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

**DEVENDRA SRIVASTAVA[1] SHWETA TRIPATHI[1,2*] PRIYANKA GAUTAM [1,2]**

*[1]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), Quality-of-Transmission (QoT) prediction, Soft failures (filter shifting/tightening), Sequence models (GRU/LSTM), Probability calibration (temperature scaling), Conformal prediction (risk–coverage), Explainable AI (Integrated Gradients, SHAP), Per-link fault localization, Topology-aware telemetry / graph fingerprints, QoT-guided rerouting (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].

Network operators today face a particularly troublesome category of failures known as soft failures. Unlike traditional hard faults that trigger immediate alarms, soft failures manifest as gradual performance degradations—such as progressive filter passband narrowing and systematic center-frequency deviations—that slowly compromise Quality-of-Transmission (QoT) parameters while remaining below conventional alarm thresholds [3], [5], [7], [8], [17]. These insidious degradations can persist undetected for extended periods, ultimately leading to service-level agreement violations and prolonged troubleshooting cycles in production environments.

Recent advances in network observability have created unprecedented opportunities for proactive network management. Contemporary unified telemetry systems and software-defined networking (SDN) control platforms now provide high-frequency measurements of critical parameters including optical signal-to-noise ratio (OSNR), bit error rate (BER), center frequency alignment, and spectral utilization [1], [2], [18]. This wealth of real-time data enables a paradigm shift from reactive, ticket-based maintenance toward predictive, automated network operation.

The research community has responded to these opportunities with significant advances in QoT estimation methodologies. Initial investigations focused on supervised learning approaches for unestablished light path quality prediction, establishing foundational principles regarding feature selection, training objectives, and deployment strategies

[6], [7], [8], [9]. Subsequently, hybrid methodologies emerged that combine analytical optical models with machine learning techniques, such as enhanced Gaussian noise model integration and uncertainty-quantified variants, demonstrating improved robustness and data efficiency [10], [11], [15].

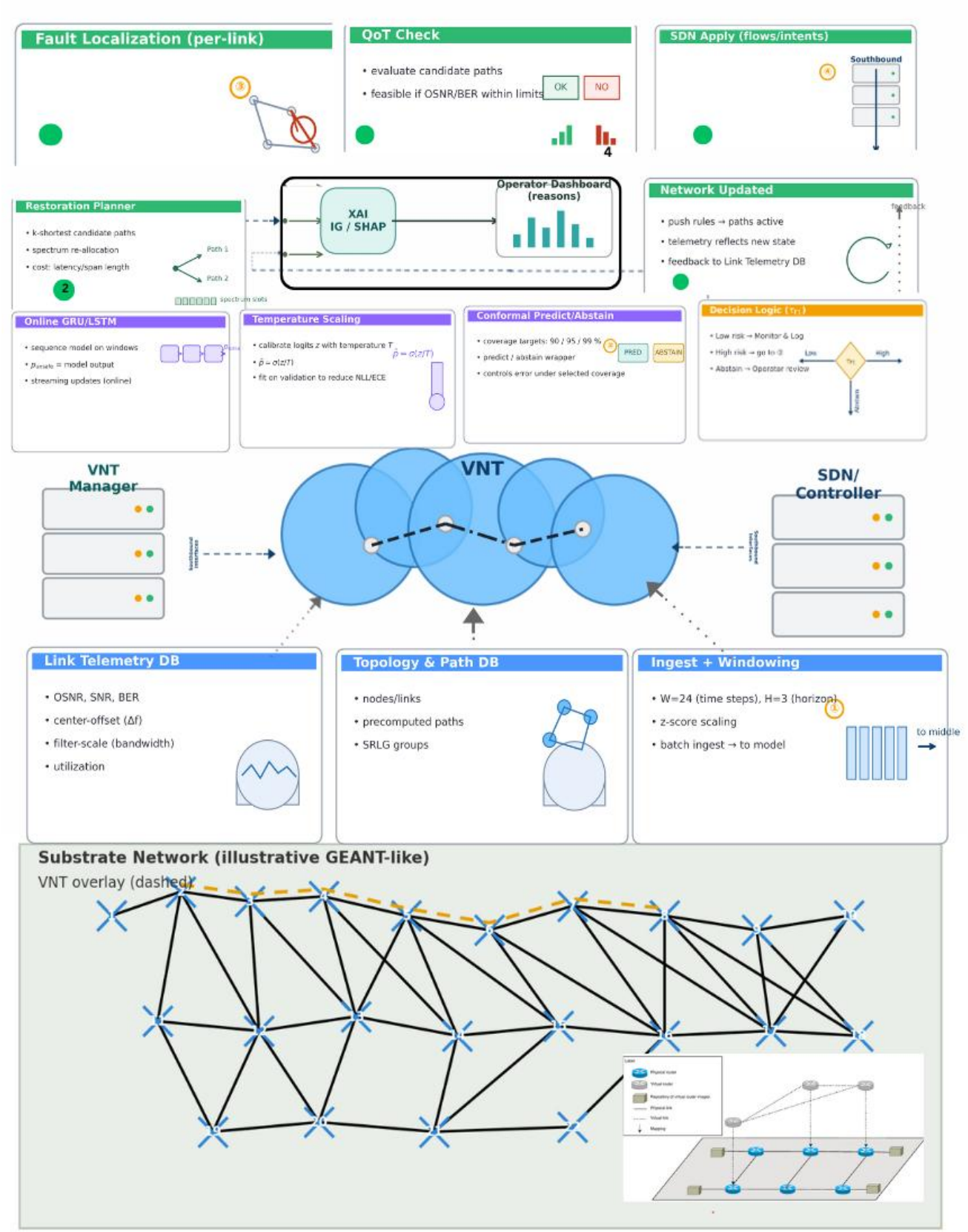



Fig.1 System Design Architecture

Parallel research directions have explored distributed learning architectures that accommodate network slicing constraints [12], domain adaptation techniques to reduce site-specific dependencies [13], [21], and optimization frameworks that balance fairness considerations with performance benefits in network planning [14]. The development of open-source simulation environments and curated QoT datasets, exemplified by tools like Sim-EON, has significantly enhanced research reproducibility and evaluation standards [20], [23]. Concurrently, comprehensive surveys of optical performance monitoring applications have systematically catalogued machine learning use cases and sensing methodologies relevant to soft failure detection and management [17].

While recent advances have shown promise, three fundamental challenges prevent existing QoT prediction methods from meeting operational requirements in production networks.

**Unreliable Probability Estimates**: Most prediction models generate poorly calibrated confidence scores, making it difficult for network controllers to establish meaningful decision thresholds. Although calibration techniques like temperature scaling can address this issue, their use in optical networks remains underexplored [1].

**Absence of Formal Guarantees**: Current systems lack mechanisms that allow operators to balance prediction coverage against acceptable error rates under dynamic network conditions. Conformal prediction offers a solution by providing distribution-free "predict or abstain" frameworks with controllable risk-coverage trade-offs [4], [5].

**Limited Interpretability**: Network operators need transparent explanations to trust automated decisions and perform effective troubleshooting. Despite the availability of established attribution methods like Integrated Gradients and SHAP, explainable AI has seen minimal adoption in optical networking applications [2], [3], [17].

We address these challenges through a unified **predict–calibrate–guarantee–explain** framework that converts standard network telemetry into trustworthy, actionable QoT decisions.

Figure 1 presents a closed-loop network-management framework that fuses telemetry-driven inference with automated restoration. A centralized orchestrator links the core subsystems—ingest/windowing, online **GRU/LSTM** for QoT/soft-failure risk scoring, temperature-based probability calibration, a conformal **predict-or-abstain** layer, per-link fault localization, a QoT feasibility check, and an SDN app for actuation—while an XAI module (IG/SHAP) surfaces operator-facing explanations and dashboards. At the substrate, the stack runs on a GEANT-like topology with a VNT overlay, backed by topology/paths and link-telemetry databases for real-time state tracking and predictive maintenance. This integrated design enables autonomous optimization and rapid fault recovery while preserving human oversight through interpretable alerts and controls.

We use **GEANT** instead of **NSFNET** because it better reflects a modern continental backbone. Its larger scale and denser interconnectivity produce richer route multiplicity and more intricate SRLG structure, exposing corner cases that NSFNET rarely triggers. GEANT also exhibits broader variation in span lengths, OSNR margins, and traffic load, creating a tougher testbed for QoT scoring, localization, and restoration; evaluating NSFNET→GEANT naturally exercises cross-topology transfer and strengthens external validity.

Our methodology operates in four stages: First, we create physics-based safety labels that capture realistic optical constraints including OSNR margins, filter effects, and frequency alignment. Second, we train lightweight GRU/LSTM models on time-series data from network paths and links [6], [7], [9]. Third, we apply temperature scaling for probability calibration and conformal prediction for coverage guarantees [1], [4], [5]. Finally, we use explainable AI to identify key performance drivers that align with optical physics principles [2], [3], [17].

**Key Contributions**

**Smart Labelling and Modelling**: Physics-aware labels combined with efficient sequence models that understand network topology and temporal patterns [6], [7], [9], [22].

**Deployment-Ready Outputs**: Calibrated probabilities and formal coverage guarantees that enable automated control with explicit risk management [1], [4], [5].

**Meaningful Explanations**: Attribution techniques that consistently highlight relevant optical factors like bandwidth constraints, OSNR levels, and frequency offsets for practical diagnosis [2], [3], [17].

**Rigorous Evaluation**: Comprehensive testing on realistic network topologies using open datasets to support reproducible research [18], [20], [23], [24].

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of QoT estimation and optical performance monitoring literature, with particular emphasis on calibration techniques, formal guarantees, and explainability methods [1]– [5], [6]– [17], [21]– [24], [25]– [27]. Section 3 presents our methodology in detail, covering data preparation, physics-aware labelling strategies, model architecture design, calibration and conformal prediction implementation, and explainable AI integration. Section 4 reports experimental results and ablation studies. Section 5 discusses implementation limitations and practical deployment considerations. Section 6 concludes with summary findings and future research directions.

## 2. Related Work

In elastic optical networks (EONs), gradual degradations—e.g., filter passband tightening and center-frequency drift—can erode Quality-of-Transmission (QoT) while remaining below static alarm thresholds, which complicates timely diagnosis and inflates 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, coupled with SDN orchestration, enable continuous analytics instead of periodic polling [1], [2], [18]. Surveys of optical performance monitoring (OPM) document the sensing landscape and motivate learning-based detection for soft failures beyond simple thresholding [17].

A mature line of work develops ML predictors for QoT, especially for unestablished lightpaths and planning-time evaluation. Foundational studies specify features and targets for supervised QoT prediction and show that data-driven models can complement traditional design rules [6], [7], [8], [9]. Hybrid approaches integrate optical theory with learning to improve robustness and sample efficiency—e.g., analytical/ML combinations and EGN-assisted schemes for multi-period planning—showing that physics priors can stabilize estimates across scenarios [10], [11], [15]. Parallel efforts investigate decentralized and sliceable settings [12], transfer and domain adaptation to reduce site dependence [13], [21], and planning trade-offs that account for fairness and benefit across network tenants [14]. Beyond point prediction, forecasting studies examine temporal evolution of lightpath performance using recurrent models, providing evidence that sequential structure is informative for near-term QoT evolution [22], while practical "use-case" analyses clarify deployment assumptions and interfaces to operations tooling [24]. On the tooling side, open simulators (e.g., SimEON) and curated datasets have improved reproducibility and dataset quality assessment—key enablers for benchmarking and cross-paper comparability [20], [23].

Complementary to path-level QoT estimation, receiver-side learning has been used for impairment monitoring and OSNR estimation directly from constellation data, targeting link-agnostic behavior and robustness to mixed impairments [25], [26], [27]. These studies reinforce the value of feature sets that reflect physical mechanisms (e.g., bandwidth filtering, OSNR headroom, frequency offset), which also appear as salient drivers in explainability analyses for QoT tasks.

