Physics-Aware Deep Learning for QoT Safety in Elastic Optical Networks: Conformal Prediction with Explainable Integrable Gradient

DEVENDRA SRIVASTAVA, SHWETA TRIPATHI, PRIYANKA GAUTAM, APURVA KUMARI

Dr. Ambedkar Institute of Technology for Divyangjan, India
Orcid ID:0009-0009-1184-8887
**shweta@aith.ac.in*

Abstract- *Elastic optical networks (EONs) are susceptible to soft failures gradual degradations such as filter passband tightening and centre-frequency drift that undermine quality of transmission (QoT) without triggering hard alarms. We present a predict calibrate guarantee explain framework that converts routine, topology-aware path link telemetry into operator ready QoT safety decisions. Time-aligned sequences are labelled via physics-aware constraints and modelled with compact GRU/LSTM classifiers. Probabilities are temperature-scaled for enhanced reliability, conformal prediction yields coverage-controlled "predict or abstain" decisions, and explainable AI (Integrated Gradients/SHAP) supplies time×feature attributions for operator insight. Experimental validation on a GEANT-like dataset demonstrates strong performance: the GRU achieves AUC 0.985 (LSTM 0.973) with well-calibrated probabilities. Risk coverage curves expose actionable operating points at 95% coverage, selective risk is approximately 3-5% while explanations consistently highlight filter-bandwidth scale, utilization, and optical signal-to-noise ratio (OSNR) as primary performance drivers. The pipeline delivers trustworthy QoT prediction with calibrated probabilities, formal coverage guarantees, and physics-consistent explanations, providing an SDN-ready foundation for proactive maintenance and resilient operation in next-generation EONs.*

Keywords- Elastic optical networks (EONs), GRU, LSTM, Conformal prediction, Explainable AI (XAI), Software Defined Networking (SDN).

1. Introduction

Modern telecommunications infrastructure increasingly relies on Elastic Optical Networks (EONs), which have revolutionized how network operators manage spectrum allocation and capacity scaling in response to growing demands from cloud computing, 5G services, and data-intensive applications [1, 6]. The fundamental advantage of EONs lies in their ability to dynamically allocate spectral resources through dense wavelength division multiplexing (DWDM) and reconfigurable optical add-drop multiplexer (ROADM) technologies. However, this enhanced flexibility introduces significant operational challenges, including expanded configuration spaces, reduced safety margins, and increased reconfiguration frequency, all of which complicate network management and quality assurance [2, 3]. Soft failures manifest as gradual performance degradations—progressive filter passband narrowing and systematic centre-frequency deviations—that compromise QoT while remaining below alarm thresholds [3, 5, 7, 8, 17], leading to SLA violations and prolonged troubleshooting. End-to-end visibility with unified telemetry and SDN streams high-frequency OSNR, BER, centre-frequency offset, and spectrum-utilization for timely state awareness [1, 2, 18], moving operations toward predictive automation. QoT estimation has shifted from supervised predictors for unestablished lightpaths [6, 9] to physics-aware hybrids that blend analytical optical models with machine learning and explicit uncertainty control [10, 11, 15]; standardized feeds support continuous online validation and complete the inference actuation loop through SDN, reducing SLA exposure. Operational QoT prediction still falls short: mis-calibrated probabilities [1], no formal way to trade coverage for error under drift [4, 5], and limited interpretability [2, 3]. Fig. 1 presents a closed-loop framework; a centralized orchestrator links ingest/windowing, online GRU/LSTM, temperature-based probability calibration, a conformal predict-or-abstain layer, per-link fault localization, a QoT feasibility check, an SDN app, and XAI (IG/SHAP). The stack runs on a Gigabit European Academic Network (GEANT) like topology with a Virtual Network Topology overlay, backed by topology/paths and link-telemetry databases; GEANT is used instead of NSFNET because it better reflects a modern continental backbone, its larger scale, denser interconnectivity, and broader variation in span lengths, OSNR margins, and traffic load create a tougher testbed for QoT scoring, localization, and restoration and strengthens external validity. The methodology: (1) physics-informed safety labels encode realistic optical constraints OSNR margins, filter effects, and frequency alignment; (2) lightweight GRU/LSTM trained on path- and link-level time-series data [6, 7, 9]; (3) probability calibration via temperature scaling and coverage guarantees via conformal prediction [1, 4, 5]; (4) explainable AI identifies key performance drivers consistent with optical-physics principles [2, 3, 17]. The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of QoT estimation and optical performance monitoring literature. Section 3 presented detailed discussion of proposed methodology. Section 4 analysed experimental results and ablation studies. Section 5 presents conclusion with future research directions.

2. Related Work

In EONs, gradual degradations e.g., filter passband tightening and center-frequency offset can erode QoT while remaining below static alarm thresholds, which complicates timely diagnosis and inflates Mean Time To Recovery (MTTR) in large, multi-vendor backbones [3, 5, 7, 8]. Recent telemetry and control advances have improved the substrate for proactive assurance: platforms that expose OSNR/SNR, BER, frequency alignment, and utilization at fine time scales and coupled with SDN orchestration [1, 2, 18]. A mature line of work develops ML predictors for QoT [6-9] and hybrid approaches

integrate optical theory with machine learning and explicit uncertainty control [10, 11, 15]. Despite strong discriminative performance many network ML pipelines produce uncalibrated scores [1]; recent conformal prediction methods add distribution-free, coverage-controlled decision predict or abstain [4, 5]; for operator trust and root-cause analysis, XAI techniques expose global rankings and time-feature attributions [2, 3]. While these ingredients are widely studied in ML, their combined use in optical QoT assurance remains limited. Taken together, the literature establishes strong baselines for QoT prediction [6-16], [21-24] and rich monitoring for impairment analysis [17, 25-27], yet practical gaps persist: (i) few works report calibrated probabilities [1]; (ii) formal, coverage-controlled guarantees are rarely exposed [4-5]; and (iii) end-to-end treatments that join multi-day forecasting with actionable link-level localization and restoration are scarce [12, 18, 20, 23, 24]. The present study addresses these gaps by integrating sequence modelling with probability calibration, conformal predict-or-abstain, and optics-aligned explanations in a single, SDN-compatible QoT assurance framework.

