

Risk-Aware Machine Learning-Based Approach for Lightpath QoT Estimation in Optical Networks

Original

Risk-Aware Machine Learning-Based Approach for Lightpath QoT Estimation in Optical Networks / Jammal, H., Ayoub, O., Bianco, A., Rottondi, C.. - ELETTRONICO. - (2026), pp. 1-6. (2026 International Conference on Optical Network Design and Modelling (ONDM) Munich (Ger) 12-15 Maggio 2026) [10.23919/ondm68511.2026.11618793].

Availability:

This version is available at: 11583/3014552 since: 2026-09-23T20:39:10Z

Publisher:

IEEE

Published

DOI:10.23919/ondm68511.2026.11618793

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

Risk-Aware Machine Learning for Soft-Failure Classification in Optical Networks

Hussein Jammal, Andrea Bianco, Cristina Rottondi

Politecnico di Torino, Torino, Italy

hussein.jammal@polito.it, andrea.bianco@polito.it, cristina.rottondi@polito.it

Abstract—Machine Learning (ML)-based approaches for soft-failure classification in optical networks are commonly evaluated using aggregate metrics such as accuracy and F1-score, which fail to capture the operational impact of classification errors. In practice, different types of misclassification carry very different costs for the network operator, since failing to detect a degraded lightpath, raising an unnecessary alarm, and attributing a failure to the wrong root cause have distinct consequences on service continuity and maintenance effort. To address this asymmetry, we introduce a risk-aware multiclass soft-failure classification framework based on a Hybrid Ensemble (HE) architecture, which partitions the feature space into regions and allocates dedicated classifiers to each region. We also propose metrics to quantify the operational cost of classification errors and its distribution across failure types and feature-space regions. To verify that the gains come from forming meaningful feature-space regions rather than from the ensemble structure alone, we compare the HE against an Unguided Hybrid Ensemble (UHE) that follows the same ensemble architecture but whose regions are obtained through generic unsupervised clustering of the input features and unconditional acceptance of the classifiers assigned to those regions. Experiments on a multiclass soft-failure dataset show that the proposed approach improves overall accuracy by approximately 5% over a baseline model trained on the full dataset, while reducing expected decision risk by 95%, risk disparity across feature-space regions by 84%, and risk disparity across failure types by 95%.

Index Terms—Soft-Failure Classification, Optical Networks;

I. INTRODUCTION

Optical networks suffer from gradual physical-layer degradations, known as *soft failures*, such as Erbium-Doped Fiber Amplifier (EDFA) gain reduction, External Cavity Laser (ECL) drift, and excessive Nonlinear Interference (NLI), which erode the Signal-to-Noise Ratio (SNR) of the affected lightpaths and may eventually cause Bit-Error-Rate (BER) violations and service disruption [1]. Modern optical network management workflows increasingly adopt Machine Learning (ML) techniques that monitor parameters such as received power, Optical SNR (OSNR), and BER traces to catch these degradations before they cross the predefined alarm thresholds. Most ML-based approaches follow a two-stage paradigm in which a first stage detects the failure and a second stage classifies the type of the failure [2]–[5]. Of the two, the classification task most directly governs recovery time, since a misidentified cause leads the operator to the wrong corrective action, which would delay the restoration of the affected Quality of Transmission (QoT) [1].

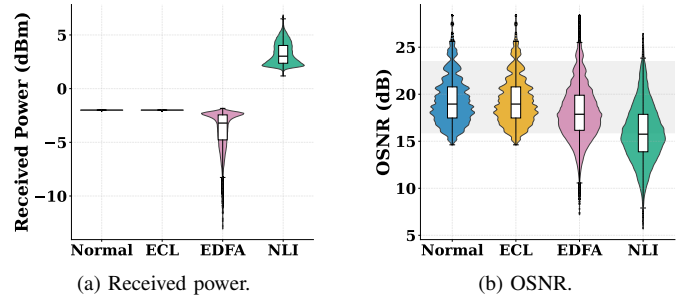


Fig. 1. Distribution of received power (dBm) and OSNR (dB) across the four classes in the soft-failure dataset [6], [7]: Normal, ECL, EDFA, and NLI.