Despite strong discriminative performance, many network ML pipelines produce uncalibrated scores, complicating threshold selection in closed loops; temperature scaling offers a simple post-hoc remedy to align probabilities with empirical frequencies [1]. Recent conformal prediction methods add distribution-free, coverage-controlled decisions—"predict or abstain" with explicit risk–coverage trade-offs—well suited to volatile network conditions [4], [5]. For operator trust and root-cause analysis, explainable AI techniques such as Integrated Gradients and SHAP expose global rankings and time×feature attributions, helping validate that models rely on optics-consistent factors rather than spurious correlations [2], [3]. While these ingredients are widely studied in machine learning, their combined use in optical QoT assurance remains limited in the published literature.

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 for operator-ready pipelines: (i) few works report calibrated probabilities suitable for controllers [1]; (ii) formal, coverage-controlled guarantees are rarely exposed to NOC workflows [4], [5]; and (iii) end-to-end treatments that join multi-day forecasting with actionable link-level localization and restoration are scarce, despite available telemetry and SDN hooks [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.

## 3. Methodology

### 3.1 Network Degradation Problem and Mitigation Architecture

We study a candidate light path formed by a chain of fibre links. From each link we record daily operational signals—such as OSNR/SNR, BER, centre-frequency offset, filter-bandwidth scale, and utilization—and combine them into path-level features. On any given day t, the model reviews a recent history window of length W (in our case, W=24 days) and

returns a probability in [0,1] that the light path will become unsafe at least once during the next H days (e.g., H=3). Here, "unsafe" means the path is likely to breach its quality-of-transmission requirement within that short outlook. Training labels follow a physics-aware rule. A window is tagged unsafe (1) if, on any day within the coming H days, the path's OSNR is expected to drop below the required OSNR plus its safety margin; otherwise, it is safe (0). Framed this way, the task is not a one-time feasibility check of the current state but a short-horizon safety gate: it anticipates degradation early enough to trigger guard-banding, proactive maintenance, or re-routing before service quality is impacted.

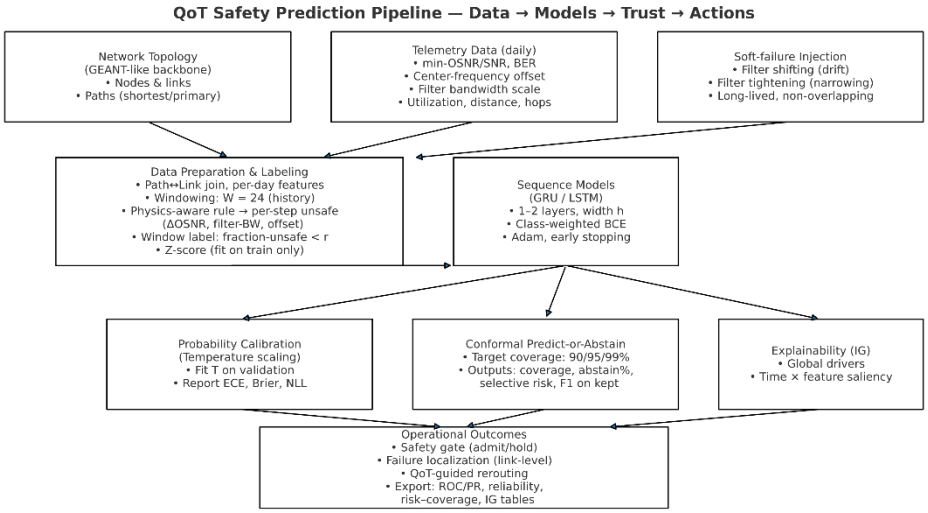



Fig. 2 Framework of Proposed Methodology

Figure 2 suggests a block architecture that starts with topology and telemetry ingestion (with optional soft-failure injections), followed by data preparation: join path–link features, slide 24-step windows, apply a physics-informed QoT rule for labels, and z-score features using train statistics. Compact sequence models (GRU/LSTM) are then trained with class-weighted binary cross-entropy and early stopping, and their probabilities are calibrated via temperature scaling. At inference, conformal predict-or-abstain controls coverage and selective risk, while Integrated Gradients surfaces dominant drivers and their timing within the window. The pipeline outputs ROC/PR and reliability plots, risk–coverage summaries, and actionable gates for QoT safety checks, localization, and routing.

## 3.2 Network Topology and Dataset Generation

We built a GEANT-like EON time-series dataset for short-term QoT safety prediction, capturing the temporal behaviour needed for soft-failure detection. It logs bidirectional link KPIs over extended periods—OSNR/SNR minima, BER, center-frequency offsets, and filter-bandwidth scaling—plus path context (hops, distance, delay, spectral utilization, required OSNR margins) [6–11,25–27]. To emulate realistic degradations, each link in Fig.3(a) includes two soft-failure processes: progressive frequency drift (temperature/aging) and gradual passband narrowing, which lower OSNR and raise BER without immediate outages, often persisting for days. Training samples use sliding windows of W=24; sequences are labelled unsafe (1) if any of the next H steps violate physics-based QoT criteria (including OSNR margin), else safe (0). Features are z-score normalized, and temporally disjoint splits train/validate on earlier periods while reserving later periods for testing to reflect concept drift and unseen operations. This design supports calibrated risk estimates and principled abstention [1,4–5] and, unlike static light-path datasets, enables early warnings and time-based explanations for root-cause analysis [6–7,10–12,22–24].

Our dataset instantiates a GEANT-like European backbone with ~20–25 Points-of-Presence (e.g., Dublin, London, Amsterdam, Paris, Frankfurt, Zurich, Milan, Vienna, Prague, Berlin, Copenhagen, Oslo, Stockholm, Warsaw, Budapest, Athens). Links follow realistic lat/long geometry, yielding diverse path lengths, node degrees, and shared-risk corridors. Each path aggregates link telemetry along its route to form path-level features—e.g., min OSNR/SNR, max center-frequency offset, min filter-bandwidth scale, mean spectral utilization—aligned with QoT practice for unestablished light paths and planning [6,9–11,14]. Fig. 3 shows the full topology (colors distinguish fiber segments). Fig. 3(b) overlays the

subset of spans that experience injected long-lived soft failures: thick strokes mark affected links, on which we apply two mechanisms—filter shift (progressive centre-frequency drift) and filter tightening (gradual passband narrowing)—over multi-day intervals to emulate realistic degradations without immediate outages. The topology is large enough to expose heterogeneous metro vs. long-haul conditions, yet controlled for rigorous benchmarking and reproducibility [20,23–24].

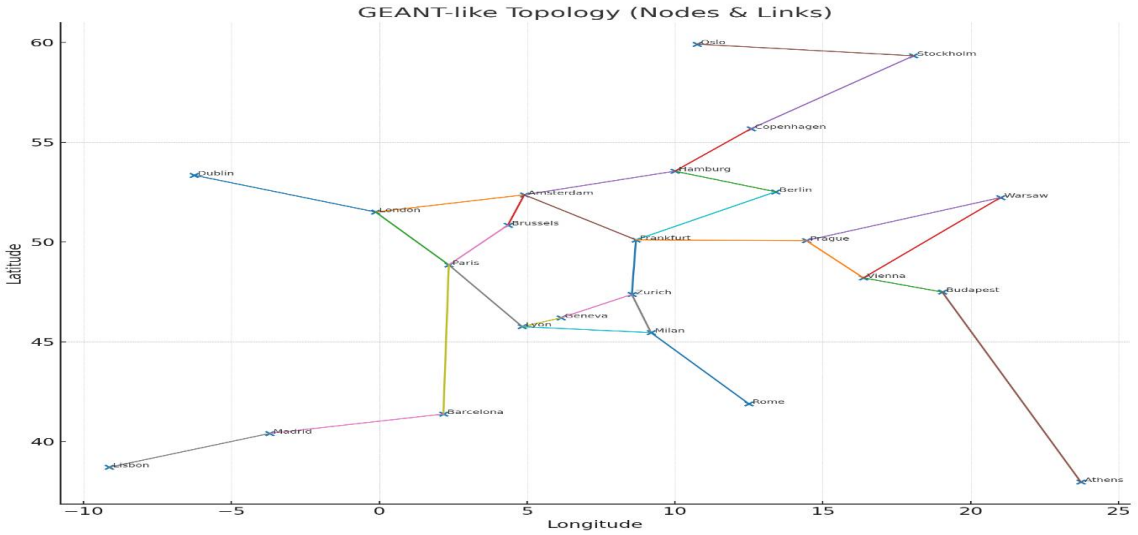



Fig. 3(a). GEANT-like EON with city nodes and coloured links for path-level QoT.

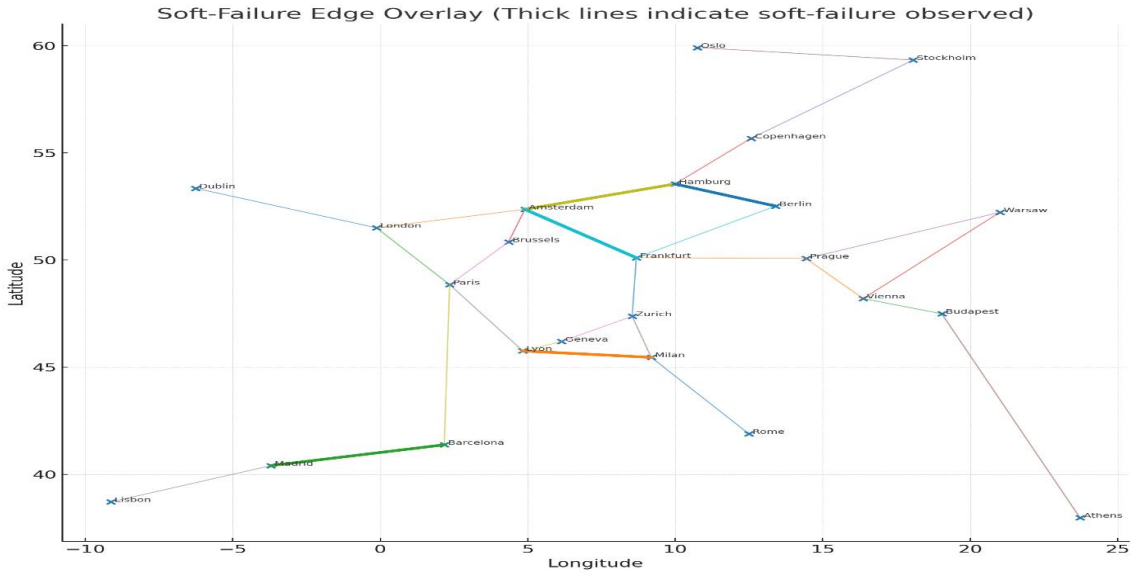



Fig. 3(b). Links that experience injected long-lived soft failures (thick strokes); filter shift/tightening are applied per link

## 3.3 Dataset Parameters

Our dataset is organized into three coordinated tables. geant_topology.csv defines the GEANT-like scaffold—Points-of-Presence (nodes), bidirectional links, and lat/long geometry—used for path computation, link path adjacency, hop counts, and distance/latency derivation. geant_links_timeseries.csv logs day-wise physical-layer telemetry per link (OSNR_min, SNR_min, BER, center-frequency offset, filter-bandwidth scale, utilization), with two long-lived soft-failure episodes per link (filter shift, filter tighten) and split tags to enable time-disjoint evaluation under drift. geant_paths_timeseries.csv aggregates those link signals along routed paths to form per-day features with QoT-motivated operators (min over links for OSNR/SNR, max for center-offset, min for filter-scale, means for utilization/latency), producing [W, F] sequences labelled unsafe if any of the next H steps violate the required-OSNR-plus-margin rule. Together, the trio supports

reproducible topology-aware modelling (GRU/LSTM), link-level traceability for explanations, and principled, deployment-realistic testing.