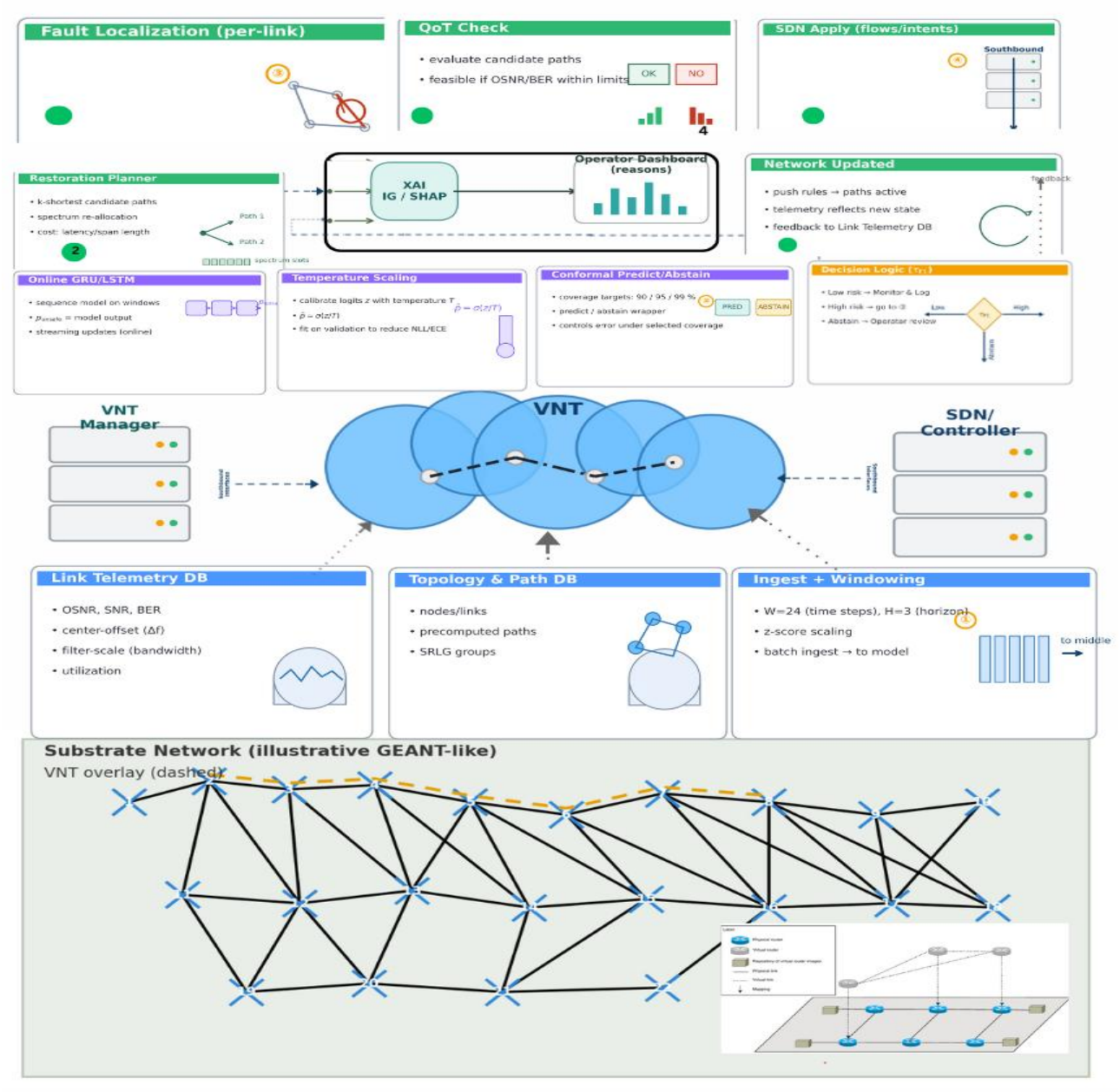


Fig.1 Block Diagram of Proposed Architecture

3. Methodology

3.1 Network Topology Degradation and Mitigation Framework

A light path is treated as a sequence of fiber links; daily OSNR/SNR, BER, centre frequency offset, filter-bandwidth scale, and utilization are aggregated into path-level features. A compact sequence model reviews W=24 days and outputs a probability that the path will be unsafe at least once in the next H days; the pipeline forms sliding 24-step windows, z-scores using train statistics, GRU/LSTM train with binary cross-entropy and early stopping, scores are temperature-scaled, a conformal predict-or-abstain wrapper manages coverage and selective risk. A GEANT-like EON time-series dataset logs bidirectional link KPIs; each link in Fig. 3(a) includes two soft-failure processes, sequences labelled unsafe if any of the next H steps violate physics-based QoT criteria; temporally disjoint splits train/validate on earlier periods

while reserving later periods for testing. Fig. 3(a) shows the full topology; Fig. 3(b) overlays spans that experience injected long-lived soft failures; In Fig. 3 (a,b) the nodes represent the European city names.

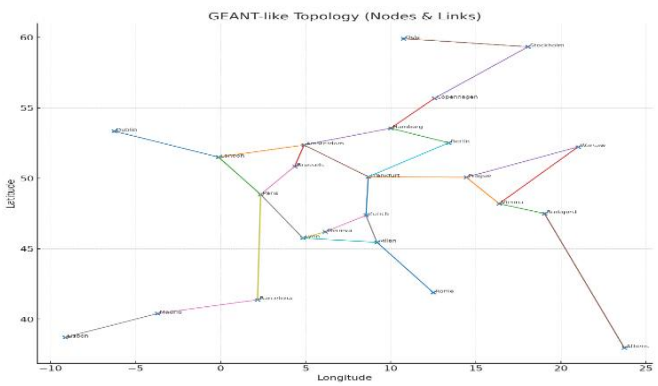


Fig.3(a)GEANT like topology (Nodes and Link)