Despite the progress made in this area, two important limitations remain. First, most existing approaches rely on aggregate metrics (e.g., accuracy, F1-score, ROC-AUC) that mask performance disparities across failure types. We define as *global model* a single classifier trained on the entire dataset to distinguish all failure types at once. Such a model performs well on clearly separable failure types and poorly on those overlapping with “Normal” operation, as illustrated in Fig. 1. For example, early-stage EDFA degradation overlaps substantially with Normal operation in terms of received power and OSNR, i.e., the parameters most diagnostic of amplifier health, whereas NLI produces clearly distinct signatures. The same overlap is observed in the other monitored parameters of the dataset [6], [7]. Consequently, the global model achieves a satisfactory aggregate performance while systematically missing a significant fraction of EDFA degradations, which are particularly critical since a single EDFA amplifies every Wavelength Division Multiplexing (WDM) channel traversing its span and its degradation propagates to all of those lightpaths simultaneously, increasing the risk of service disruption and Service Level Agreement (SLA) violations.

The second limitation concerns the asymmetric cost of classification errors, which standard metrics treat as equally harmful. We define the *operational cost* of a classification error as its downstream impact on the network, including service downtime, unnecessary maintenance, delayed repairs, and SLA violations. Three distinct error types arise in multiclass soft-failure classification. A *missed failure* labels a degraded lightpath as Normal, leaving the fault unaddressed and allowing the degradation to propagate across every lightpath sharing the affected device; a *false alarm* labels a healthy lightpath as degraded, triggering an unnecessary maintenance intervention

that wastes technicians' time; and a *misdiagnosis* attributes a real failure to the wrong root cause, sending the repair crew to the wrong component and delaying the actual repair while the true degradation continues to develop.

In this paper, we address both limitations by proposing a *Hybrid Ensemble (HE)* framework for multiclass soft-failure classification. The HE introduces *task-aligned partitioning* where it partitions the multidimensional feature space into regions that reflect classification difficulty, rather than generic feature similarity alone, so that lightpaths with similar discrimination characteristics are grouped together. The HE then assigns a specialized classifier, hereafter referred to as an *expert*, to each region. An expert selector then accepts an expert only when its performance on a separate validation set improves over that of the global model. To evaluate the operational behavior of the HE, we further introduce two risk-aware metrics: (i) a multiclass Expected Decision Risk (EDR), parameterized by a cost vector (α, β, γ) , respectively penalizing missed failures, false alarms, and misdiagnoses, and (ii) a multiclass Risk Disparity (RD), quantifying how unevenly the operational cost is distributed across regions and failure types. Experiments on the synthetically generated soft-failure dataset [6], [7] show that the HE achieves approximately 5% improvement in overall accuracy over the global model, together with a substantial reduction in the operational cost of classification errors and in its distribution across regions and failure types.

The remainder of this paper is organized as follows. Sec. II reviews the related work. Sec. III formalizes the problem and presents the HE methodology and risk-aware metrics. Sec. IV describes the experimental setup and baselines. Sec. V presents and discusses the results, and Sec. VI concludes the paper.

II. RELATED WORK

ML-based methods for soft-failure detection and identification in optical networks have been widely studied. Hybrid supervised/unsupervised architectures have been proposed for multi-domain detection with limited abnormal samples [8], while GAN-based methods have reduced the need for failure samples during training [9]. Other studies have used auto-encoder features with support vector machines [10], attention-based deep learning [5], and recurrent models for amplifier fault classification or soft-failure evolution prediction [11], [12]. SHAP-based explanations have also been used to identify influential telemetry features in failure detection and prediction pipelines [13], [14]. These works demonstrate the effectiveness of ML for optical-network failure management, but they mainly assess performance through aggregate indicators such as accuracy, F1-score, or ROC-AUC.

A. Limitations of Existing Work

Despite this progress, three limitations remain. First, aggregate metrics can hide large disparities across failure types, especially when early-stage degradations overlap with Normal operation. Second, existing classifiers typically treat all misclassifications as equally harmful, although missed failures,

false alarms, and misdiagnoses have different operational consequences. Third, prior work has not combined explainability-guided feature projection, region-specific expert specialization, and cost-aware evaluation within a single multiclass soft-failure classification framework.