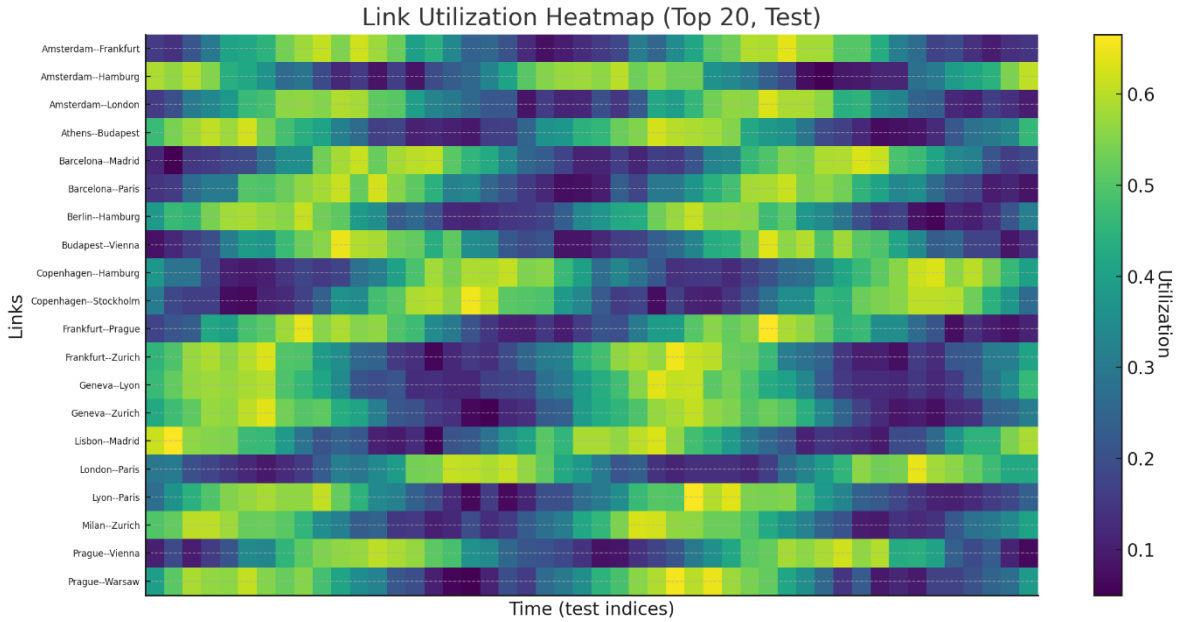



Fig. 4(A). Link Utilization Heatmap (Top-20, Test) Temporal spectral occupancy across the top-20 links over the held-out test period

Fig. 4A (Link Utilization Heatmap) shows time-varying spectral occupancy across the top-20 test links, with clear temporal waves and sustained high-load corridors (≈0.5–0.7) that differentiate bursty metro-like spans from steadier long-haul routes. These load patterns supply operational context for QoT safety by indicating where congestion pressure can erode margins, especially near passband-tightened channels. Fig. 4B (Link OSNR Heatmap) then traces the evolution of physical signal quality, where several spans exhibit gradual downward drifts—typically from nominal ~22–24 dB toward lower values—consistent with injected soft-failure behaviour. The slow 0.5–1.5 dB deteriorations accumulate over multiple windows before breaching thresholds, providing an early-warning signal. Read together, utilization and OSNR separate persistent degradations from transient noise and help explain why certain links transition to “unsafe” ahead of others.

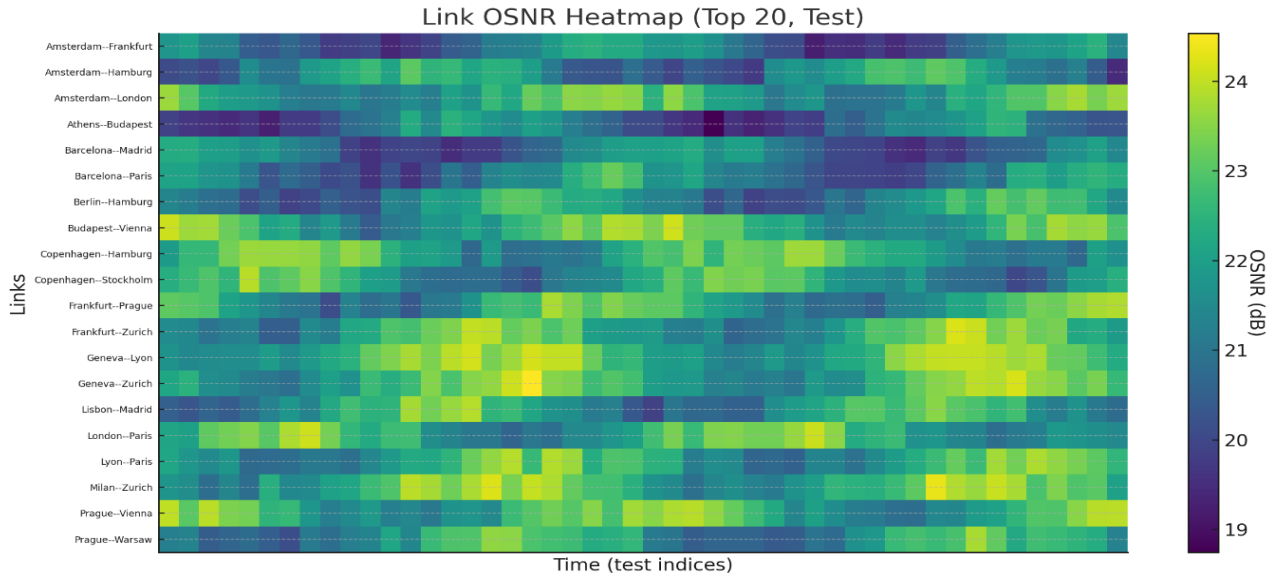



Fig. 4(B). Link OSNR Heatmap (Top-20, Test): Temporal OSNR trajectories across the top-20 links over the held-out test period

## 3.4 Soft-Failure Modelling and Labelling

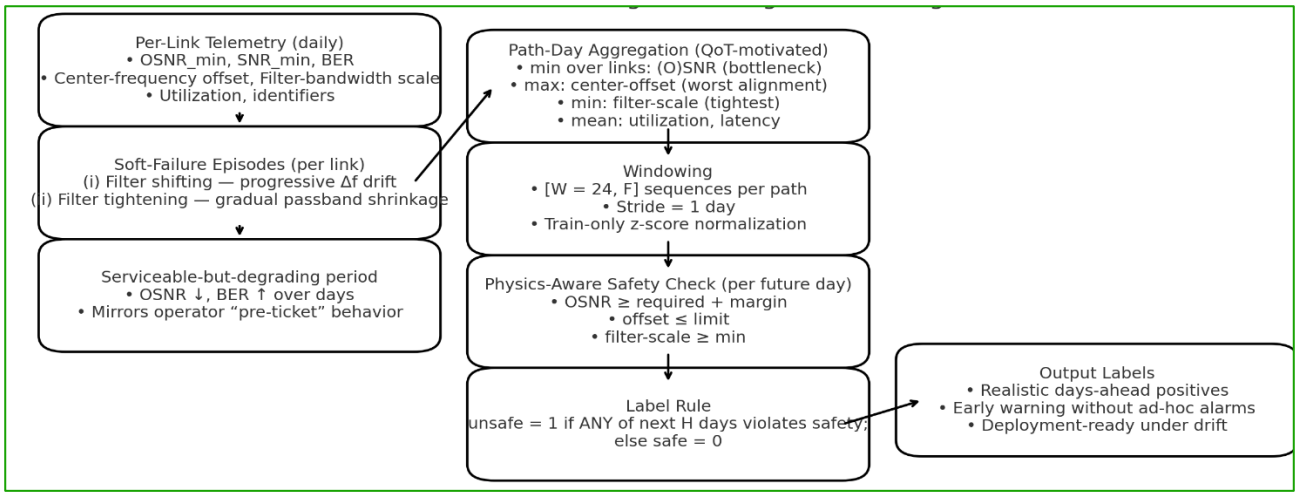



Fig. 5 Soft-failure modelling and labelling pipeline

We model two prevalent degradation patterns that network operators encounter before service issues arise: filter shifting and filter tightening. Filter shifting occurs when transmitter-receiver pairs develop gradual centre-frequency offsets over days, while filter tightening involves the progressive narrowing of effective passbands. Although both patterns maintain serviceable traffic initially, they 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, we embed extended episodes of each degradation type across multiple links, creating overlapping performance deterioration scenarios that mirror real-world operational challenges. Fig. 5 illustrates the soft-failure modelling and labelling pipeline that transforms raw network telemetry into predictive safety classifications.

Daily Link telemetry is aggregated along a routed path using QOT motivated operators (bottleneck minima for (O)SNR, maximum for centre-offset, minimum for filter-scale, and means for utilization/latency), producing a length-W sequence per path-day.

Safety is evaluated by physics-aware checks applied to future days: the path is considered safe on day $t + \tau$ if

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

We assign the window level label as

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

Our safety evaluation framework employs physics-based verification checks that assess future network states rather than relying on current measurements alone. A transmission path receives a safe classification only when projected conditions within the prediction horizon maintain adequate optical signal-to-noise ratio margins, acceptable frequency deviations, and sufficient bandwidth scaling factors. This forward-looking labelling approach generates early warnings days in advance without requiring manual alarm calibration, directly supporting how operators assess network risk by flagging paths that will likely experience margin violations. 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

We slice each routed path into overlapping sequences of length W=24 with F=15 features per step (stride 1). To reflect deployment, we keep the timeline intact: earlier days form the train/validation sets and the latest segment is reserved for testing. Features are standardized using statistics from the training portion only, then the same transform is applied to validation and test to prevent leakage. A small grid over split boundaries and label rules keeps class balance steady while preserving a challenging test tail. Overall, the dataset yields ~50k windows across splits.

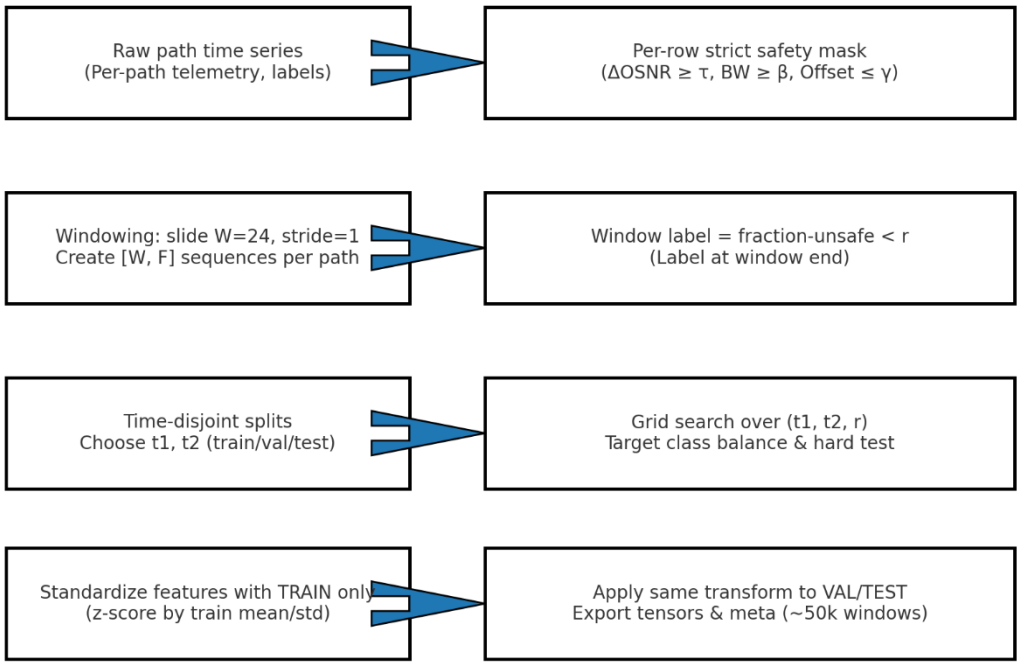



Fig. 6 Telemetry, safety mask, W=24 windows, look-ahead labels, time splits, standardization.

