfect accuracy (100%) on the training set, indicating successful learning of the underlying patterns associated with amplifier gain degradation. In the validation phase, the LSTM model continues to demonstrate superior generalization with an accuracy of 99%, followed closely by LSTNet at 97% and CNN at 96%. These results suggest that the LSTM model is particularly effective in capturing temporal dependencies in the data, which may contribute to its improved performance. A similar trend is observed in the test set, where the LSTM model again achieves the highest accuracy at 99%, while LSTNet model improves to 98% and CNN model shows consistent performance with 96%. This consistency across validation and test sets indicates that all models generalize well, with the LSTM showing the most robust performance. In general, comparative analysis shows the effectiveness of deep learning approaches in accurately identifying patterns in the dataset for failure prediction in optical amplifiers.

Figures 5, 6, and 7 present heatmaps that illustrate key performance metrics, precision, recall, and the F1 score for the CNN, LSTM, and LSTNet models, respectively. These

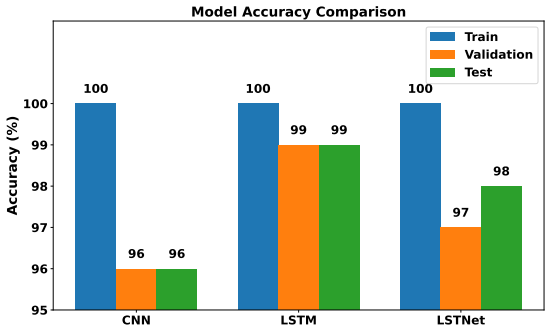

Fig. 4: Bar graph displaying accuracy comparison of CNN, LSTM and LSTNet models

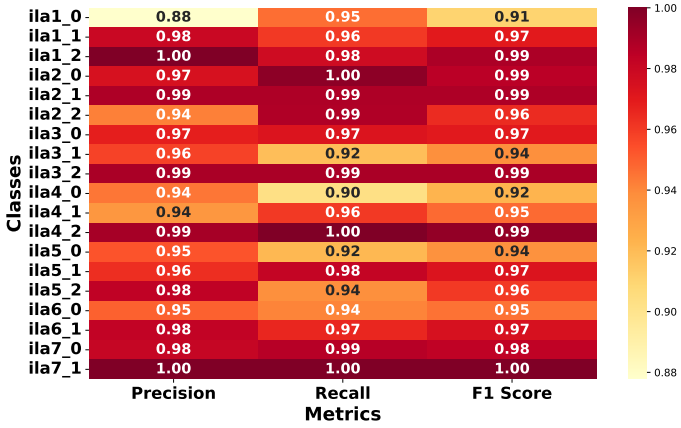

Fig. 5: Heatmap displaying Precision, Recall and F1-Score across all Classes predicted by the CNN model, illustrating class-wise classification performance.

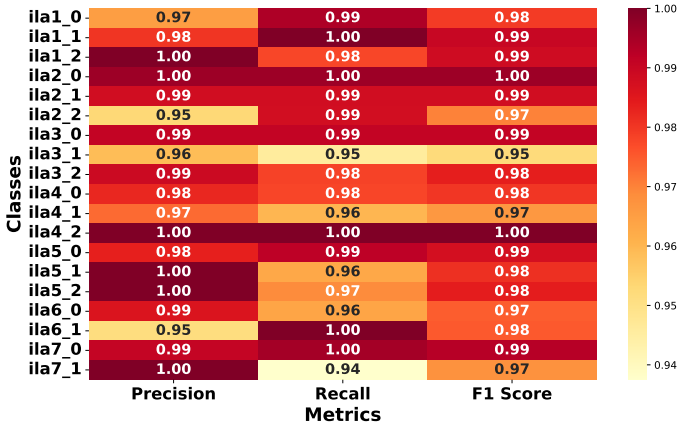

Fig. 6: Heatmap displaying Precision, Recall and F1-Score across all Classes predicted by the LSTM model, illustrating class-wise classification performance.