Our work addresses these gaps by proposing a HE model that partitions the feature space using SHAP-weighted Principal Component Analysis (PCA) projection [15], [16] and uncertainty-derived indicators, assigns specialized experts to disjoint regions, and retains only experts whose local performance exceeds that of the global model. We also introduce risk-aware multiclass metrics that quantify both the operational cost of classification errors and its distribution across failure types and feature-space regions.

III. PROBLEM FORMULATION AND METHODOLOGY

A. Problem Formulation

We refer to the problem addressed in this work as *risk-aware multiclass soft-failure classification*. It formally extends the classical supervised multiclass classification setting by explicitly accounting for the asymmetric operational cost of different types of misclassifications.

The underlying ML problem is formulated as follows. Given a dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$, where $\mathbf{X}_i \in \mathbb{R}^{T \times f}$ is a window of T consecutive observations of f monitored parameters collected on lightpath i , and $y_i \in \mathcal{Y} = \{\text{Normal, ECL, EDFA, NLI}\}$ denotes its class label, the objective is to learn a classifier $\mathcal{C} : \mathbb{R}^{T \times f} \rightarrow \mathcal{Y}$ that maps each monitoring window to its corresponding class.

Let $\hat{y}_i = \mathcal{C}(\mathbf{X}_i)$ denote the predicted label. For a given prediction $\hat{y}_i$, evaluating its operational impact requires distinguishing not only whether it is correct, but also which type of error it represents when it is not. The three error types introduced in Sec. I can now be defined formally in terms of y_i and $\hat{y}_i$ as follows. A *missed failure (MF)* occurs when $\hat{y}_i = \text{Normal}$ and $y_i \neq \text{Normal}$; a *false alarm (FA)* occurs when $\hat{y}_i \neq \text{Normal}$ and $y_i = \text{Normal}$; and a *misdiagnosis (MD)* occurs when $\hat{y}_i \neq y_i$ with both $y_i \neq \text{Normal}$ and $\hat{y}_i \neq \text{Normal}$. Each error type is weighted by a non-negative cost, collected in a cost vector (α, β, γ) , where α , β , and γ penalize MF, FA, and MD, respectively, with $\alpha \geq \gamma \geq \beta > 0$, reflecting the operational severity ordering since a missed failure leaves an active degradation undetected, a misdiagnosis directs repair toward the wrong root cause, and a false alarm triggers unnecessary intervention on a healthy lightpath.

B. Methodology

The proposed *Hybrid Ensemble (HE)* decomposes the multiclass classification task into a set of region-specific classification tasks, each handled by a dedicated expert whose contribution to the ensemble is retained only when it improves over the global model. The overall workflow is illustrated in Fig. 2.

A partitioning function $\mathcal{P} : \mathbb{R}^{T \times f} \rightarrow \mathcal{G}$ assigns each monitoring window $\mathbf{X}_i$ to one of K disjoint regions $\mathcal{G} = \{g_1, g_2, \dots, g_K\}$ that partition the input feature space. The