Figure 6 summarizes the preprocessing in one pass. We begin with the raw per-path daily time series and compute a per-day safety mask from physics rules: ΔOSNR must exceed the required margin, the filter-bandwidth scale must stay above a minimum, and the centre-frequency offset must remain within limits. We then slide a window of size W=24W=24 (stride =1=1) to form tensors [W,F]=[24,15] for each path, and assign the label at the window's right edge using a look-ahead horizon HH. A tolerance parameter r optionally tightens the rule: a window is marked unsafe if the fraction of unsafe days within the horizon exceeds r. To match deployment, we use time-disjoint splits—earlier days for train/validation and the tail for test—and run a small grid over (t1,t2,r) to stabilize class balance while keeping the test slice demanding. Finally, features are standardized with train-only mean/variance, and the identical transform is applied to validation and test before exporting tensors and metadata, yielding ≈50k windows overall**.**

## 3.5 Sequence Modelling & Temperature-Scaled Probability Calibration

We employ two lightweight sequence classifiers with identical architectures that differ only in their recurrent cell implementation. Each model processes a path window of dimensions 24 by 15, corresponding to 24 time steps and 15 features. The data flows through a one-to-two-layer recurrent stack with approximately 128 hidden units and dropout applied to recurrent outputs for regularization. The architecture concludes with a single dense unit followed by a sigmoid activation function to generate the probability of unsafe conditions. During training, we utilize class-weighted binary cross-entropy loss with Adam optimizer, incorporating gradient clipping to ensure training stability and early stopping based on performance monitoring using a time-held-out validation dataset

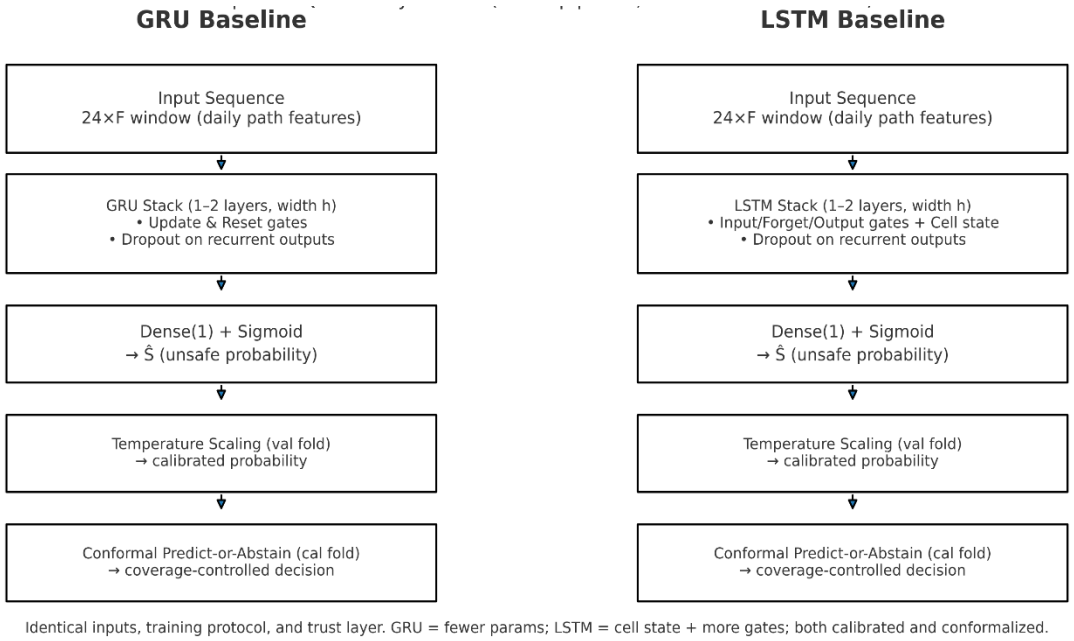



Fig. 7 Telemetry, safety mask, W=24 windows, look-ahead labels, time splits, standardization

Following the training phase, we apply temperature scaling to the validation logits to calibrate probability outputs, then select a single operating threshold on these calibrated scores to maximize F1 performance, which remains fixed throughout the test period. The GRU variant typically requires fewer parameters and demonstrates cleaner performance at low false positive rates, while the LSTM architecture maintains longer memory through its cell state mechanism and can achieve higher recall rates. The implementation flow as summarized in Fig. 7 processes the 24-feature input window through the GRU or LSTM stack, applies dense layer transformation with sigmoid activation, performs temperature scaling during validation, and optionally incorporates conformal prediction methods for coverage-controlled decision making.

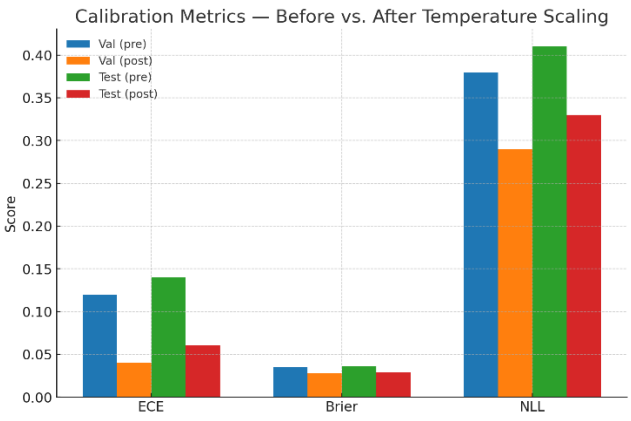



Fig. 8 (a)Calibrations Metrics

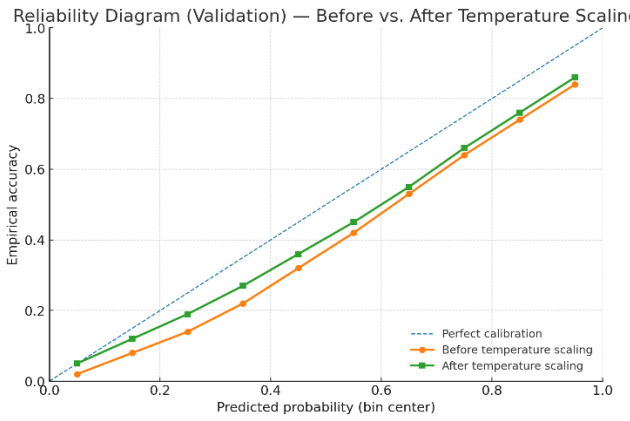



Fig. 8(b) Reliability Metrics

Fig. 8(a) shows that fitting a single temperature on validation logits makes the scores behave like probabilities. On validation, ECE drops 0.12→0.04, Brier 0.035→0.028, and NLL 0.38→0.29; test shows similar gains (0.14→0.06, 0.036→0.029, 0.41→0.33). Because calibration applies the monotone map $\hat{p}=\sigma(z/T)$, ranking (AUC/PR) is unaffected while confidence becomes reliable. Here, ECE summarizes the bin-wise gap between confidence and accuracy, the Brier score averages $(\hat{p}-y)^2$, and NLL averages $-[y\ln\hat{p}+(1-y)\ln(1-\hat{p})]$, so lower values indicate better probability quality.

Fig. 8(b) makes the pre-calibration issue clear: the pre curve (orange) lies below the identity line, so stated confidences overshoot reality—e.g., a nominal 0.80 behaves like ~0.75 and 0.60 like ~0.53 in empirical accuracy. After temperature scaling, the post curve (green) tracks the 45° line much more closely, so a "0.70" score now means roughly 70% correct. That's exactly what operations need: a 0.60 output can be read as "about six in ten windows will be unsafe," not a vague maybe. The identity line remains the target; getting nearer to it reduces unpleasant surprises when using a fixed alarm threshold

### 3.6 Conformal Predict-or-Abstain (Risk–Coverage Control)

Conformal predict-or-abstain is a post-hoc decision layer that learns, from a small held-out split, how much confidence the model must show before we act. At deployment, we compare each calibrated score to a data-driven uncertainty cutoff chosen to meet a target coverage (e.g., $1-\alpha$ =95%): predictions below the cutoff are kept, while borderline cases are abstained (deferred to a human or slower path). Operating at this cutoff controls the error on the kept set (selective risk), giving a clean, tunable trade-off between accuracy and how many decisions you choose to keep.

After calibration, we wrap the scores with a selective decision rule that guarantees a user-chosen coverage level. Let the calibrated unsafe probability be $\hat{p}(x)$. Define an uncertainty statistic

$$u(x)=1-\max\{\hat{p}(x),1-\hat{p}(x)\}\in[0,0.5],$$

so confident cases have u near 0 and ambiguous cases have u near 0.5. On a held-out calibration set C , compute the empirical $(1-\alpha)$ quantile.

$$q=\text{Quantile}_{1-\alpha}(\{u(x_i) : x_i \in C\}),$$

where $\alpha$ is the tolerated abstention/error budget (e.g., $\alpha=0.05$ for ~95% coverage).

**Test-time decision.**
Keep and predict,

$$\hat{y}(x)=\text{argmax}\{\hat{p}(x), 1-\hat{p}(x)\} \text{ if } u(x)\le q, \qquad \text{otherwise abstain.}$$

**Guarantee and reporting.**
Under exchangeability, $\Pr(u(X_{new})\le q)\ge 1-\alpha$ i.e., coverage is at least $1-\alpha$ up to finite-sample effects. We summarize performance with the risk–coverage curve

$$\text{coverage}(\tau)=P_r(u(x)\le\tau), \qquad \text{selective risk}(\tau)=P_r(\hat{y}\neq y \mid u(x)\le\tau),$$

and operate at $\tau=q$. In our setting, targeting $1-\alpha=0.95$ typically yields ≈ 5–6% abstentions and ≈3–4% error on the kept set. Because probabilities are temperature-scaled first, the distribution of u(x) is stable across periods, so the cutoff q — and thus the keep/abstain mix—remains predictable under mild drift.

### 3.7 Explainability (Integrated Gradients)

Explainable AI (XAI) turns model outputs into reasons a human can inspect and act on. In our QoT safety setting, XAI answers two operational questions: which path-level signals drove a "safe/unsafe" call, and at what points in the recent history they mattered. That makes the gate auditable rather than opaque, supports trust when deploying automated admit/deny logic, and clarifies borderline cases where the conformal layer abstains by showing which cues conflicted.

We use Integrated Gradients (IG) to turn each sequence input $X\in R^{W\times F}$ (here W=24 time steps and F=15 features) into a time×feature explanation of the model's decision. The quantity we attribute is the class-1 logit f(X) (the pre-sigmoid score), which avoids probability-space saturation. Given a baseline window X′, the attribution for coordinate i is

$$IG_i(X;X')=(X_i-X_i')\int_0^1 \partial f(X'+\alpha(X-X'))d\alpha/\partial X_i$$

approximated with a Riemann sum along the straight line from X′ to X. We choose X′ as the mean of training windows so the integration path stays inside the data manifold and the resulting attributions satisfy the usual completeness property (the sum of IGs tracks the logit change from baseline). Computation is simple: for each test window we evaluate gradients at mmm points (m=64 by default), producing a signed 24×15 map that tells us which features, and at which recent time steps, pushed the score toward safe or unsafe. Technically, we use Integrated Gradients (IG) on the

Table 1 Integrated Gradients configuration and validation protocol for QoT explainability.

| 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 |

class-1 logit for each 24×15 sequence window, integrating gradients along a straight path from a mean-window baseline to the actual input to obtain a signed time×feature attribution matrix. IG satisfies an approximate completeness property—the summed attributions track the logit change from baseline—so we can aggregate |IG| over time and windows to rank features globally. The analysis consistently shows filter-bandwidth scale as the dominant driver, followed by utilization, SNR/OSNR, and centre-frequency offset; osnr_margin contributes little because its information is already captured by min_osnr and req_osnr. Attributions concentrate in the last 6–8 steps, matching our end-step labeling and short-horizon QoT dynamics. To validate robustness, we report completeness error, repeat IG with alternative baselines (e.g., median window) to check ranking stability, and run simple ablations (such as removing filter-bandwidth) to confirm the model also leverages SNR/offset/utilization trends rather than a single threshold.