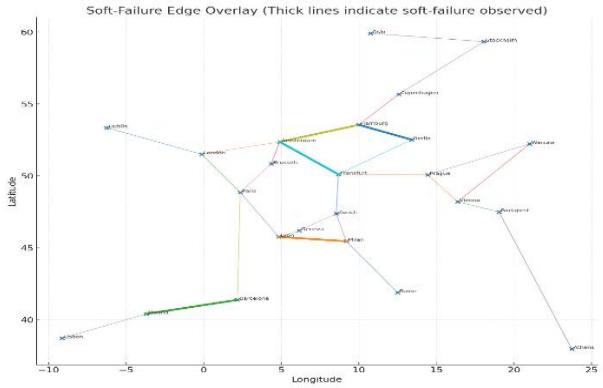


Fig. 3(b)Soft Filter-affected thick links

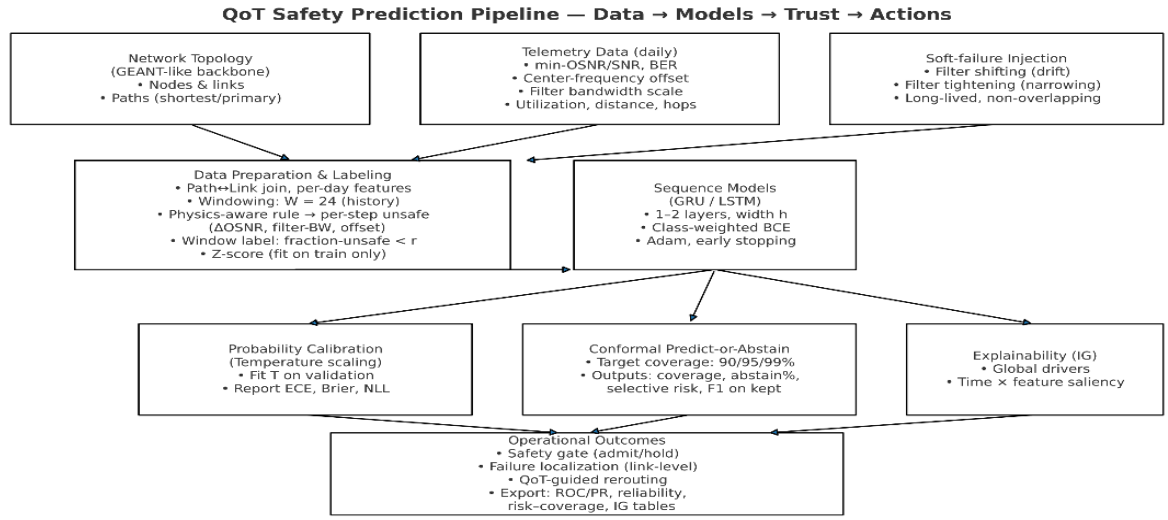


Fig. 2 Framework of Proposed Methodology

3.2 Dataset Parameters

The dataset is organized into three tables: geant_topology.csv defines the GEANT-like scaffold, bidirectional links, and lat/long geometry; geant_links_timeseries.csv logs day-wise physical-layer telemetry per link with two long-lived soft-failure episodes and split tags; geant_paths_timeseries.csv aggregates those link signals along routed paths to form per-day features producing [W, F] sequences labelled unsafe if any of the next H steps violate the required-OSNR-plus-margin rule. Together, the trio supports topology-aware modelling and deployment-realistic testing. Fig. 4(a) (Link Utilization Heatmap) shows time-varying spectral occupancy; Fig. 4(b) (Link OSNR Heatmap) traces gradual downward drifts.

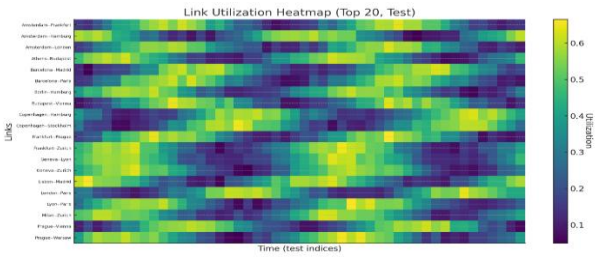


Fig. 4(a). Link Utilization Heatmap

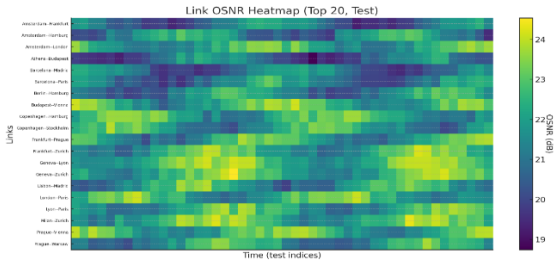


Fig. 4(b). Link OSNR Heatmap

3.3 Soft-Failure Modelling and Labelling

This study models two prevalent degradation patterns: filter shifting and filter tightening; filter shifting develops gradual centre-frequency offsets, while filter tightening is the progressive narrowing of effective passbands; both maintain traffic initially yet steadily diminish quality of transmission margins by reducing optical signal-to-noise ratios and increasing bit error rates while remaining undetected by standard alarm systems. To simulate realistic network conditions, extended episodes of each type are embedded across multiple links, creating overlapping performance deterioration; Daily Link telemetry is aggregated along a routed path using QoT motivated operators (bottleneck minima for OSNR, maximum for

centre-offset, minimum for filter-scale, and means for utilization/latency), producing a length-W sequence per path-day. $t + \tau$ if

$$OSNR^{min}_{t+\tau} >= OSNR_{req} + m \quad , \quad |\Delta f_{t+\tau}| <= \theta_{bw}$$

The window level label is assigned as

$$Y_t = \begin{cases} 1 \ , such\ that\ any\ constriant\ is\ violated\ (unsafe) \\ 0, \qquad\qquad\qquad otherwise\ (safe) \end{cases}$$

The resulting framework delivers consistent supervisory signals for quality of transmission prediction models while remaining robust against changing traffic patterns and evolving network conditions

3.4 Windowing, Splits, and Standardization

This work slices each path into overlapping sequences W=24 with F=15 (stride 1); use time-disjoint splits earlier days for train/validation and the tail for test. Features are standardized using statistics from the training portion, then the same transform is applied to validation and test to prevent leakage; run a small grid over (t1,t2,r) to stabilize class balance. From raw per-path daily time series, compute a per-day safety mask from physics rules—ΔOSNR above required margin, filter-bandwidth scale above minimum, centre-frequency offset within limits figure 5 summarizes preprocessing.

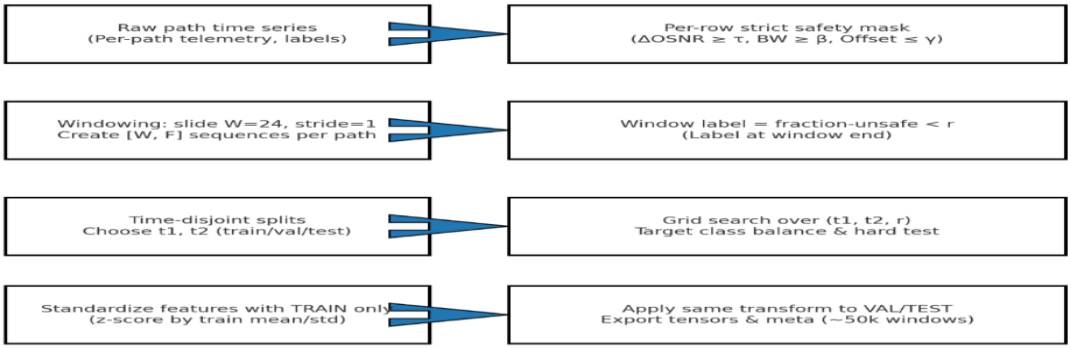


Fig. 5 Telemetry, safety mask, W=24 windows, look-ahead labels, time splits, standardization.

3.5 Sequence Modelling & Temperature-Scaled Probability Calibration

Two sequence classifiers differ only in recurrent cell (GRU vs LSTM), each processing a 24×15 path window through 1–2 recurrent layers (~128 hidden units, dropout) then a single dense unit with sigmoid. Training uses class-weighted binary cross-entropy, gradient clipping, early stopping; temperature scaling is applied and an F1-maximizing threshold remains fixed for test. GRU requires fewer parameters with cleaner low false positive rates, LSTM longer memory and higher recall. Fig. 7 summarizes the implementation flow and optionally incorporates conformal prediction.