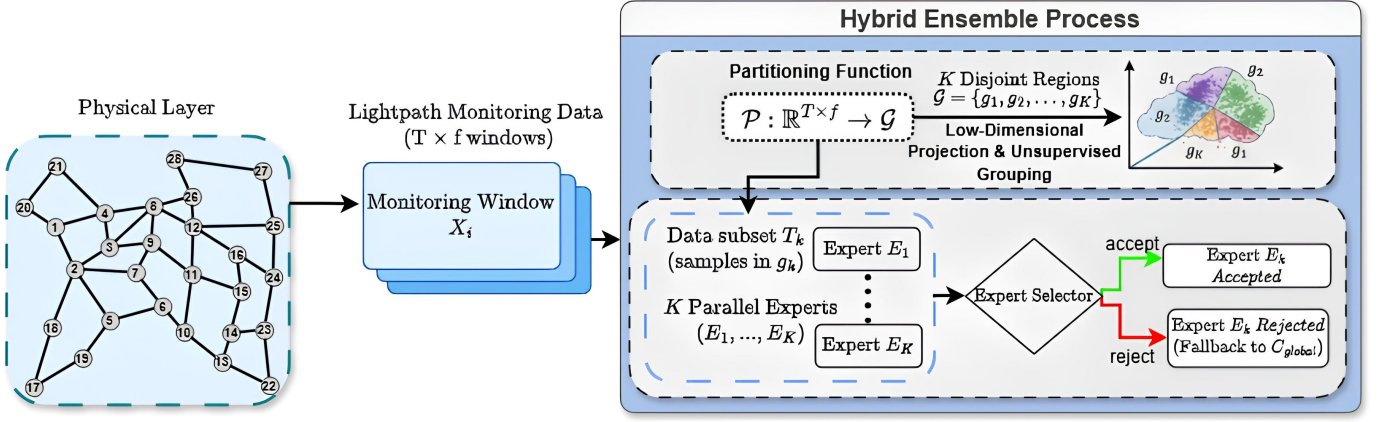


Fig. 2. Building the Hybrid Ensemble model.

partitioning is obtained by projecting the input features into a low-dimensional space designed to reflect classification difficulty across the full class set $\mathcal{Y}$, and then grouping the projected samples into K non-overlapping regions through an unsupervised procedure. The resulting regions group together lightpaths that share similar discrimination characteristics with respect to the class set $\mathcal{Y}$.

Let $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ denote the training and validation sets, respectively. For each region g_k , we define the region-specific training subset as $\mathcal{T}_k = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_{\text{train}} : \mathcal{P}(\mathbf{X}_i) = g_k\}$ and the corresponding validation subset as $\mathcal{V}_k = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_{\text{val}} : \mathcal{P}(\mathbf{X}_i) = g_k\}$. A region-specific expert $\mathcal{E}_k$ is trained on $\mathcal{T}_k$ to distinguish all classes in $\mathcal{Y}$ within the region. All experts share the same architectural family as the global model and are independently tuned on their own region, so that each expert can adapt to the local difficulty of its region while the comparison with the global model remains fair.

To prevent a poorly trained expert from degrading the overall performance of the ensemble, each expert is filtered through an expert selector. The expert $\mathcal{E}_k$ is accepted if and only if its macro F1 score on $\mathcal{V}_k$ exceeds that of the global model $\mathcal{C}_{\text{global}}$ on the same samples; otherwise $\mathcal{E}_k$ is rejected and the predictions for region g_k fall back to $\mathcal{C}_{\text{global}}$. We denote by $\mathcal{A} \subseteq \{1, \dots, K\}$ the set of region indices whose experts pass this expert selector. By construction, this rule ensures that the HE is at least as good as the global model in every region on the validation set.

At inference time, given a monitoring window $\mathbf{X}$, the HE first determines its region $g_k = \mathcal{P}(\mathbf{X})$ and returns

$$\mathcal{C}_{\text{HE}}(\mathbf{X}) = \begin{cases} \mathcal{E}_k(\mathbf{X}), & k \in \mathcal{A}, \\ \mathcal{C}_{\text{global}}(\mathbf{X}), & k \notin \mathcal{A}. \end{cases} \quad (1)$$

We remark that the proposed HE is agnostic to the specific choice of partitioning projection, unsupervised grouping algorithm, and base classifier architecture, and supports any partitioning strategy that correlates with classification difficulty.

C. Evaluation Metrics

We introduce two risk-aware metrics that quantify the operational cost of the classifier's errors and how unevenly

this cost is distributed across the feature space and failure types.

Expected Decision Risk (EDR). Extending the binary formulation of our previous work [17] to the multiclass setting, we define the EDR as the average per-sample cost incurred by the classifier:

$$\text{EDR} = \frac{1}{N} \sum_{i=1}^N \left[\alpha \mathbb{I}_{\text{MF}}(i) + \beta \mathbb{I}_{\text{FA}}(i) + \gamma \mathbb{I}_{\text{MD}}(i) \right], \quad (2)$$

where $\mathbb{I}_{\text{MF}}(i)$, $\mathbb{I}_{\text{FA}}(i)$, and $\mathbb{I}_{\text{MD}}(i)$ are indicator functions for the three error events defined in Sec. III-A. A lower EDR corresponds to a classifier whose errors carry a smaller expected operational cost.