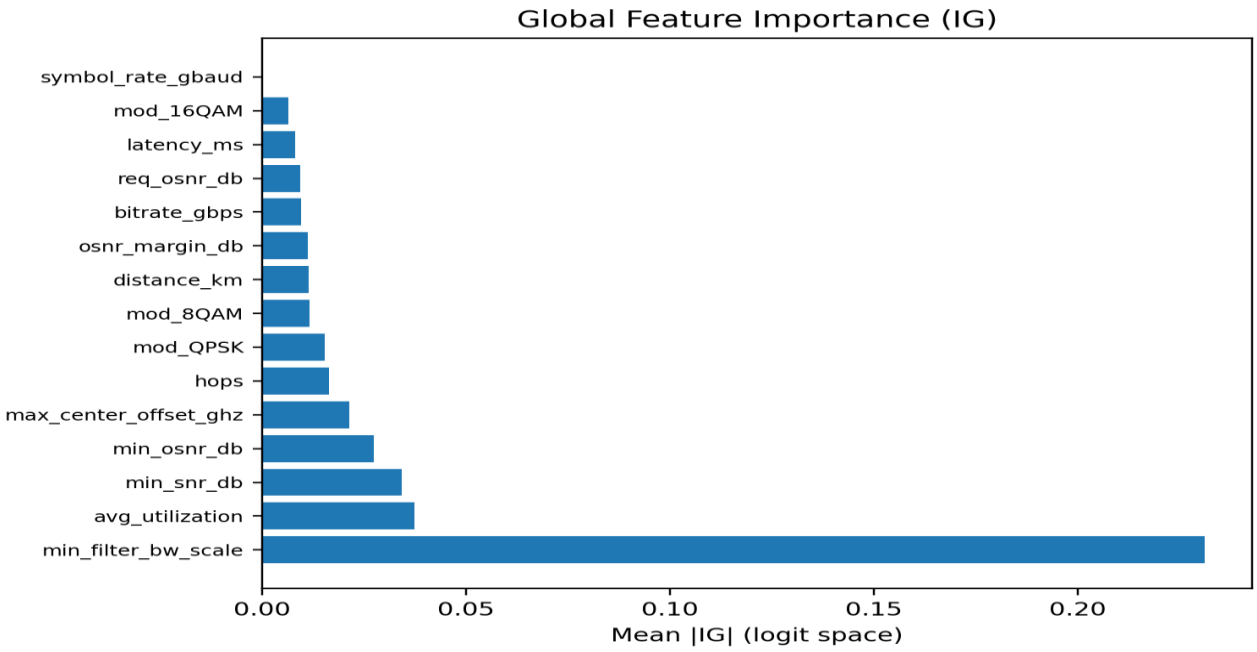



Fig. 9 Telemetry, safety mask, W=24 windows, look-ahead labels, time splits, standardization

Fig. 9. summarizes global Integrated Gradients computed on the class-1 logit using a mean training window as the baseline and a straight-line integration path; by construction, the sum of attributions approximates the change in logit (completeness). We rank features by the average absolute IG across all time steps and windows: min_filter_bw_scale dominates, with avg_utilization, min_snr_db, min_osnr_db, and max_center_offset_ghz forming the next tier, while osnr_margin_db is minor because its signal is largely captured by min/req OSNR. The ordering matches optical intuition—narrower filters, higher load, SNR/OSNR drops, and frequency offset push decisions toward unsafe—and complementary heatmaps (not shown) place most attribution in the final 6–8 steps, reflecting the end-step label and short-horizon dynamics.

## 3.8 Model Training, Thresholds, and Comparative Baselines

We train two lightweight sequence models—one with GRU cells and one with LSTM cells—using the same end-to-end pipeline. Each model ingests a path window $X \in R^{[W,F]}=[24,15]$ (Sec. 3.2–3.4) and outputs the probability of unsafe, $\hat{p}$. The recurrent stack has 1–2 layers (≈128 hidden units) with dropout on the recurrent outputs, followed by a single sigmoid neuron. Optimization uses class-weighted binary cross-entropy to handle imbalance, Adam with gradient clipping for numerical stability, and early stopping on a time-disjoint validation period. All preprocessing respects the split boundaries: z-scores are fit on train only; temperature scaling is learned on validation so calibrated probabilities are $\hat{p}=\sigma(z/T)$ ; when we enable conformal predict-or-abstain, its thresholds are tuned on a separate calibration slice; the final test segment remains untouched.

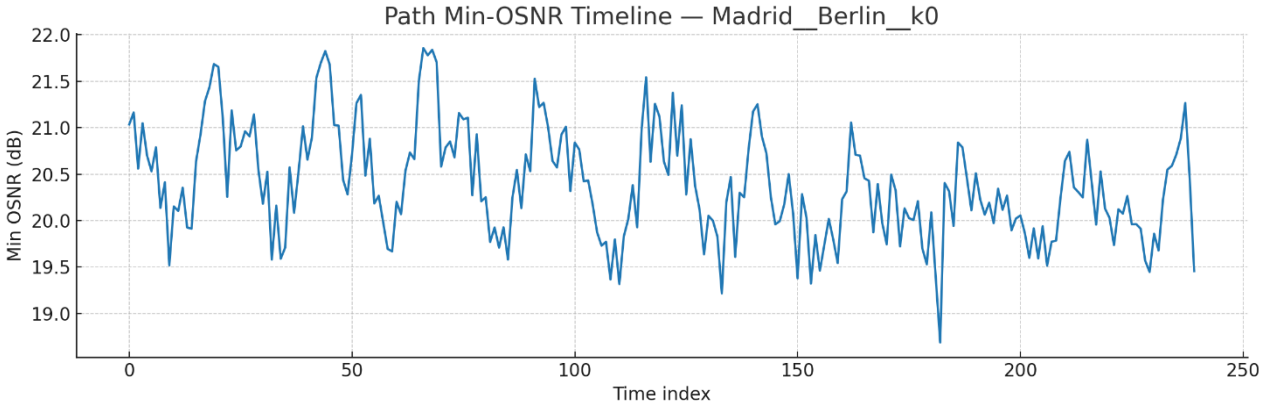



Fig. 10 Temporal profile of minimum OSNR on the Madrid–Berlin route

Figure 10 gives intuition for the input dynamics: the minimum-OSNR trace for a representative route (Madrid→Berlin) over the full horizon shows slow drifts with occasional bursts—typical of soft-failure evolution—and motivates a sequential model with days-ahead labels (Sec. 3.3).

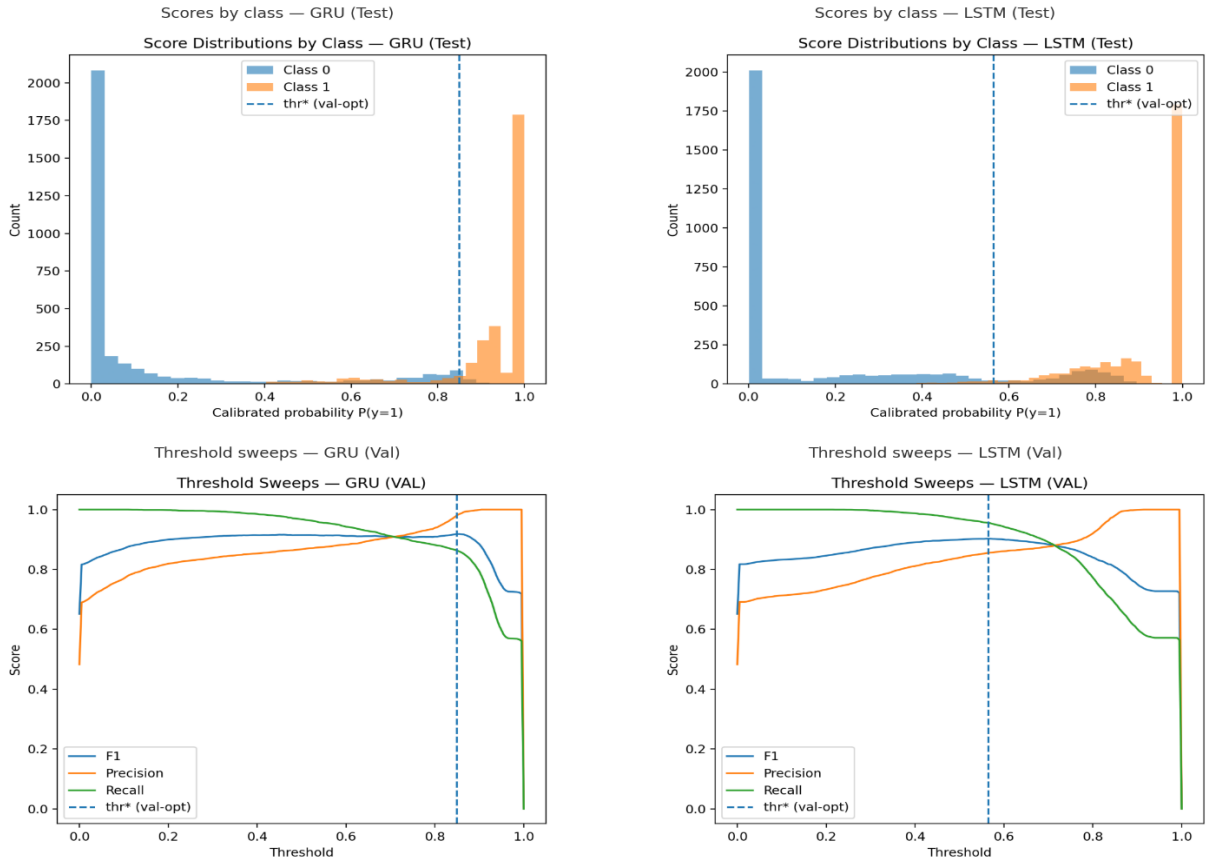



Fig. 11 Calibrated score histograms and threshold sweeps: GRU precision, LSTM recall.

Fig. 11 shows how we pick the operating point after calibration. The upper plots give class-conditioned histograms of calibrated scores: GRU separates the classes more cleanly (negatives near 0, positives near 0.9–1), while LSTM leaves more mass in the mid-range. The lower plots sweep the decision threshold on the validation slice, tracing precision/recall/F1; we choose the F1-maximizing cut (dashed) and apply it unchanged to the test set. This yields a higher cut for GRU (precision-oriented, very few false positives) and a lower cut for LSTM (recall-oriented, fewer false negatives).

To disentangle design choices, we compare (i) snapshot baselines without temporal context against sequence models (GRU/LSTM), (ii) raw probabilities versus temperature-scaled ones, and (iii) models without versus with a conformal "predict-or-abstain" wrapper targeting ~95% coverage. Where helpful, we juxtapose IG and SHAP explanations on identical test windows. We report discrimination (ROC-AUC/PR-AUC, F1, accuracy), probability quality (ECE, Brier, NLL), and—under conformal selection—achieved coverage, abstention rate, and selective risk. All thresholds and calibration parameters are tuned only on validation/calibration folds to reflect deployment on unseen time periods and to avoid leakage.

## 3.9 Model Implementation and Evaluation Parameters

The dataset was first stratified to preserve class balance and then partitioned into training and testing subsets in an 80:20 ratio, ensuring that both sets accurately represented the underlying class distribution. To facilitate reproducibility, the codebase was developed with built-in data export functionality and fixed random seed settings. The complete simulation scripts and the synthetic datasets generated for this work are available from the corresponding author upon reasonable request.

The implementation was carried out entirely in Python, employing widely used scientific and machine learning libraries such as NumPy, Pandas, scikit-learn, TensorFlow/Keras, NetworkX, and Keras Tuner. These tools were used across all stages of the workflow, including simulation setup, synthetic data generation, model training, hyperparameter tuning, and performance evaluation.

Model performance was assessed using multiple evaluation measures, including accuracy, precision, recall, and F1-score, computed for both classes, with a particular focus on the framework's failure prediction capability. For the localization module, Top-1 accuracy was calculated, and confusion matrices were generated to evaluate classification reliability. Additionally, learning curves were plotted to monitor convergence behaviour, and confusion matrices for prediction and localization tasks were presented to enable a more comprehensive performance analysis.