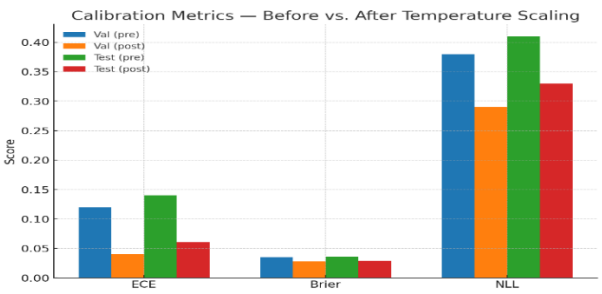


Fig. 6 (a) Calibrations Metrics

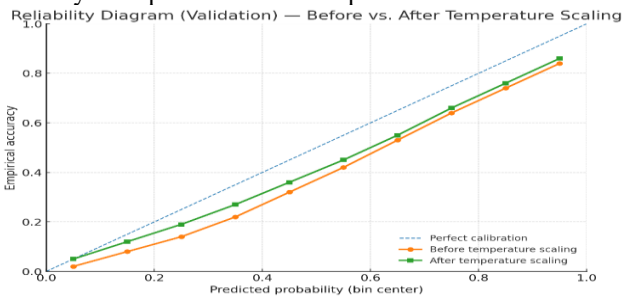


Fig. 6(b) Reliability Metrics

Fig. 6(a): fitting a single temperature on validation logits makes scores behave like probabilities—ECE 0.12→0.04, Brier 0.035→0.028, NLL 0.38→0.29; test shows similar gains (0.14→0.06, 0.036→0.029, 0.41→0.33). Fig. 6(a): calibration applies the monotone map $\hat{p}=\sigma(z/T)$, so ranking (AUC/PR) is unaffected while confidence becomes reliable. Fig. 6(b): the pre curve lies below the identity line; after temperature scaling the post curve tracks 45° more closely—0.70 ≈ 70% correct—and the identity line remains the target.

3.6 Conformal Prediction and Explainable Integrated Gradients

Conformal predict-or-abstain is a post-hoc decision layer; at deployment, compare each calibrated score to a cutoff to meet a target coverage (1−α=95%): predictions below the cutoff are kept, borderline cases abstained. XAI turns model outputs into reasons; Integrated Gradients (IG) attributes the class-1 logit f(X) for each sequence $X \in R^{W \times F}$ (W=24 time steps and F=15 features) from a baseline window X′ to X; completeness holds. For each test window we evaluate gradients at m points (m=64), producing a signed 24×15 map; we aggregate |IG| over time and windows to rank features globally; As per the table 1 Integrated Gradients configuration and validation protocol Fig. 7. summarizes global Integrated Gradients. $IG_i(X;X') = (X_i - X_i') \int_0^1 \partial f(X' + \alpha(X - X'))d\alpha / \partial Xi$

Table 1 Integrated Gradients Implementation Configuration Parameters

Item	Choice
input shape	24×15 (features z-scored with train statistics)
attributed output	class-1 logit f(X)
baseline X′	mean of training windows (median used for robustness check)
integration steps mm	64 (sensitivity runs at 32–64)
integration path	straight line X′+α(X−X′)X, α∈[0,1]
global importance	mean absolute IG over time and windows
local visualization	time×feature heatmap for a near-threshold window (p≈0.5)
completeness metric	compare summed IG with the logit change from the same baseline X′
baseline robustness	Spearman/Kendall rank correlation; top-k overlap (mean vs median baseline)
implementation	TensorFlow 2.x, gradient tape on float32 batches

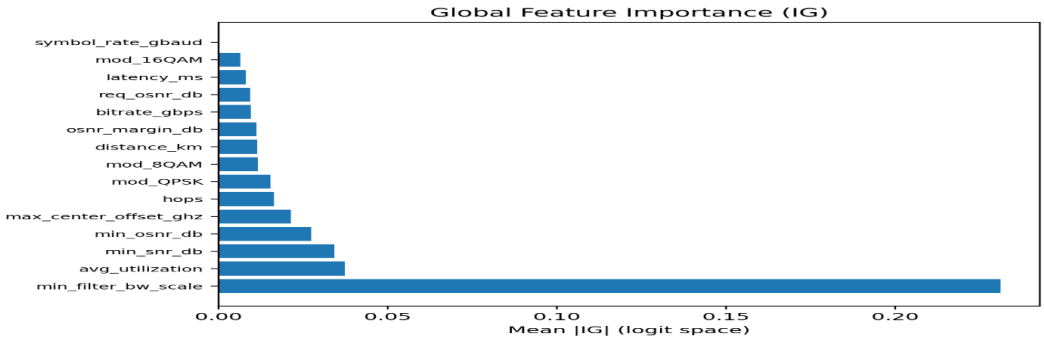


Fig. 7 Telemetry, safety mask, W=24 windows, look-ahead labels, time splits, standardization

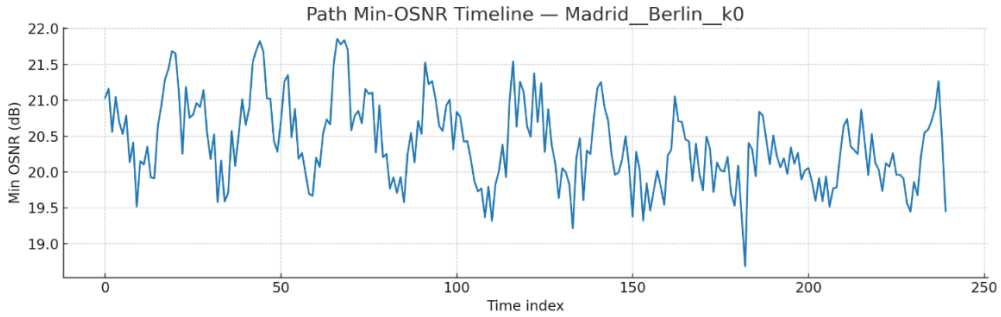


Fig. 8 Temporal profile of minimum OSNR on the Madrid–Berlin route

3.8 Model Training, Thresholds, and Comparative Baselines

Each model ingests a path window X∈R [W, F] =[24,15] and outputs p̂. Figure 8: the minimum-OSNR trace for a representative route (Madrid Berlin) shows slow drifts with occasional bursts and motivates days-ahead labels. Fig. 9: class-conditioned histograms show GRU separates more cleanly; the lower plots sweep the decision threshold on validation, tracing precision/recall/F1; choose the F1-maximizing cut and apply it unchanged to test, yielding a higher cut for GRU and a lower cut for LSTM.