Risk Disparity (RD). Following the definition introduced in [17], RD measures the gap between the maximum and the minimum EDR computed over a given partition of the samples. Unlike the binary case, the multiclass setting exposes two distinct axes along which the operational cost can be unevenly distributed, and we therefore report RD along both. The first, $\text{RD}_{\text{region}}$, is obtained by computing the EDR separately on each of the K regions produced by the partitioning function $\mathcal{P}$ and taking the gap between the highest and the lowest value. The second, RD_{class} , is obtained analogously by computing the EDR separately on the samples of each failure class in $\mathcal{Y} \setminus \{\text{Normal}\}$. A low value of both $\text{RD}_{\text{region}}$ and RD_{class} is meaningful only when the overall EDR is itself low, since together they indicate a classifier that minimizes operational cost while distributing it evenly across the network and across failure types.

IV. EXPERIMENTAL SETUP

Dataset. We evaluate the proposed approach on the synthetic soft-failure dataset [6], [7], generated via physical-layer simulation of a software-defined optical network over the Euro28 backbone topology (28 nodes, 41 links). One lightpath per source–destination pair is routed over the shortest path, yielding 756 lightpaths in total. Each node is equipped with a coherent receiver providing optical performance monitoring capabilities and a tunable ECL-based Dense WDM transceiver. All lightpaths use Polarization Multiplexed Quadrature Phase

Shift Keying (PM-QPSK) at 32 GBaud. Fiber links use standard single-mode fiber (0.22 dB/km attenuation, $1.3 \text{ W}^{-1}\text{km}^{-1}$ nonlinear coefficient, 16.7 ps/nm/km dispersion), with inline EDFAs (5 dB noise figure, 17.6 dB nominal gain) placed every 80 km, and preamplifiers and booster amplifiers at each node.

QoT parameters (ECL operating current, received optical power, OSNR, and BER) are recorded at 10-minute intervals under four conditions: normal operation, ECL malfunction (threshold current increase), EDFA gain degradation (pump power decrease), and NLI impairment (transmit power increase). The training set contains 900 time samples per lightpath per label (150 hours), and the test set contains 300 time samples per lightpath per label (50 hours), yielding roughly 2.7M training samples and 900k test samples balanced across the four classes. Each sample is a window of 10 consecutive observations of the four time-varying parameters, augmented with static lightpath length as a context feature.

Feature Preparation and Partitioning. We instantiate the partitioning function $\mathcal{P}$ using the full set of available input features.¹ Specifically, the partitioning space is a five-dimensional representation combining two complementary sets of indicators: (i) a SHAP-weighted PCA projection onto three components that emphasizes the directions most informative for discriminating the hard classes, and (ii) two proxy indicators of classification difficulty extracted from the global model, namely the predicted probability of the Normal class and the entropy of the softmax distribution. The SHAP-weighted PCA is constructed by rescaling each input feature by its normalized SHAP importance score prior to projection, so that the principal axes align with the directions most relevant for class discrimination rather than those of maximum variance. The optimal number of regions K is selected by evaluating the silhouette score over candidate values $K \in \{2, 3, 4, 5, 6\}$ on the training partition, yielding $K = 2$. Samples are then grouped into the two disjoint regions using the K -means algorithm applied in the partitioning space.

Baselines. We compare the proposed HE against two baselines. The first is the *Global LSTM*, a single LSTM trained on the entire dataset without any partitioning, which isolates the gain attributable to region-specific specialization. The second is a *Unguided Hybrid Ensemble (UHE)*, an ensemble that follows the same procedure as the HE but replaces the task-aligned partitioning with an unweighted PCA projection onto the raw standardized features, omitting the SHAP weighting and the proxy difficulty indicators. The UHE uses $K = 4$ regions selected by silhouette score and accepts all experts unconditionally, without an expert selector. In this way, the contribution of task-aligned partitioning can be quantified: any performance difference between the UHE and the proposed HE is attributable solely to the partitioning strategy.

¹Before defining the final partitioning space, we also evaluated a SHAP-driven recursive feature elimination procedure [15] to remove features with the weakest contribution to class discrimination. The procedure retained all five available features, so no feature was discarded.