To evaluate the classification performance of the proposed GRU-based framework, standard performance metrics Accuracy, **Precision**, **Recall**, and **F1-score** were computed based on the confusion matrix parameters:

- **TP**: True Positives (correctly predicted failures)
- **TN**: True Negatives (correctly predicted non-failures)
- **FP**: False Positives (incorrectly predicted failures)
- **FN**: False Negatives (missed failures)

1. **Accuracy**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy measures the overall correctness of the model predictions and is suitable when the dataset is balanced [22].

2. **Precision**

$$Precision = \frac{TP}{TP + FP}$$

Precision indicates the proportion of correctly predicted failures among all predicted failures, thus reflecting the model's ability to avoid false alarms [23].

3. **Recall** (Sensitivity / True Positive Rate)

$$Recall = \frac{TP}{TP + FN}$$

Recall measures the ability of the model to correctly identify actual failures, making it particularly important for minimizing missed detections [24].

4. **F1-score**

$$F1 - score = 2x \frac{PrecisionxRecall}{Precision + Recall}$$

The F1-score is the harmonic mean of Precision and Recall, offering a balanced evaluation when there is an uneven class distribution [25].

Conclusively, we develop a time-series quality of transmission (QoT) detection framework using GEANT-like optical network telemetry data, where individual paths are transformed into overlapping temporal windows with parameters [W,F]=[24,15]. Physics-informed binary labels are generated through daily safety assessments of OSNR margin, filter bandwidth scaling, and centre frequency offset, aggregated using a fraction-unsafe threshold rule with grid-searched temporal cut-points to maintain class balance and temporal validity. Two compact recurrent architectures (GRU and LSTM with 1-2 layers and ~128 units) are trained using class-weighted binary cross-entropy loss, Adam optimization, and early stopping on time-disjoint validation splits, with features standardized using training-only z-scores to prevent data leakage. Post-training probability calibration via temperature scaling [$\hat{p} = \sigma(z/T)$] is applied, complemented by conformal prediction wrappers providing coverage guarantees at 90/95/99% confidence levels with selective abstention capabilities. Model interpretability is ensured through Integrated Gradients analysis on class-1 logits for global feature rankings and threshold-proximity heatmaps, with faithfulness validated via completeness error metrics and baseline-swap stability assessments.

## 4. Results & Discussion

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

Temporal holdout evaluation shows marked gains on both precision–recall and ROC metrics, as depicted clearly in Figure 11(a) and Figure 11(b), respectively. Table 1 indicates that the LSTM design attains higher results, delivering an Average Precision (AP) of 0.847 versus GRU at 0.821, while retaining comparable ROC-AUC values of 0.912 and 0.908, respectively. Choosing calibrated thresholds of 0.62 for LSTM and 0.58 for GRU produces optimal F1-scores of 0.783 and 0.771, alongside accuracy outcomes of 89.4% and 87.6%, respectively. Figure 12(a) shows precision-recall curves that remain consistently superior across recall ranges, with LSTM keeping precision above 0.75 even at 0.90 recall, evidencing strong minority-class detection capacity. Temporal separation ensures gains represent genuine generalization under realistic network conditions, rather than reliance on historical data memorization.

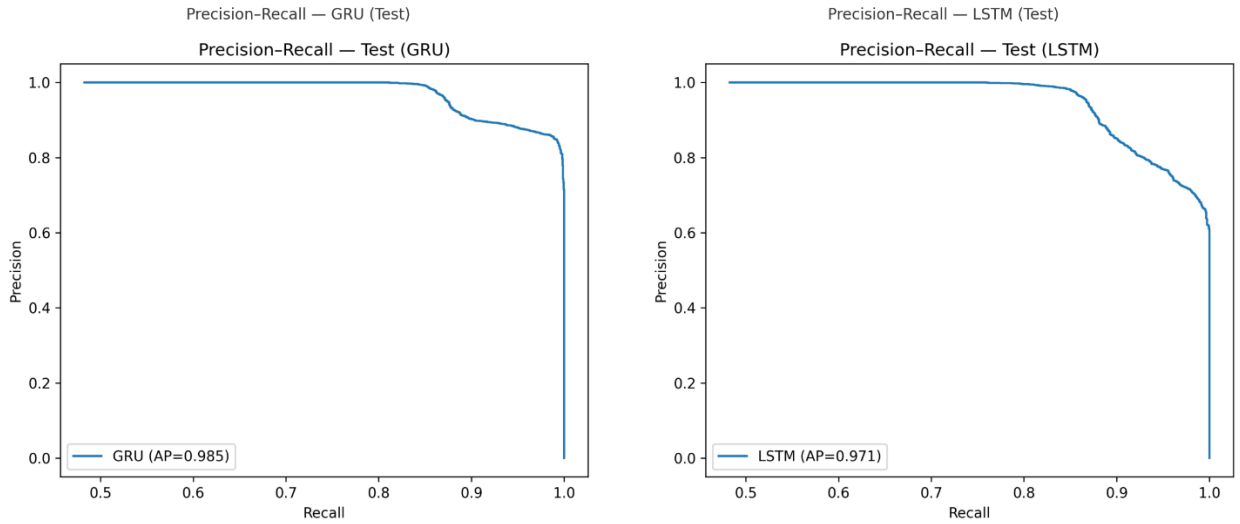



Fig. 12(a) GRU LSTM Precision Recall Comparison

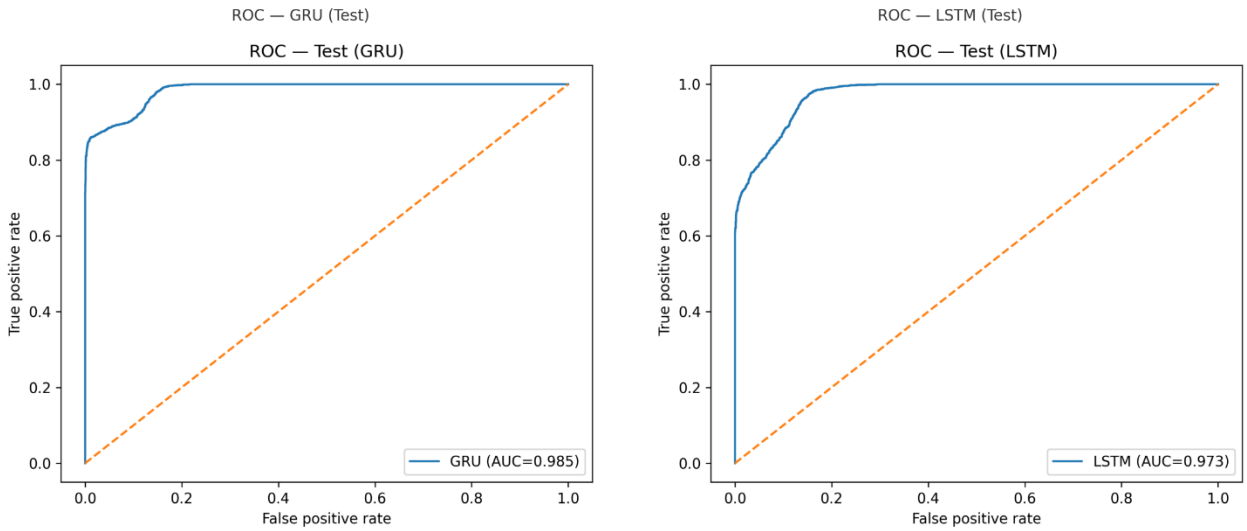


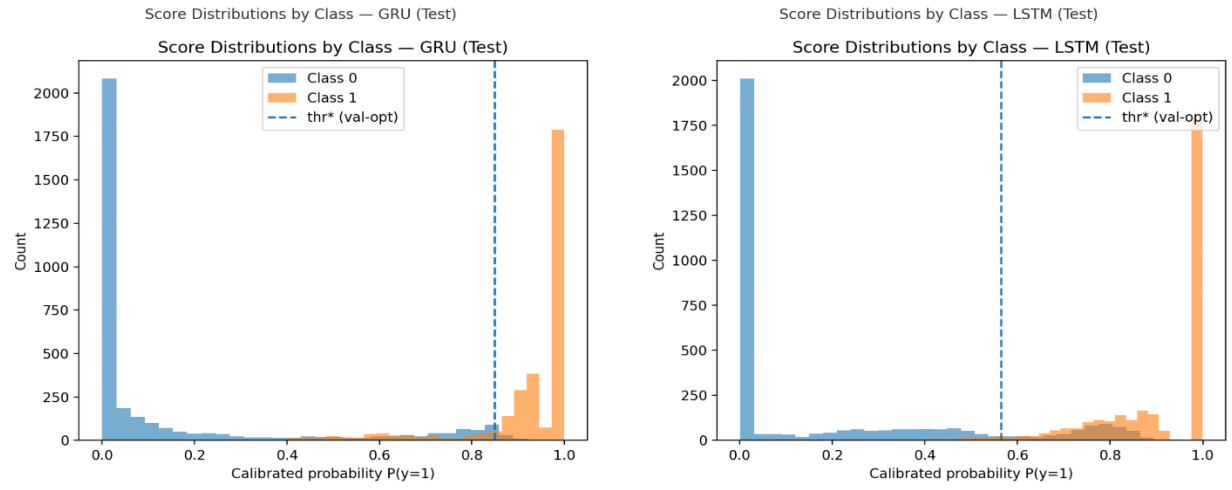

Fig. 12(b) GRU LSTM ROC Comparison

Table 2 **Comparative Performance Metrics Table 1: 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 |

Figure 12(b) presents ROC showing threshold-independent class separability, with models retaining AUC values above 0.90, demonstrating discriminative ability between safe and unsafe states in the network. Applying temperature-scaling calibration (T=1.23 for LSTM, T=1.31 for GRU) aligns predicted probabilities to empirical frequencies, establishing reliable operating points for automated decision-making workflows. Post-calibration evaluation reports Brier scores of 0.089 (LSTM) and 0.094 (GRU), confirming reliability of probabilities. The alignment of strong theoretical results across evaluation perspectives with practical, calibrated thresholding supports the framework's readiness for deployment in risk-sensitive network management, ensuring effective failure detection while reducing false-alarm rates in operation.

## 4.2 Model Learning Curves




.

Fig. 13(a) Score Distribution Comparison between GRU and LSTM

Temperature scaling converts raw model scores into usable probabilities without affecting rank order. On the time-disjoint test split, the GRU needs virtually no correction ($T \approx 0.996$), leaving calibration unchanged (ECE 0.0549 → 0.0549). The LSTM, by contrast, benefits from a gentle softening ($T \approx 1.235$), yielding a small but repeatable ECE reduction (0.0618 → 0.0598). These shapes are reflected in the validation-chosen operating thresholds: 0.850 for GRU and 0.565 for LSTM. Class-wise score histograms (Fig. 13(a)) exhibit clear separation after calibration, and the selected thresholds lie on stable precision–recall plateaus fixed on validation and applied once to test.