metrics are used to evaluate the capability of the models in identifying faulty amplifiers and the corresponding fault levels, as categorized in Table I. Each class label is constructed by combining the unique identifier of the amplifier and its associated fault level, in the format ila$x$_y, where x denotes the amplifier number and y represents the fault level. In the heatmaps, darker shades correspond to higher metric values, indicating better model performance. Across all three models, class-wise precision, recall, and F1 scores consistently exceed 0.95 for most amplifier fault combinations. To evaluate how well each model discriminates between the various fault severity levels, we further describe their classification performance. For severity level 0 (no fault condition), all models achieve strong results, though the CNN shows a small deviation that remains within acceptable performance levels (e.g., ila1_0 shows $\approx$ 0.88 precision, ila4_0 achieves $\approx$ 0.90 recall, and ila1_0 records $\approx$ 0.91 F1 score). In contrast, both LSTM and LSTNet remain highly stable, frequently attaining values close to 1.00 across all metrics. This consistency demonstrates their ability to reliably identify normal amplifier states, thereby

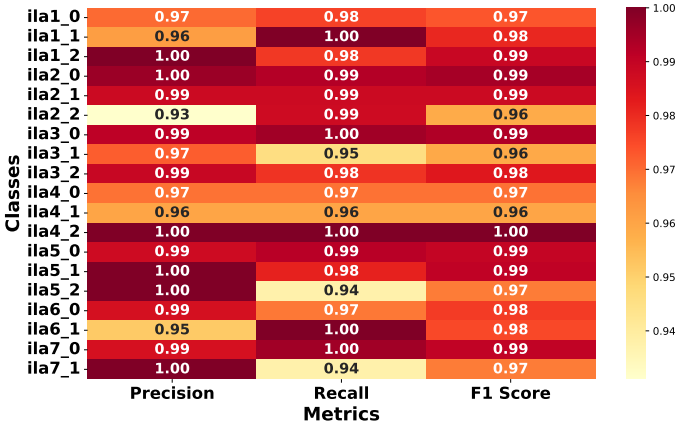

Fig. 7: Heatmap displaying Precision, Recall and F1-Score across all classes predicted by the LSTNet model, illustrating class-wise classification performance.

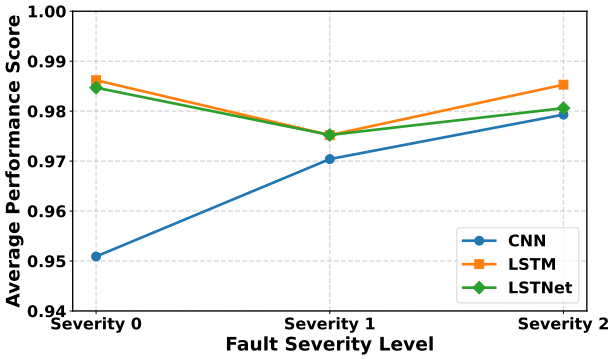

Fig. 8: Average performance score Vs Fault severity levels across all models).

reducing the likelihood of false alarms. Model performance becomes increasingly robust with higher levels of impairment. For severity level 1 (minor faults), all three architectures perform notably well. The CNN achieves precision ≥ 0.94, recall ≥ 0.92, and F1 ≥ 0.94. The LSTM maintains similarly high performance, with precision and F1 ≥ 0.95 and recall ≥ 0.94. LSTNet continues to excel, reporting precision ≥ 0.95, recall ≥ 0.94, and F1 ≥ 0.96. These results illustrate that all the models particularly LSTM and LSTNet are highly capable of identifying subtle fault signatures. At severity level 2 (major faults), all models reach near perfect performance, with precision, recall, and F1 scores approaching 1.00. This indicates that all the models are performing strongly for the major fault which is very important. Fig. 8 depicts the combined average performance score computed from precision, recall, and F1 score for all models across different fault severity levels. Overall, the comparison across severity levels shows that although all models demonstrate strong performance, LSTM and LSTNet consistently deliver the most robust and reliable results. Their superior handling of both minor and major faults underscores the advantages of temporal sequence modeling over purely convolutional approaches for accurately distinguishing amplifier fault severity.