TABLE I
OPTIMIZED HYPERPARAMETERS FOR ALL MODELS.

Model	Region	Hidden	Layers	Dropout	LR
Global	—	256	2	0.40	4.6×10^{-3}
HE	g_1	64	1	0.30	9.0×10^{-3}
	g_2	32	2	0.30	1.0×10^{-3}
UHE	g_1	64	1	0.00	8.8×10^{-3}
	g_2	32	2	0.30	1.0×10^{-3}
	g_3	256	3	0.20	1.7×10^{-4}
	g_4	256	1	0.10	1.7×10^{-3}

Models. All models are implemented as LSTM classifiers, which naturally match the sequential structure of the monitoring windows. The hyperparameters of the global model and of each expert are tuned independently on their respective training subsets through Bayesian optimization [18], with window size fixed at 10 across all models. The resulting hyperparameters are reported in Tab. I.

The HE experts are notably compact relative to both the global model and the UHE experts, with hidden sizes of 64 and 32 units, respectively, reflecting the focused nature of the subproblems each region presents. The relationship between region composition and model size is examined in detail later in Sec. V.

Evaluation. Models are assessed using overall accuracy, macro F1 score, and ROC-AUC as standard classification metrics, together with the risk-aware metrics EDR, $\text{RD}_{\text{region}}$, and RD_{class} introduced in Sec. III. Since EDR is linear in (α, β, γ) , we report the normalized form EDR/α , which removes the dependence on the absolute scale of α and reduces the cost vector to the two ratios β/α and γ/α ; the two RD metrics are normalized in the same way. To explore different operational priorities, results are evaluated under six cost vector configurations spanning from missed-failure-dominated regimes (small $\beta/\alpha, \gamma/\alpha$) to equal-weight regimes ($\beta/\alpha = \gamma/\alpha = 1$).

V. RESULTS

We compare the numerical results across the three methods, namely, the Global LSTM baseline, the UHE baseline, and the proposed HE. We first examine standard classification metrics, then per-class behavior and error composition together with the risk-aware metrics, and finally the robustness of the gains across cost vector configurations.

Standard Classification Metrics. Tab. II reports overall accuracy, macro F1 score, and ROC-AUC for all three methods. The UHE degrades performance relative to the Global model on every metric, with accuracy falling from 0.9450 to 0.9046 (−4.27% relative to the Global model) and macro F1 decreasing by −3.97%, confirming that the ensemble structure itself

TABLE II
STANDARD CLASSIFICATION METRICS. BEST RESULTS IN BOLD.

Method	Accuracy	Macro F1-Score	ROC-AUC
Global	0.9450	0.9444	0.9928
UHE	0.9046	0.9069	0.9917
HE (proposed)	0.9973	0.9973	0.9992
Δ HE vs Global (%)	+5.53	+5.60	+0.65

Global LSTM					UHE					HE				
T	Pred.				T	Pred.				T	Pred.			
	Normal	ECL	EDFA	NLI		Normal	ECL	EDFA	NLI		Normal	ECL	EDFA	NLI
Normal	100.0	0.0	0.0	0.0	Normal	100.0	0.0	0.0	0.0	Normal	100.0	0.0	0.0	0.0
ECL	0.5	99.5	0.0	0.0	ECL	23.8	76.2	0.0	0.0	ECL	0.1	99.9	0.0	0.0
EDFA	21.5	0.0	78.5	0.0	EDFA	14.3	0.0	85.7	0.0	EDFA	1.0	0.0	99.0	0.0
NLI	0.0	0.0	0.0	100.0	NLI	0.0	0.05	0.0	99.95	NLI	0.0	0.0	0.0	100.0

Fig. 3. Row-normalized confusion matrices (%). T = true label, Pred = predicted label. Off-diagonal entries in the Normal column are missed failures; the orange cell is a misdiagnosis.

TABLE III
CLASS COMPOSITION (%) OF EACH REGION FOR THE UHE AND THE PROPOSED HE MODELS.