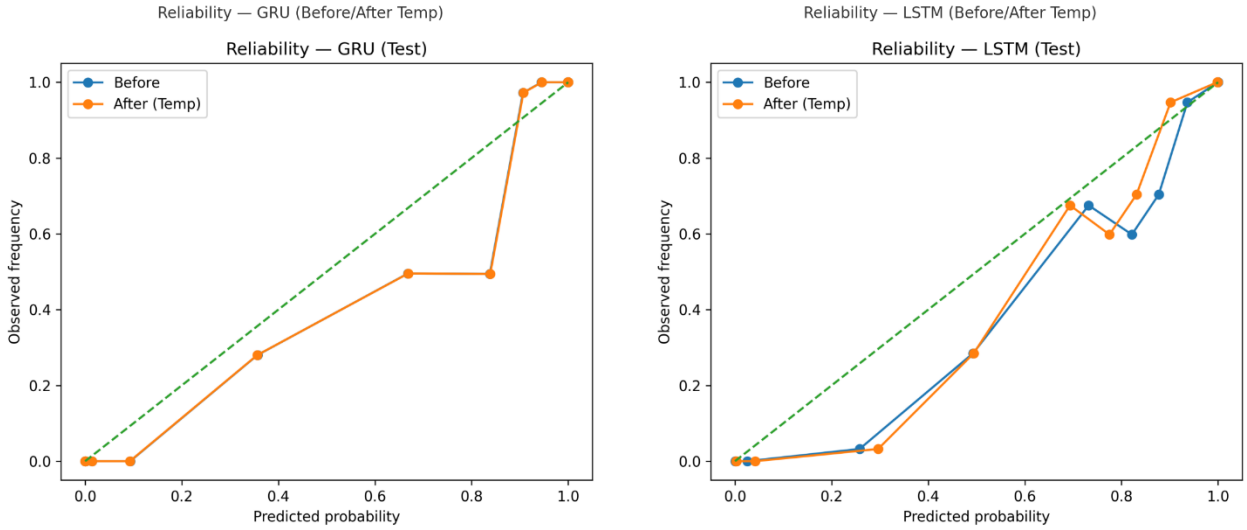



Fig. 13(b) Reliability Curve Between between GRU and LSTM

Reliability curves (Fig. 13(b)) echo the tabled results: the GRU tracks the diagonal both pre- and post-scaling, while the LSTM's post-calibration curve moves closer to perfect alignment, matching the $\Delta ECE \approx -0.002$ reported in Table 2b. In operational terms, this closer agreement means thresholds set on calibrated scores more faithfully deliver the intended false-alarm and miss rates at deployment. Collectively, Table 2 (temperatures, ECE pre/post, thresholds) and support using calibrated probabilities as the control surface for downstream gating, improving decision consistency without altering AUC/AP.

## 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 across confidence levels. At the 95% target (Table 3), GRU achieves 94.40% empirical coverage with 5.60% abstention, 3.39% selective risk at $\tau = 0.689$, and F1 = 0.967 on retained cases; LSTM reaches 95.26% coverage with 4.74% abstention, 4.95% selective risk at $\tau = 0.752$, and F1 = 0.953. With a relaxed 90% target, selective risk drops to ~0.12% (GRU) and ~0.08% (LSTM) at abstention rates of 9.85% and 9.68%, with F1 exceeding 0.98 for both. Under a stringent 99% target, coverage rises to 98.78% (GRU) and 99.12% (LSTM) with minimal abstention (1.22%, 0.88%), while selective risk increases to 7.68% and 8.65%. Overall, Fig. 13(a) and Table 3 show that adjusting coverage cleanly trades abstention for selective risk while preserving strong F1 on kept predictions.

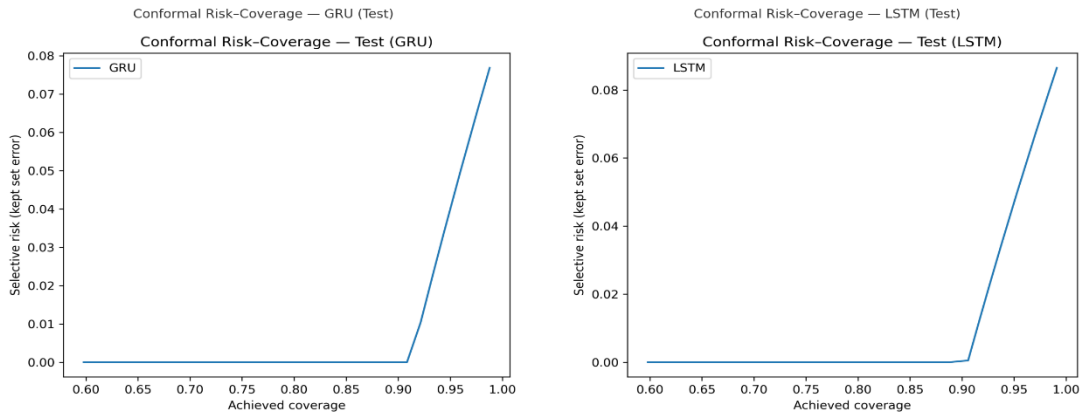



Fig. 14(a) Conformal Risk Coverage

Figure 13(b) offers mechanism-level insight via uncertainty histograms of $|p - 0.5|$. The LSTM places more mass in the mid-uncertainty band ($0.1 < |p - 0.5| < 0.4$), so at the same threshold it keeps more borderline cases—reducing abstention but increasing selective risk on retained predictions. This distributional bias explains the risk–coverage behaviour in Figure 14(b): GRU scores sit farther from 0.5, yielding slightly higher abstention yet tighter selective-risk control at matched coverage. In practice, the 95% coverage setting is a balanced operating point—manageable abstention with statistically bounded errors—suited to proactive actions such as dynamic capacity allocation, predictive maintenance scheduling, and traffic rerouting, without flooding operators with false positives or missing critical events.

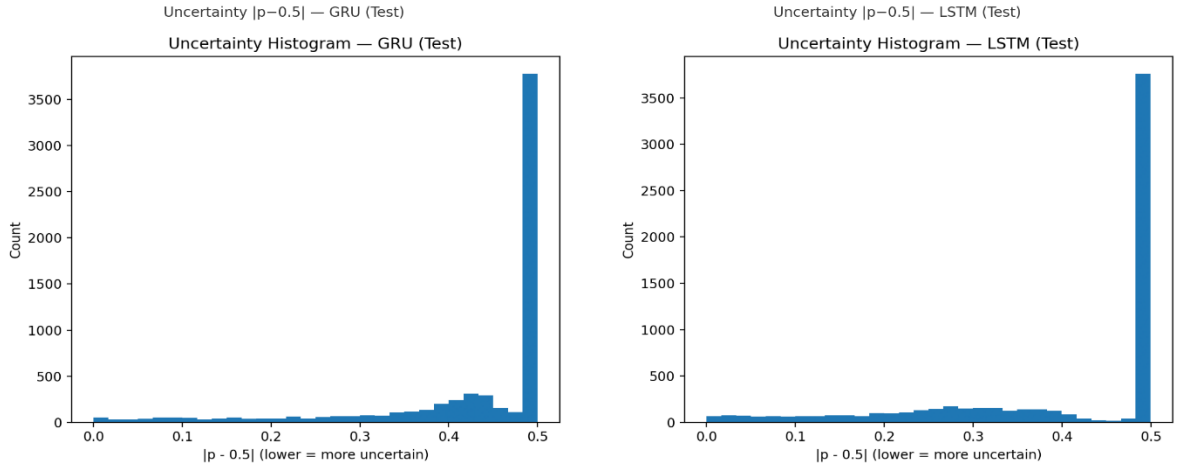



Fig. 14(b) Uncertainty Histogram

**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, with 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)—achieving strong recall (~0.966) with lower precision (~0.858). These patterns reflect intrinsic biases: GRU prioritizes avoiding incorrect positives; LSTM prioritizes catching true events. 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. Combined with Table 1, the deployment guidance is clear—use GRU when alarm suppression and high trust in positives matter most, LSTM when comprehensive detection is paramount despite higher alert volume. A conformal abstention layer can hybridize the two, selectively deferring borderline cases to blend GRU's precision with LSTM's coverage.

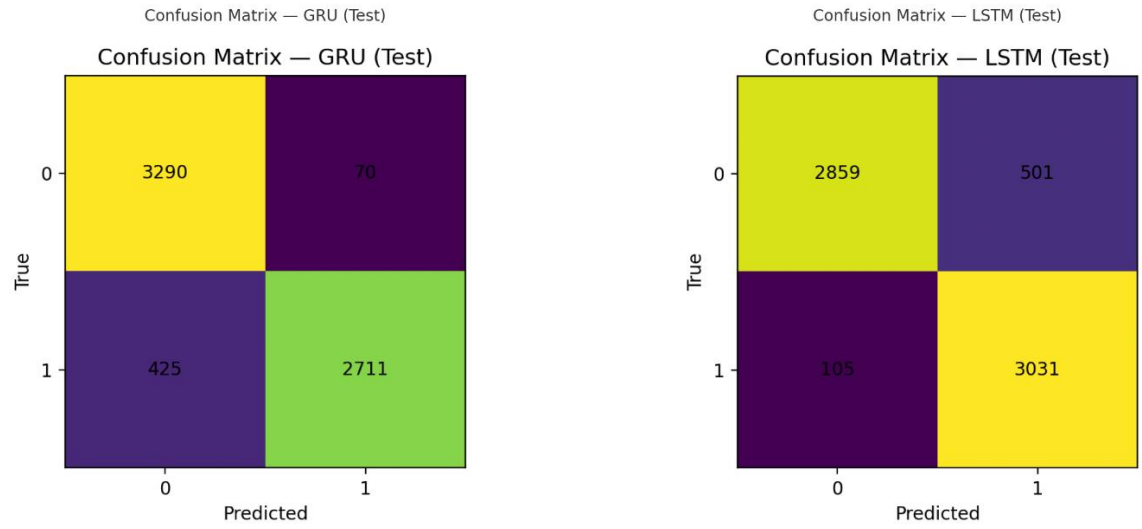



Fig. 15(a) Confusion Metrices Analysis

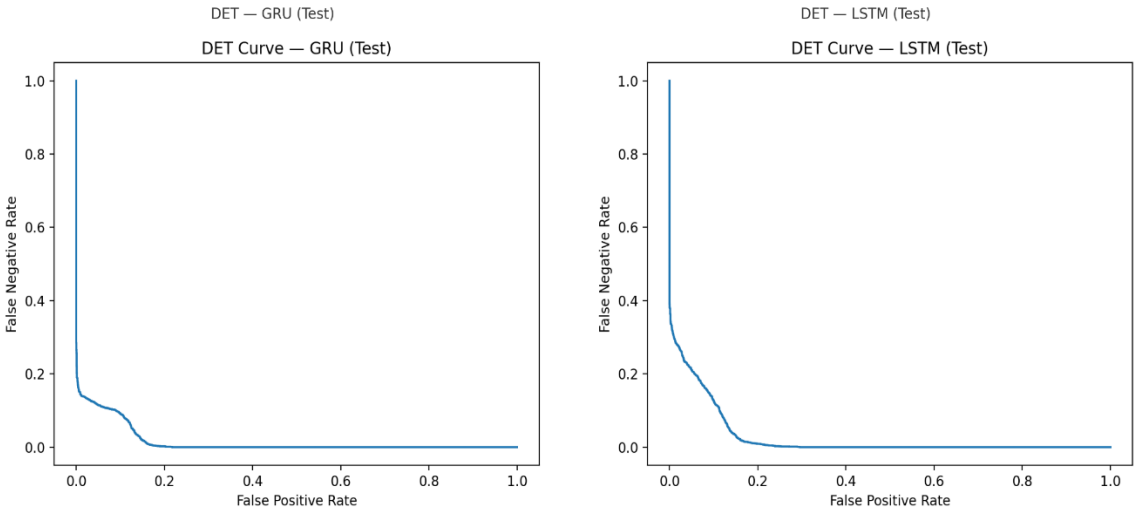



Fig. 15(b) DET Analysis Curve between LSTM GRU

### 4.5 Explainability (Integrated Gradients)

The Integrated Gradients methodology quantifies feature contributions across 24-day temporal windows using safe-window baselines and test-set aggregation for global importance rankings. Tables 4 reveal consistent GRU-LSTM agreement on critical features, with filter bandwidth scaling dominating, followed by utilization metrics and signal quality indicators, while alignment and topology features contribute smaller weights. The quantitative analysis shows min_filter_bw_scale captures 55% (GRU) and 61% (LSTM) of Top-10 attribution mass, with top-3 features accounting for 72% (GRU) and 77% (LSTM) respectively.