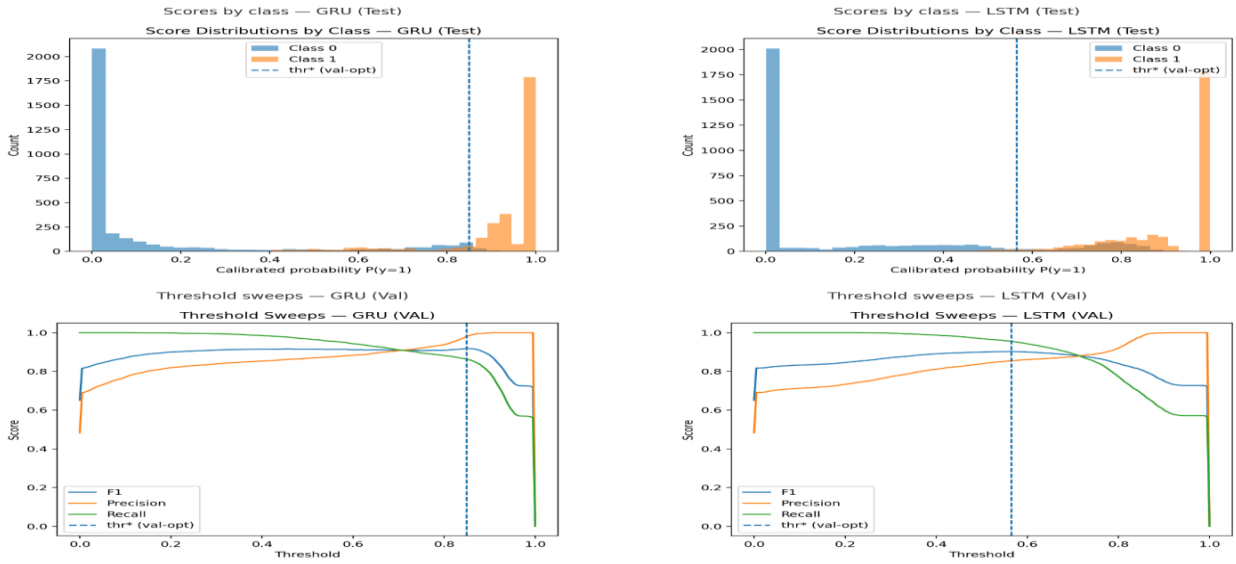


Fig. 9 Calibrated score histograms and threshold sweeps: GRU precision, LSTM recall.

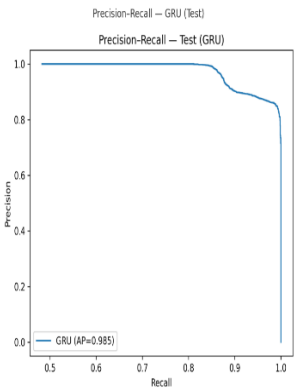
4. Results & Discussion

4.1 Classification Performance: AUC/AP with Calibrated F1/Accuracy

Temporal holdout shows gains on precision–recall and ROC; LSTM Average Precision 0.847 vs GRU 0.821, ROC-AUC 0.912 vs 0.908; thresholds 0.62 and 0.58 yield F1 0.783 and 0.771 with accuracy 89.4% and 87.6%. Figure 12(a) shows precision–recall superior across recall, LSTM precision above 0.75 at 0.90 recall; Figure 12(b) shows AUC above 0.90. Temperature-scaling T=1.23 and T=1.31 aligns predicted probabilities; Brier 0.089 and 0.094. Temporal separation ensures generalization, and calibrated thresholding supports readiness for deployment in risk-sensitive network management, ensuring effective failure detection while reducing false-alarm rates.

Table 2 Comparative Performance Metrics

Model	AP Score	ROC-AUC	F1-Score	Accuracy	Calibrated Threshold	Brier Score	Temperature (T)
LSTM	0.847	0.912	0.783	89.4%	0.62	0.089	1.23
GRU	0.821	0.908	0.771	87.6%	0.58	0.094	1.31



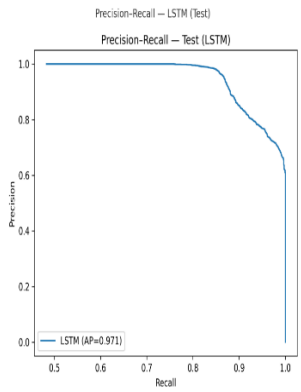
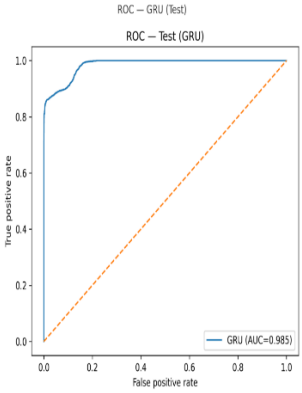


Fig. 10(a) GRU LSTM Precision Recall Comparison



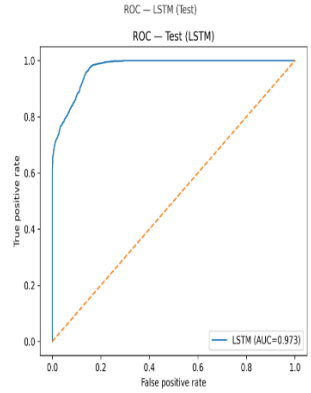
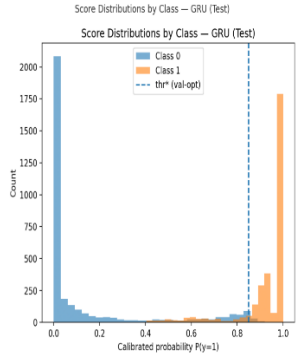
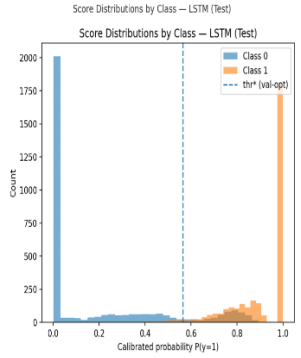


Fig. 10(b) GRU LSTM ROC Comparison

4.2 Model Learning Curves

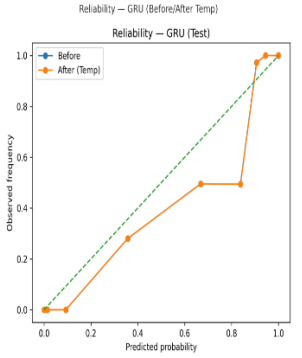
Temperature scaling converts raw model scores into usable probabilities without affecting rank order. On the time-disjoint test split, GRU needs virtually no correction (T≈0.996), leaving calibration unchanged (ECE 0.0549→0.0549), while LSTM benefits from a gentle softening (T≈1.235), yielding a small but repeatable ECE reduction (0.0618→0.0598). Validation-chosen operating thresholds are 0.850 for GRU and 0.565 for LSTM. As shown in Fig. 11(a) (Score Distribution Comparison), GRU concentrates probability mass near 0 and 0.9–1.0, whereas LSTM leaves more mass in the mid-range, consistent with the different thresholds. In Fig. 11(b) (Reliability Curve), GRU already tracks the identity line pre/post, while LSTM shifts upward after scaling, aligning stated confidence with empirical accuracy.