Method	Region	Normal	ECL	EDFA	NLI
HE	g_1 (failure-dense)	0.0	33.4	33.1	33.5
	g_2 (Normal-adj.)	98.9	0.2	0.9	0.0
UHE	g_1	27.0	26.9	25.6	20.5
	g_2	32.8	32.9	34.3	0.0
	g_3	0.0	0.0	0.0	100.0
	g_4	29.3	29.3	25.1	16.4

confers no benefit without task-aligned routing. In contrast, the proposed HE raises accuracy to 0.9973 (+5.53% relative to the Global model) and macro F1 to 0.9973 (+5.60%). The ROC-AUC improvement over the Global model is modest in relative terms (+0.65%) because the global model is already near-optimal on this metric.

Takeaway 1: Unguided partitioning degrades performance relative to a single global classifier, demonstrating that the value of the HE architecture lies entirely in how the partition is constructed, not in the presence of multiple experts.

Partition Structure and Per-Class Behavior. Tab. III reveals the structural mechanism behind the aggregate results discussed above. The HE’s partition procedure produces two regions: g_1 , which concentrates the failure classes exclusively (ECL:33.4%, EDFA: 33.1%, NLI: 33.5%), and g_2 , which is Normal-dominated (98.9%), capturing the Normal/EDFA onset boundary where the global model fails. By contrast, the UHE regions mix the classes and do not isolate the failure subspace, so their experts inherit the same overlap faced by the Global model. This explains why the HE experts can be smaller (64 and 32 hidden units) than the UHE experts in g_3 and g_4 (256 hidden units; Tab. I).

The per-class accuracy of these two partitioning criteria is shown in Fig. 4. The HE’s regions enable each expert

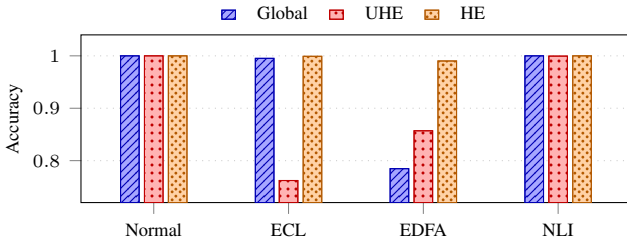


Fig. 4. Per-class accuracy of all three methods on the test set.

to specialize on a specific subproblem: g_1 ’s expert learns to discriminate between failure types without the distraction of Normal samples, while g_2 ’s expert focuses exclusively on the detection boundary where early EDFA degradation is hardest to distinguish from normal operation. The result is a simultaneous improvement on both classes: EDFA accuracy rises from 0.7847 to 0.9902 (+26.06%), and ECL from 0.9952 to 0.9991 (+0.39%), while Normal and NLI remain at almost perfect accuracy. The UHE’s mixed regions produce the opposite outcome: ECL accuracy collapses from 0.9952 to 0.7621 (−23.4%) and the marginal EDFA gain (0.7847 to 0.8570) is achieved only by redistributing errors from EDFA onto ECL, which is a trade that increases total misclassification rather than reducing it.

Takeaway 2: Task-aligned partitioning resolves the fundamental limitation of a global classifier, namely its inability to simultaneously handle classes with similar characteristics and classes that are easily separable, by decomposing a single difficult classification problem into a set of easier subproblems. Specialization, when properly guided, does not trade performance of one class for another but raises the performance floor across all classes simultaneously.

Error Decomposition and Operational Cost. The error type breakdown is reported via confusion matrices in Fig. 3. In both the Global model and the HE, every misclassification takes the form of a missed failure, i.e. a degraded lightpath predicted as Normal. No false alarms and no misdiagnoses arise in either method. The UHE introduces a small NLI to ECL misdiagnosis rate (0.05%) and substantially worsens the missed failure rate for ECL (0.5% to 23.8%) while only modestly improving EDFA (21.5% to 14.3%).

Tab. IV then translates this error structure into operational cost at $\beta/\alpha = \gamma/\alpha = 0.1$. The UHE raises the missed failure rate from 5.50% to 9.52% (+73.0%), pushing EDR/α from 0.055 to 0.0952, i.e., nearly doubling the expected operational cost per monitored lightpath. The HE moves decisively in the opposite direction. The missed failure rate falls to 0.27% (−95.17%), EDR/α drops to 0.0027, $\text{RD}_{\text{region}}/\alpha$ to 0.0105

TABLE IV
ERROR RATE DECOMPOSITION AND RISK-AWARE METRICS AT $\beta/\alpha = \gamma/\alpha = 0.1$. BEST RESULTS IN BOLD.