Following the comparison of standard performance metrics,

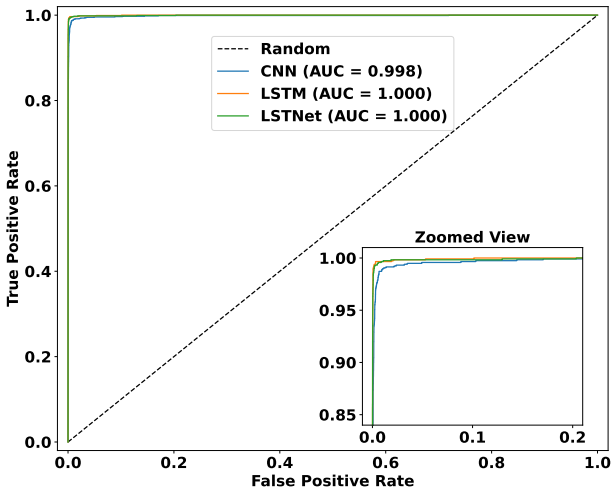

Fig. 9: ROC Curves illustrating the trade-off between True Positive and False Positive rates, and AUC metrics providing a quantitative measure of overall model performance across CNN, LSTM and LSTNet models.

the Receiver Operating Characteristic (ROC) curve shown in Fig. 9. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds, effectively illustrating the trade-off between correctly identifying faults and incorrectly flagging normal instances. These thresholds refer to the cutoff values used to convert the predicted probabilities of the model into class labels. By systematically varying the threshold from 0 to 1, the ROC curve captures how the performance of the model changes, revealing how more or fewer instances are classified as positive. A model that makes random predictions would trace a diagonal line (shown as a dotted line) from the bottom left corner to the top right corner of the plot, yielding an Area Under the Curve (AUC) of 0.5. The AUC metric, which represents the integral of the ROC curve, quantifies the ability of the model to rank positive instances higher than negative ones. An AUC of 1.0 indicates perfect separability, whereas an AUC of 0.5 implies no discriminative power. In this study, the LSTM and LSTNet models achieved an ideal AUC of 1.0, demonstrating a flawless distinction between faulty and normal amplifier behaviors across all thresholds. The CNN model followed closely with an AUC of 0.998, indicating high classification performance, although slightly less robust than the recurrent models.

To gain a more comprehensive understanding of the overall effectiveness of the models, it is essential to visualize the trade-offs between the predictive performance metrics discussed above and the computational efficiency. Although these metrics capture how well a model distinguishes between classes, they do not reflect the time and resources required to train and deploy the model. In real-world applications, especially those involving large-scale deployments or resource-limited environments, factors like training time and inference speed become an equally important consideration. To capture this balance, a radar plot is shown in Fig.10 to compare each model across four axes: accuracy, F1 score, training time and

inference time. This allows for a multifaceted performance evaluation across accuracy, generalization, and computational cost.

The training time is quantified using a normalized metric based on the observed range of training times across all models. The training time, T_{train}, is normalized as:

$$T_{\text{train}} = \frac{t_{\text{train}}}{t_{\max}} \tag{14}$$

where:

- t_{train} is the training time of a given model,
- $t_{\max}$ is the maximum training time observed across all models.

Similarly, inference time, T_{inf}, is normalized to evaluate the prediction speed of the model. A lower inference time implies better efficiency. The metric is normalized as:

$$T_{\text{inf}} = \frac{t_{\text{infer}}}{t_{\max}} \tag{15}$$

where:

- t_{infer} is the average inference time per sample for a given model,
- $t_{\max}$ is the maximum inference time observed across all models.

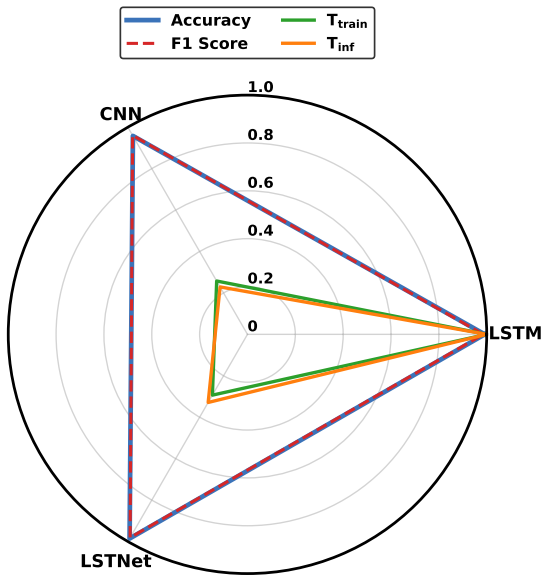

Fig. 10: Radar plot visualizing accuracy, F1 score, training time, T_{train} and inference time T_{inf} to provide a multidimensional comparison of model performance across CNN, LSTM and LSTNet models.