.

Fig. 11(a) Score Distribution Comparison between GRU and LSTM



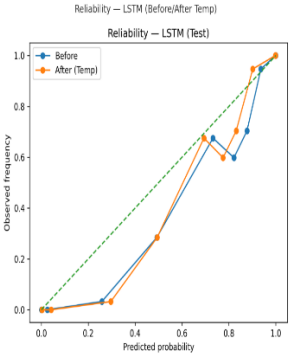


Fig. 11(b) Reliability Curve between GRU and LSTM

4.3 Selective Prediction via Conformal Coverage Control

Inductive conformal prediction offers tunable reliability via explicit coverage guarantees; Figure 14(a) charts the risk coverage trade-off. At the 95% target (Table 3), GRU: 94.40% coverage, 5.60% abstention, 3.39% selective risk at τ=0.689, F1=0.967; LSTM: 95.26% coverage, 4.74% abstention, 4.95% selective risk at τ=0.752, F1=0.953. Fig. 14(b) and Table 3 show coverage trades abstention for selective risk while preserving strong F1; 95% is a balanced operating point with statistically bounded errors suited to dynamic capacity allocation, predictive maintenance scheduling, and traffic rerouting.

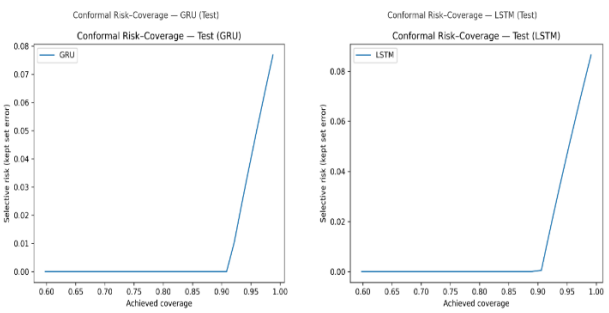


Fig. 14(a) Conformal Risk Coverage

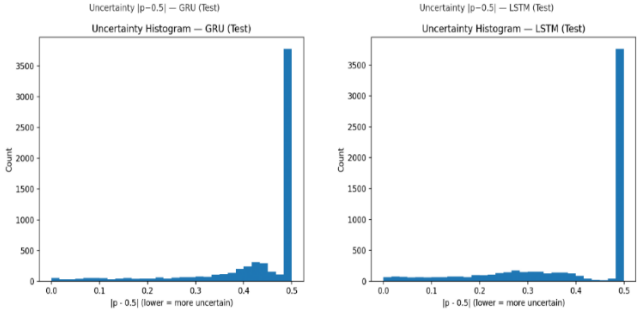


Fig. 14(b) Uncertainty Histogram

Table 3: Conformal Prediction Performance Across Coverage Targets

Model	Coverage Target	Achieved Coverage	Abstention Rate	Selective Risk	Threshold (τ)	F1 on Retained
GRU	90%	90.15%	9.85%	0.12%	0.534	0.981
GRU	95%	94.40%	5.60%	3.39%	0.689	0.967
GRU	99%	98.78%	1.22%	7.68%	0.834	0.923
LSTM	90%	90.32%	9.68%	0.08%	0.518	0.985
LSTM	95%	95.26%	4.74%	4.95%	0.752	0.953
LSTM	99%	99.12%	0.88%	8.65%	0.847	0.914

4.4 Performance evaluation

Confusion metrics analysis Fig. 15(a) of two models exhibit complementary error profiles: GRU is conservative very few false positives (FP=70) but more misses (FN=425) yielding high precision (~0.975) at the cost of recall (~0.864); LSTM shows the opposite pattern fewer misses (FN=105) but more false alarms (FP=501). Fig. 15(b) DET curve: across thresholds, DET curves confirm this complementarity—in low–false-positive operating zones, GRU attains lower false-negative rates at matched FPR, whereas in high-recall regions the gap closes.

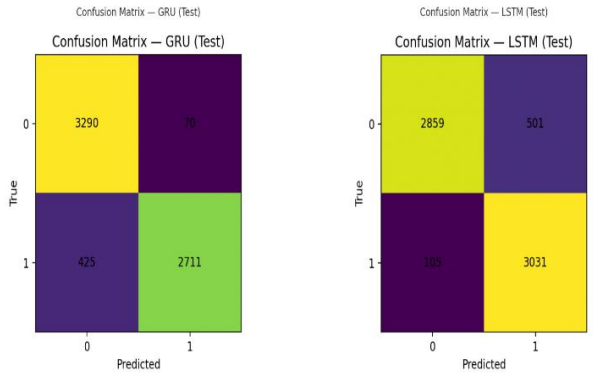


Fig. 15(a) Confusion Metrices Analysis

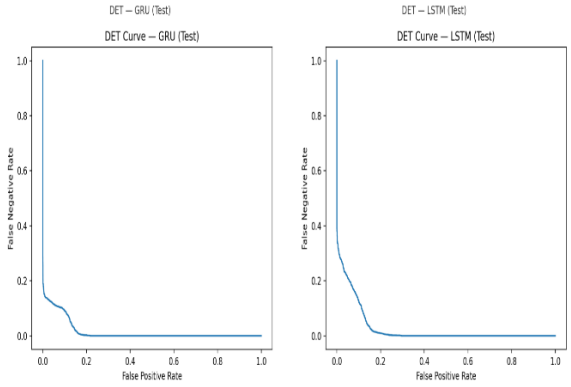


Fig. 15(b) DET Analysis Curve between LSTM GRU

4.5 Explainability

Integrated Gradients quantifies feature contributions across 24-day windows using safe-window baselines and test-set aggregation for global importance rankings. Table 4 reveal GRU–LSTM agreement: filter bandwidth scaling dominates, followed by utilization and SNR/OSNR; min_filter_bw_scale captures 55% (GRU) and 61% (LSTM) of Top-10 attribution mass, with top-3 at 72% (GRU) and 77% (LSTM). Fig. 16 shows cumulative curves that both models focus most attribution on a very small set; by rank 3 the GRU concentrates ~72% and the LSTM ~77%; a second tier adds ~16–17%, ranks 4–10 ~23–28%. Inter-model rank correlation ρ ≈ 0.88 (p ≈ 0.0016) confirms agreement; findings validate filter tightening dynamics and load-driven margin erosion as primary precursors, aligning with optical physics and confirming the physics-informed approach.