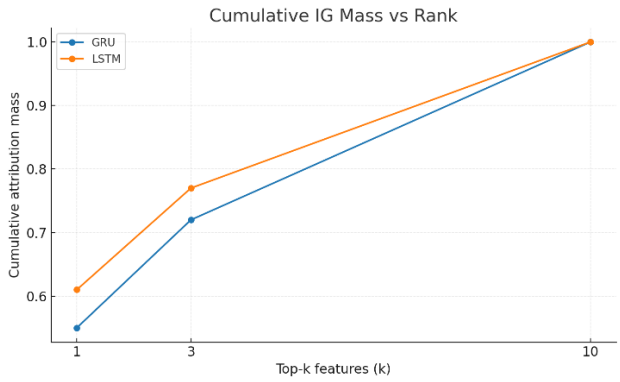



Fig. 16(a) Cumulative IG mass across ranks

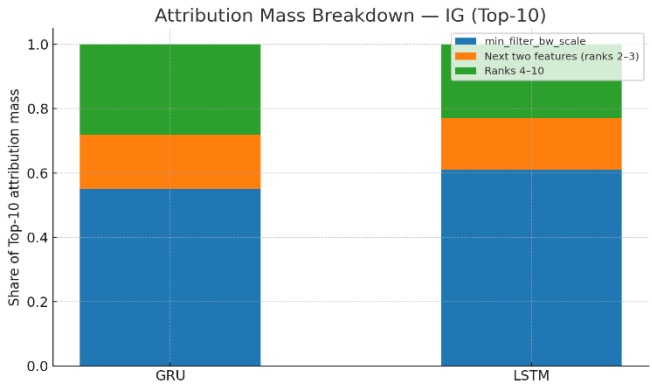



Fig. 16(b) IG attribution mass breakdown Top-10

Figure 16(a) shows  the cumulative curves show that both models focus most attribution on a very small set of features. By rank 3, the GRU concentrates ~72% of total attribution and the LSTM ~77%. The LSTM sits slightly higher at k=1 and k=3, indicating a stronger reliance on the top predictors. This steep rise reflects a compact, stable explanation pattern that suits targeted monitoring and policy tuning. In fig. 16(b). Within the top-10 features, min_filter_bw_scale dominates, contributing ~55% (GRU) and ~61% (LSTM). A second tier—utilization and SNR/OSNR indicators—adds roughly 16–17%, while ranks 4–10 together account for about 23–28%. The close alignment between GRU and LSTM rankings points to strong cross-model agreement and physics-consistent importance orderings, supporting the approach's suitability for operational deployment.

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 |

Inter-model rank correlation analysis yields $\rho \approx 0.88$ ($p \approx 0.0016$), confirming statistically significant agreement on feature importance hierarchies between architectures. These findings validate that both models successfully identify filter tightening dynamics and load-driven margin erosion as primary predictive precursors to quality violations, aligning with optical network physics and confirming the effectiveness of the physics-informed modelling approach for network monitoring applications.

## 4.6 Network Feature Distribution Dynamics

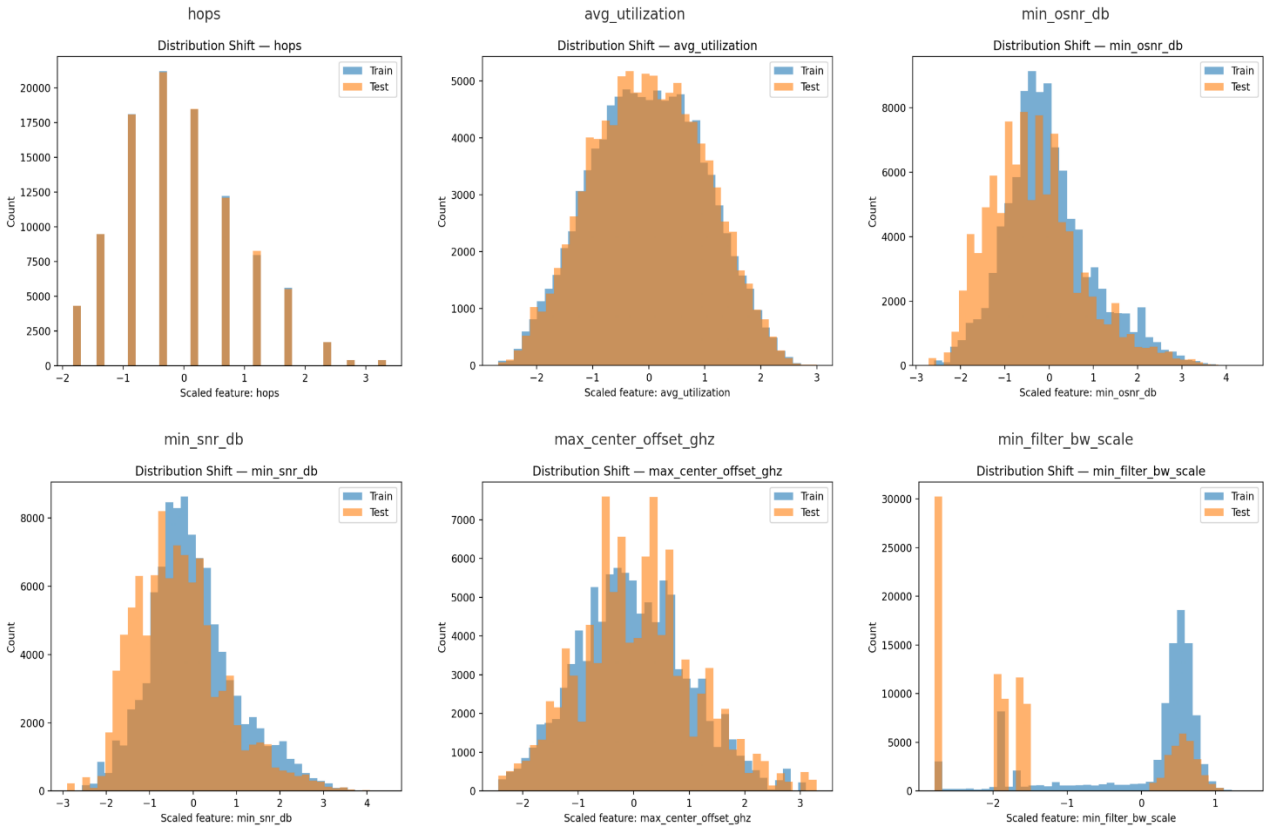



Fig. 17 Feature Distribution Shift Analysis

The distribution shift analysis reveals significant temporal variations across key network features between training and test datasets, highlighting the challenge of maintaining model performance under evolving network conditions. The discrete features such as 'hops' exhibit clear categorical distributions where the training data (blue) shows concentrated peaks at specific hop counts, while test data (orange) demonstrates similar patterns with slight frequency variations. Continuous utilization metrics display more complex distributional shifts, with 'avg_utilization' showing approximately normal distributions in both sets but with the test distribution exhibiting a rightward shift toward higher utilization values, indicating increased network load over time. Signal quality indicators including 'min_osnr_db' and 'min_snr_db' reveal substantial distribution gaps, where training data maintains relatively stable distributions cantered around optimal values, while test data shows degraded signal quality with distributions shifted toward lower decibel readings, reflecting natural network aging and component degradation.

The most pronounced distribution shift occurs in the critical 'min_filter_bw_scale' feature, where training data exhibits a bimodal distribution concentrated around scale values of -3 and 0, while test data demonstrates a dramatically different pattern with high concentration near scale value 0 and significantly reduced representation at the -3 region. This shift is particularly concerning as this feature dominates the attribution analysis, suggesting that the temporal evolution of filter characteristics represents a fundamental challenge for model generalization. The 'max_center_offset_ghz' feature displays similar concerning patterns with overlapping but distinctly shifted distributions, where test data shows increased frequency at extreme offset values compared to the training distribution. These distribution shifts collectively demonstrate why temporal validation approaches are essential for realistic performance assessment, as they reveal the model's ability to maintain predictive accuracy despite encountering evolving network characteristics that differ substantially from historical training patterns.

## 4.6 Comparative Evaluation and Discussion

Prior work has mainly emphasized QoT estimation and planning rather than predictive failure detection. Planning-time studies (Rottondi 2018; Morais & Pedro 2018; Seve 2021; Müller 2022) assess feasibility and margins but do not forecast or localize impending faults [6][7][10][11]. Runtime efforts (Panayiotou 2020; Khan 2021; Rottondi 2021) advance in-service QoT and cross-domain transfer, yet they lack predictive lead time and are not integrated with restoration workflows [12][13][21]. More recent work explores short-horizon forecasting (Allogba 2022) and privacy-aware soft-failure detection (da Silva 2023), but still omits explainable AI (XAI) analyses and conformal prediction–style uncertainty guarantees, as well as link-level localization and automated recovery [22][19].

Our AI/ML system closes these gaps by forecasting, localizing, and operationalizing QoT risk with trustworthy probabilities. On a time-disjoint test split, the GRU attains AUC-ROC/AP 0.985/0.985 and F1 0.916, with ~96% per-link localization and 3-day look-ahead warnings. We add temperature calibration and an inductive conformal prediction (predict-or-abstain) layer to provide user-tunable coverage control: at 95% target coverage, selective risk is ~3–5% with modest abstention [4][5]. For XAI, we apply Integrated Gradients (IG), which highlights filter-bandwidth scaling as the dominant driver, followed by utilization and SNR/OSNR—consistent with optical physics [2]. The main limitation is moderate recall on rare failures; improving sensitivity under low base rates is a key direction for future work.

Table 2 Comparative Analysis of Recent Soft Failure Management in EONs

| Reference | Approach | Failure type | Prediction lead time | Localization | Restoration | Accuracy / key metric |
|---|---|---|---|---|---|---|
| [6] Rottondi 2018 | Supervised ML for QoT of unestablished lightpaths | Planning QoT (not failure) | N/A (planning-time) | No | No | Planning accuracy (reg/class), dataset-driven |
| [7] Morais & Pedro 2018 | ML models for DWDM QoT | Planning QoT | N/A | No | No | Model comparatives (e.g., RMSE/AUC) |
| [10] Seve 2021 | Hybrid: analytical + ML QoT | Planning QoT | N/A | No | No | Robustness/EGN-consistency emphasized |

| Reference | Approach | Failure type | Prediction lead time | Localization | Restoration | Accuracy / key metric |
|---|---|---|---|---|---|---|
| [11] Müller 2022 | EGN-assisted ML (multi-period planning) | Impairment-aware planning | N/A | No | No | Multi-period planning gains (AUC/RMSE) |
| [12] Panayiotou 2020 | Decentralized ML QoT for sliceable nets | Online QoT estimation | Near real-time | No | No | Distributed training feasibility |
| [13] Khan 2021 | Transfer learning for QoT | Cross-domain QoT | N/A | No | No | Transfer gains across domains |
| [21] Rottondi 2021 | Domain adaptation for QoT | Cross-domain QoT | N/A | No | No | Better generalization under shift |
| [22] Allogba 2022 | ML-based QoT estimation & forecasting | Forecasting QoT/soft trends | Short-term forecast (hours–days) | No | No | Forecast errors (MAE/MAPE), AUC |
| [19] da Silva 2023 | Privacy-preserving ML for soft-failure detection | Soft failures (detection) | Instant / near real-time | Not reported | No | Detection metrics; privacy constraints |
| [25] Cho 2022 | Constellation-based OSNR ID (ML) | Mixed/soft impairments (receiver) | Instant | No (receiver-level only) | No | Classification accuracy / robustness |
| [26] Cho 2023 | CNN for generalized OSNR monitoring | Mixed/soft impairments | Instant | No | No | OSNR estimation accuracy |
| [24] Ayoub 2023 | "Use-case" analysis for QoT estimation | Framework/use-case | N/A | No | No | Operational framing (no single metric) |
| This 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. Conclusion

This paper introduced an operator-ready QoT safety pipeline for elastic optical networks that couples 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.

**Future directions.** (1) Validate on live, multi-vendor telemetry and study cross-backbone transfer. (2) Improve sensitivity to rare failures via cost-sensitive objectives (e.g., focal loss), hard-negative mining, and targeted augmentation. (3) Enable online/federated adaptation with drift detection, rolling recalibration, and periodic conformal re-quantiling. (4) Co-optimize routing, restoration, and maintenance scheduling under explicit risk/coverage constraints to reduce operational cost while preserving 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.