Both T_{train} and T_{inf} range from 0 to 1, where 0 indicates the most efficient model and 1 indicates the least efficient among the compared models. Intermediate values are linearly scaled based on their relative position in the range. As observed in the radar plot, the CNN model achieved the best T_{train} and T_{inf} score of 0.26 and 0.23 respectively, followed by LSTNet with scores of 0.29 and 0.33 for T_{train} and T_{inf} respectively, while LSTM scored 1 for both due to its relatively longer training and inference duration. However, the LSTM model exhibits the highest classification performance with 99% accuracy and an F1 score of 0.98, closely followed by the LSTNet model at 98% accuracy and F1 score of 0.98. The CNN model, while slightly less accurate (96% accuracy, 0.96 F1 score), trained significantly faster and provided inference with far fewer computational demands.

This comparative analysis highlights the distinct strengths of each model. The LSTNet model achieves a balanced trade-off among accuracy, generalization, and computational efficiency, making it a practical choice for network operators with constrained resources while maintaining performance requirements. The LSTM model delivers the highest prediction accuracy, making it ideal for scenarios where precision is paramount and computational resources are not a limiting factor. In contrast, despite its slightly lower predictive accuracy, the CNN model exhibiting the highest computational efficiency among the evaluated models remains suitable for scenarios requiring rapid training and lightweight deployment, such as localized ML implementations on controller cards in the chassis hosting amplifier devices, in lieu of centralized server-based systems. By integrating training efficiency into the performance evaluation, this analysis provides a more realistic foundation for selecting models based on deployment requirements. However in this work, we employ majority vote-based decision mechanism, since high prediction accuracy is fundamental to the reliable deployment of ML models for failure management especially in low power margin optical networks provided the availability of computational resources.

V. Conclusions and Future Work

This study presents an integrated and scalable framework for predictive maintenance in optical networks, which combines DTs with the predictive capabilities of DL models. Using GNPy, an open-source framework, a DT was developed to emulate behavior of the amplifier under both normal and faulty conditions. This enabled the generation of synthetic datasets that capture QoT degradation and amplifier failure patterns, which were then used to train CNN, LSTM, and LSTNet models for fault prediction. DL models predict the degradation of individual amplifiers by classifying the amplifier states into three fault levels, normal, minor, and major, allowing for precise localization of amplifier soft-failures in optical networks. In addition, repeated prediction of a major fault in a specific amplifier can act as an early warning signal for a potential hard failure. Each model demonstrates distinct strengths in terms of accuracy, generalization, and computational efficiency. The output of these individual models facilitates a majority voting mechanism, improving the overall reliability and robustness of the fault predictions provided to network operators. In general, the proposed DT-DL framework enables proactive fault management by supporting early failure detection and prediction, thereby minimizing network downtime and enhancing the resilience and operational intelligence of optical networks. Future work can be extended to additional network topologies such as mesh networks and with varying channel occupancies. This will enable deploying the proposed framework in a live network monitoring environment.

ACKNOWLEDGMENTS

This publication has received funding and support from the project PNRR-NGEU (MUR–DM117/2023) and from the European Union's Horizon Europe research and innovation programme under projects ALLEGRO (GA: 101092766) and SENSEI (GA: 101189545).

REFERENCES

[1] Sergio Cruzes. Failure management overview in optical networks. *IEEE Access*, 12:169170–169193, 2024.

[2] Danshi Wang, Chunyu Zhang, Wenbin Chen, Hui Yang, Min Zhang, and Alan Pak Tao Lau. A review of machine learning-based failure management in optical networks. *Science China Information Sciences*, 65(11):211302, Oct 2022.

[3] Danish Rafique and Luis Velasco. Machine learning for network automation: overview, architecture, and applications [invited tutorial]. *Journal of Optical Communications and Networking*, 10(10):D126–D143, 2018.