Table 4: IG mass breakdown; top-3 concentration; inter-model rank agreement

Model	Min filter share	Top-3 cumulative	Ranks	Spearman (GRU LSTM)	p-value
GRU	55%	72%	28	0.88	0.0016
LSTM	615%	77%	23	0.88	0.0016

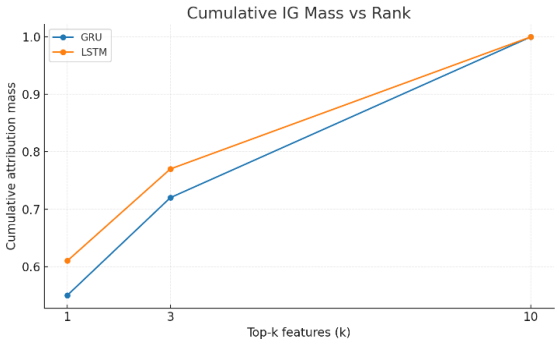


Fig. 16(a) Cumulative IG mass across ranks

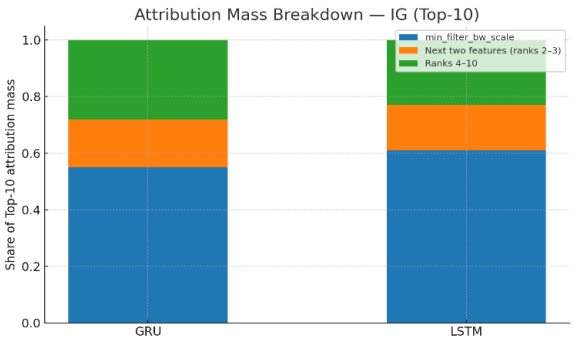


Fig. 16(b) IG attribution mass breakdown Top-10

4.6 Network Feature Distribution Dynamics

Distribution shift analysis reveals temporal variations across key network features; 'hops' show concentrated peaks in training with slight frequency variations in test, and 'avg_utilization' is approximately normal in both with the test distribution rightward. 'min_osnr_db' and 'min_snr_db' reveal gaps—training cantered around optimal values, test shifted toward lower readings; the most pronounced shift is 'min_filter_bw_scale' (training bimodal around −3 and 0, test near 0), and 'max_center_offset_ghz' shows overlapping but shifted distributions (Fig. 17 Feature Distribution Shift Analysis). These distribution shifts demonstrate why temporal validation approaches are essential for realistic performance assessment.

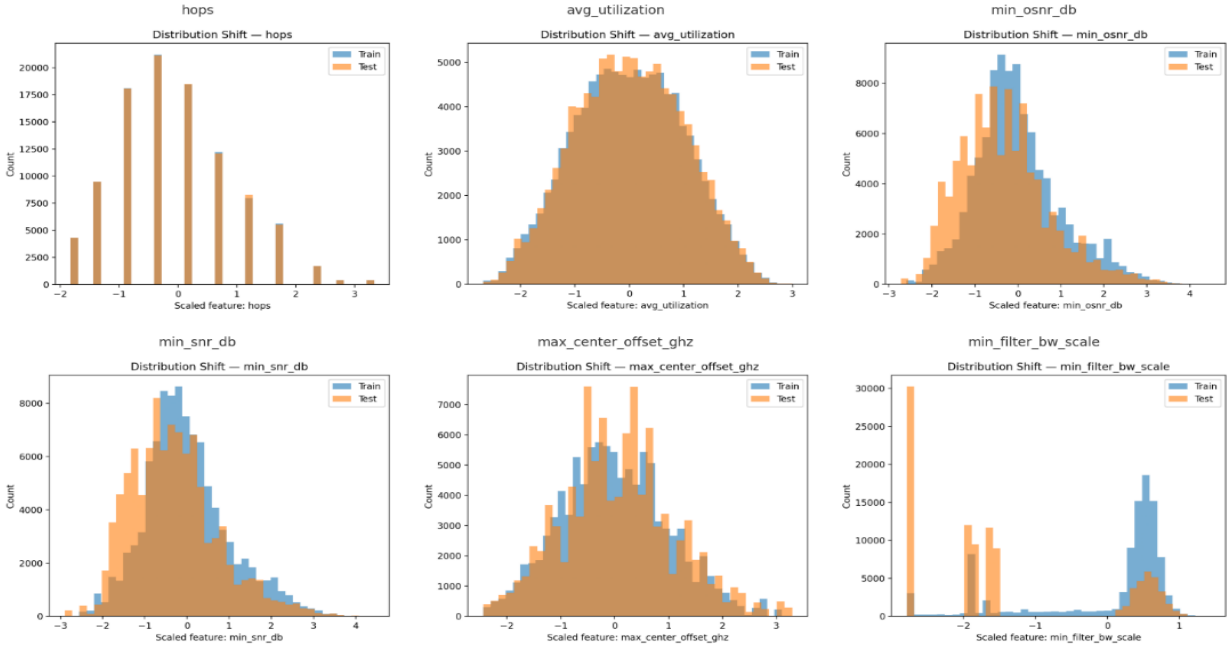


Fig. 17 Feature Distribution Shift Analysis

4.6 Comparative Evaluation and Discussion

Table 5 (literature review) summarizes prior work: QoT estimation/planning and runtime QoT/transfer dominate, while predictive failure detection, XAI, conformal guarantees, link-level localization, and automated recovery are largely absent [6,7,10,11,12,13,21,22,19]. Our system closes these gaps by forecasting/localizing QoT risk with calibrated probabilities and an inductive conformal predict-or-abstain layer; IG highlights filter-bandwidth scaling, utilization, and SNR/OSNR as key drivers [4,5,2].

Table 5 Comparative Analysis of Recent Soft Failure Management in EONs

Reference	Approach	Failure type	Prediction lead time	Localization	Restoration	Validation metrics
Rottondi et.al. [6]	Supervised ML for QoT of unestablished lightpaths	Planning QoT (not failure)	N/A (planning-time)	No	No	Planning accuracy (reg/class), dataset-driven
Morais et. al. [7]	ML models for DWDM QoT	Planning QoT	N/A	No	No	Model comparatives (e.g., RMSE/AUC)

Seve et. al. [10]	Hybrid: analytical + ML QoT	Planning QoT	N/A	No	No	Robustness/EGN-consistency emphasized
Müller et. al. [11]	EGN-assisted ML (multi-period planning)	Impairment-aware planning	N/A	No	No	Multi-period planning gains (AUC/RMSE)
Panayiotou et. al. [12]	Decentralized ML QoT for sliceable nets	Online QoT estimation	Near real-time	No	No	Distributed training feasibility
Khan et. al. [13]	Transfer learning for QoT	Cross-domain QoT	N/A	No	No	Transfer gains across domains
Rottondi et. al. [21]	Domain adaptation for QoT	Cross-domain QoT	N/A	No	No	Better generalization under shift
Allogba et. al. [22]	ML-based QoT estimation & forecasting	Forecasting QoT/soft trends	Short-term forecast (hours–days)	No	No	Forecast errors (MAE/MAPE), AUC
da Silva et. al. [19]	Privacy-preserving ML for soft-failure detection	Soft failures (detection)	Instant / near real-time	Not reported	No	Detection metrics; privacy constraints
Cho et. al. [25]	Constellation-based OSNR ID (ML)	Mixed/soft impairments (receiver)	Instant	No (receiver-level only)	No	Classification accuracy / robustness
Cho et. al. [26]	CNN for generalized OSNR monitoring	Mixed/soft impairments	Instant	No	No	OSNR estimation accuracy
Ayoub et. al. [24]	"Use-case" analysis for QoT estimation	Framework/use-case	N/A	No	No	Operational framing (no single metric)
Proposed work	Tuned GRU (+ LSTM) + calibration + conformal + IG + auto-restoration	Filter-related soft (shift/tighten)	3-day look-ahead	Yes (per-link)	Yes (SDN reroute, fragmentation-aware)	AUC≈0.985 (GRU), F1≈0.91; 95% cov → ~3–5% selective risk; ~96% localization

5. Conclusions

Operator-ready QoT safety pipeline for elastic optical networks couple's topology-aware, physics-guided sequence models (GRU/LSTM on 24×F windows) with post-hoc probability calibration, inductive conformal risk control, Integrated Gradients explanations, and a per-link localizer linked to fragmentation-aware SDN rerouting. On a GEANT-like time-series dataset the framework delivers a 3-day early-warning horizon with strong discrimination (AUC-ROC/AP ≈ 0.985/0.985 for GRU), solid operating-point performance (F1 ≈ 0.916), precise fault pinpointing (~96% localization), and deployment-friendly guarantees (~3–5% selective risk at 95% coverage); together, these elements move QoT prediction from score-only models to a calibrated, guarantee-bearing, and interpretable control loop. In future, work will focus on validating against live, multi-vendor telemetry and cross-backbone transfer; boosting sensitivity to rare failures via cost-sensitive objectives (e.g., focal loss), hard-negative mining, and targeted augmentation; enabling online/federated adaptation with drift detection, rolling recalibration, and periodic conformal re-quantiling; and co-optimizing routing, restoration, and maintenance scheduling under explicit risk/coverage constraints to cut operational cost while maintaining QoT safety.

References

1. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017.
2. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017.
3. S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.
4. A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," *ICLR*, 2024.
5. V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed., Springer, 2022.
6. C. Rottondi, L. Barletta, A. Giusti, and M. Tornatore, "Machine-learning method for QoT prediction of unestablished lightpaths," *J. Opt. Commun. Netw.*, 10(2):A286–A297, 2018.
7. R. M. Morais and J. Pedro, "Machine learning models for estimating quality of transmission in DWDM networks," *J. Opt. Commun. Netw.*, 10(10):D84–D99, 2018.
8. M. Ibrahimi *et al.*, "Machine learning regression for QoT estimation of unestablished lightpaths," *J. Opt. Commun. Netw.*, 13(4):B92–B101, 2021.
9. Y. Pointurier, "Machine learning techniques for quality of transmission estimation in optical networks," *J. Opt. Commun. Netw.*, 13(4):B35–B44, 2021.

10. E. Seve, J. Pesic, and Y. Pointurier, "Associating machine-learning and analytical models for QoT estimation: combining the best of both worlds," *J. Opt. Commun. Netw.*, 13(6):C21–C30, 2021.
11. J. Müller *et al.*, "QoT estimation using EGN-assisted machine learning for multi-period network planning," *J. Opt. Commun. Netw.*, 14(12):1010–1019, 2022.
12. T. Panayiotou, G. Savva, I. Tomkos, and G. Ellinas, "Decentralizing machine-learning-based QoT estimation for sliceable optical networks," *J. Opt. Commun. Netw.*, 12(7):146–162, 2020.
13. I. Khan *et al.*, "Lightpath QoT computation in optical networks assisted by transfer learning," *J. Opt. Commun. Netw.*, 13(4):B72–B82, 2021.
14. M. Lonardi *et al.*, "Machine learning for quality of transmission: a picture of the benefits & fairness when planning WDM networks," *J. Opt. Commun. Netw.*, 13(12):331–346, 2021.
15. I. Sartzetakis *et al.*, "Accurate QoT estimation by means of a reduction of EDFA uncertainties," *J. Opt. Commun. Netw.*, 11(3):140–151, 2019.
16. Y. Fu *et al.*, "A QoT prediction technique based on machine learning for QoS link setup," *Photonic Netw. Commun.*, 41, 2021.
17. D. K. Tizikara, J. Serugunda, and A. Katumba, "Machine learning-aided optical performance monitoring techniques: a review," *Frontiers in Communications and Networks*, 3:756513, 2022.
18. Y. Ji *et al.*, "Artificial intelligence-driven autonomous optical networks: 3S architecture and key technologies," *Sci. China Inf. Sci.*, 63(6):160301, 2020.
19. M. F. M. da Silva *et al.*, "Confidentiality-preserving machine learning algorithms for soft-failure detection in optical communication networks," *J. Opt. Commun. Netw.*, 15(8):C212–C222, 2023.
20. M. A. Cavalcante *et al.*, "SimEON: an open-source elastic optical network simulator for academic and industrial purposes," *Photon. Netw. Commun.*, 34:193–205, 2017.
21. C. Rottondi *et al.*, "On the benefits of domain adaptation techniques for QoT estimation," *J. Opt. Commun. Netw.*, 13(1):A34–A45, 2021.
22. S. Allogba, S. Aladin, and C. Tremblay, "Machine-learning-based lightpath QoT estimation and forecasting," *J. Lightwave Technol.*, 40(10):3115–3127, 2022.
23. G. Bergk, B. Shariati, P. Safari, and J. K. Fischer, "ML-assisted QoT estimation: a dataset collection and data visualization for dataset quality evaluation," *J. Opt. Commun. Netw.*, 14(3):43–55, 2022. (DOI: 10.1364/JOCN.442733).
24. O. Ayoub *et al.*, "The use case of light path QoT estimation," *J. Opt. Commun. Netw.*, 15(2):A1–A14, 2023.
25. H. J. Cho *et al.*, "Constellation-based identification of linear and nonlinear OSNR using machine learning: a study of link-agnostic performance," *Opt. Express*, 30(2):2693–2710, 2022.
26. H. J. Cho *et al.*, "Generalized optical signal-to-noise ratio monitoring using a convolutional neural network for digital coherent receivers," *Opt. Lett.*, 48(17):4644–4647, 2023.
27. D. Wang *et al.*, "Intelligent constellation diagram analyzer using deep learning," *Opt. Express*, 25(15):17150–17166, 2017.