[4] Francesco Musumeci, Cristina Rottondi, Giorgio Corani, Shahin Shahkarami, Filippo Cugini, and Massimo Tornatore. A tutorial on machine learning for failure management in optical networks. *Journal of Lightwave Technology*, 37(16):4125–4139, 2019.

[5] Behnam Shariati, Marc Ruiz, Jaume Comellas, and Luis Velasco. Learning from the optical spectrum: Failure detection and identification. *Journal of Lightwave Technology*, 37(2):433–440, 2019.

[6] Camille Delezoide, Petros Ramantanis, and Patricia Layec. Streamlined failure localization method and application to network health monitoring. *Journal of Lightwave Technology*, 41(19):6119–6125, 2023.

[7] Jatin Babbar, Ahmed Triki, Reda Ayassi, and Maxime Laye. Machine learning models for alarm classification and failure localization in optical transport networks. *Journal of Optical Communications and Networking*, 14(8):621–628, 2022.

[8] Josh W. Nevin, Sam Nallaperuma, Nikita A. Shevchenko, Xiang Li, Md. Saifuddin Faruk, and Seb J. Savory. Machine learning for optical fiber communication systems: An introduction and overview. *APL Photonics*, 6(12):121101, 12 2021.

[9] Faisal Nadeem Khan, Qirui Fan, Chao Lu, and Alan Pak Tao Lau. An optical communication's perspective on machine learning and its applications. *Journal of Lightwave Technology*, 37(2):493–516, 2019.

[10] Danshi Wang, Yidi Wang, Xiaotian Jiang, Yao Zhang, Yue Pang, and Min Zhang. When large language models meet optical networks: Paving the way for automation. *Electronics*, 13(13), 2024.

[11] Lars E. Kruse, Sebastian Kühl, Annika Dochhan, and Stephan Pachnicke. Experimental validation of machine learning-based joint failure management and quality of transmission estimation. *IEEE Photonics Journal*, 15(6):1–9, 2023.

[12] Shahin Shahkarami, Francesco Musumeci, Filippo Cugini, and Massimo Tornatore. Machine-learning-based soft-failure detection and identification in optical networks. In *2018 Optical Fiber Communications Conference and Exposition (OFC)*, pages 1–3, 2018.

[13] A. P. Vela, B. Shariati, M. Ruiz, F. Cugini, A. Castro, H. Lu, R. Proietti, J. Comellas, P. Castoldi, S. J. B. Yoo, and L. Velasco. Soft failure localization during commissioning testing and lightpath operation. *Journal of Optical Communications and Networking*, 10(1):A27–A36, 2018.

[14] Moisés Felipe Silva, Alessandro Pacini, Andrea Sgambelluri, and Luca Valcarenghi. Learning long- and short-term temporal patterns for ml-driven fault management in optical communication networks. *IEEE Transactions on Network and Service Management*, 19(3):2195–2206, 2022.

[15] Huazhi Lun, Mengfan Fu, Yihao Zhang, Hexun Jiang, Lilin Yi, Weisheng Hu, and Qunbi Zhuge. A gan based soft failure detection and identification framework for long-haul coherent optical communication systems. *Journal of Lightwave Technology*, 41(8):2312–2322, 2023.

[16] Mashboob Cheruvakkadu Mohamed, Muhammad Umar Masood, Imran Chowdhury Dipto, Renato Ambrosone, Gulmina Malik, Stefano Straullu, Sai Kishore Bhyri, Gabriele Maria Galimberti, João Pedro, Antonio Napoli, Walid Wakim, and Vittorio Curri. Network data based transfer learning failure prediction agent pre-trained using digital twin. In *2025 International Conference on Optical Network Design and Modeling (ONDM)*, pages 1–6, 2025.

[17] Lareb Zar Khan, João Pedro, Nelson Costa, Andrea Sgambelluri, Antonio Napoli, and Nicola Sambo. Model and data-centric machine learning algorithms to address data scarcity for failure identification. *J. Opt. Commun. Netw.*, 16(3):369–381, Mar 2024.

[18] Cheng Xing, Chunyu Zhang, Bing Ye, Danshi Wang, Yinqiu Jia, Jin Li, and Min Zhang. Failure data augmentation for optical network equipment using time-series generative adversarial networks. In *Optical Fiber Communication Conference (OFC) 2023*, page M3G.4. Optica Publishing Group, 2023.

[19] John Vickers. Digital twins in a nutshell: A new era in engineering and manufacturing. Presentation, doe–nsf workshop on digital twins for manufacturing, NASA, Marshall Space Flight Center, Storrs, CT, November 11 2024. NASA NTRS Document ID 20240013986.

[20] Kayol S. Mayer, Jonathan A. Soares, Rossano P. Pinto, Christian E. Rothenberg, Dalton S. Arantes, and Darli A. A. Mello. Soft failure localization using machine learning with sdn-based network-wide telemetry. In *2020 European Conference on Optical Communications (ECOC)*, pages 1–4, 2020.

[21] Kayol S. Mayer, Jonathan A. Soares, Rossano P. Pinto, Christian E. Rothenberg, Dalton S. Arantes, and Darli A. A. Mello. Machine-learning-based soft-failure localization with partial software-defined networking telemetry. *J. Opt. Commun. Netw.*, 13(10):E122–E131, Oct 2021.

[22] Kayol S. Mayer, Rossano P. Pinto, Jonathan A. Soares, Dalton S. Arantes, Christian E. Rothenberg, Vinicius Cavalcante, Leonardo L. Santos, Filipe D. Moraes, and Darli A. A. Mello. Demonstration of ml-assisted soft-failure localization based on network digital twins. *J. Lightwave Technol.*, 40(14):4514–4520, Jul 2022.

[23] Mashboob Cheruvakkadu Mohamed, Muhammad Umar Masood, Renato Ambrosone, Gulmina Malik, Rocco D'Ingillo, Stefano Straullu, Sai Kishore Bhyri, Gabriele Maria Galimberti, João Pedro, Antonio Napoli, Walid Wakim, and Vittorio Curri. Machine learning agents leveraging digital twins for failure prediction in optical networks. In *2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, pages 1–6, 2025.

[24] Mashboob Cheruvakkadu Mohamed, Renato Ambrosone, Muhammad Umar Masood, Gulmina Malik, Stefano Straullu, Sai Kishore Bhyri, João Pedro, Antonio Napoli, Gabriele Maria Galimberti, Walid Wakim, and Vittorio Curri. Digital twin-integrated binary classifier ml model for edfa failure prediction. In *2025 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6, 2025.

[25] Vittorio Curri. Gnpy model of the physical layer for open and disaggregated optical networking [invited]. *Journal of Optical Communications and Networking*, 14(6):C92–C104, 2022.

[26] Andrea Carena, Vittorio Curri, Gabriella Bosco, Pierluigi Poggiolini, and F Forghieri. Modeling of the impact of nonlinear propagation effects in uncompensated optical coherent transmission links. *Journal of Lightwave technology*, 30(10):1524–1539, 2012.

[27] Alexander William Setiawan Putra, Minoru Yamada, Hiroyuki Tsuda, and Sumiaty Ambran. Theoretical analysis of noise in erbium doped fiber amplifier. *IEEE Journal of Quantum Electronics*, 53(4):1–8, 2017.

[28] Anxu Zhang, Junjie Li, Lipeng Feng, Kai Lv, Fei Yan, Yusen Yang, Haiqiang Wang, Qifang Yang, Lingquan Wang, Xiaolei Zhang, Shibao Ding, Min Liao, Yi Yu, and Liangchuan Li. Field trial of 24-Tb/s (60×400 Gb/s) DWDM transmission over a 1910-km G.654.E fiber link with 6-THz-bandwidth C-band EDFAs,. *Opt. Express*, 29(26):43811–43818, Dec 2021.

[29] Zhen Zuo, Bing Shuai, Gang Wang, Xiao Liu, Xingxing Wang, Bing Wang, and Yushi Chen. Learning contextual dependence with convolutional hierarchical recurrent neural networks. *IEEE Transactions on Image Processing*, 25(7):2983–2996, 2016.

[30] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[31] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[32] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.

[33] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.

[34] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.