Method	MF (%)	FA (%)	MD (%)	EDR/α	$\text{RD}_{\text{region}}/\alpha$	$\text{RD}_{\text{class}}/\alpha$
Global	5.50	0.00	0.00	0.0550	0.0665	0.2153
UHE	9.52	0.00	≈0.01	0.0952	0.4920	0.2379
HE (proposed)	0.27	0.00	0.00	0.0027	0.0105	0.0098
Δ HE vs Global (%)	−95.17	—	—	−95.17	−84.17	−95.46

TABLE V
CHANGE (%) IN EDR AND RD_{class} RELATIVE TO THE GLOBAL MODEL
ACROSS SIX COST VECTOR CONFIGURATIONS. NEGATIVE VALUES
INDICATE REDUCTION; POSITIVE VALUES INDICATE INCREASE.

$(\beta/\alpha, \gamma/\alpha)$	HE (proposed)		UHE	
	ΔEDR	ΔRD_{class}	ΔEDR	ΔRD_{class}
(0.05, 0.05)	-95.17	-95.46	+73.02	+10.48
(0.10, 0.10)	-95.17	-95.46	+73.03	+10.48
(0.10, 0.50)	-95.17	-95.46	+73.13	+10.48
(0.50, 0.10)	-95.17	-95.46	+73.03	+10.48
(0.50, 0.50)	-95.17	-95.46	+73.13	+10.48
(1.00, 1.00)	-95.17	-95.46	+73.25	+10.48

(-84.17%), and RD_{class}/α to 0.0098 (-95.46%). Critically, these reductions are concentrated in missed failures, i.e., the most costly error type under $\alpha \geq \gamma \geq \beta$.

Takeaway 3: Aggregate accuracy is an insufficient basis for evaluating soft-failure classifiers because it assigns equal weight to all misclassification types regardless of their operational consequence. The risk-aware metrics expose what accuracy conceals: a classifier can appear competitive on standard benchmarks while simultaneously inflating the expected operational cost that network operators must absorb in practice.

Cost Vector Sensitivity. Tab. V evaluates whether the observed rankings hold across a range of operator cost assumptions, from missed-failure-dominated regimes ($\beta/\alpha = \gamma/\alpha = 0.05$) to equal-weight settings ($\beta/\alpha = \gamma/\alpha = 1.0$). For the HE, both EDR and RD_{class} reductions are invariant across all six configurations at -95.17% and -95.46%, with zero spread. This invariance has a precise cause: since neither the Global model nor the HE produces false alarms or misdiagnoses, the EDR of both methods is determined entirely by their missed failure rates. The relative improvement between them therefore does not depend on how false alarms and misdiagnoses are weighted, because neither method produces them. The same reasoning applies to both RD metrics. For the UHE, the EDR deterioration grows marginally with γ/α (+73.0% to +73.2%) because its small misdiagnosis rate becomes slightly more penalized as the misdiagnosis cost increases, while RD_{class} remains constant at +10.5%. Across all configurations, the spread is below 0.25 percentage points for all methods, confirming that the performance ordering is consistent regardless of the cost regime.

Takeaway 4: The proposed approach provides consistent risk reduction across all cost configurations.

VI. CONCLUSION

This paper proposed a risk-aware Hybrid Ensemble (HE) framework for multiclass soft-failure classification in optical networks together with two risk-aware evaluation metrics, namely Expected Decision Risk (EDR) and Risk Disparity (RD), that assess classifiers according to the operational impact of classification errors rather than aggregate performance alone. The proposed framework partitions the feature space according to classification difficulty, enabling specialized classifiers to address simpler and more homogeneous subproblems

while reverting to a global model whenever specialization is not beneficial.

Experimental results on a multiclass soft-failure dataset show that the proposed approach improves classification accuracy by approximately 5% compared with a single global classifier, while reducing EDR by about 95% and substantially decreasing RD across both feature-space regions and failure types. The UHE comparison further indicates that these improvements are driven primarily by task-aligned partitioning rather than by the ensemble architecture alone.

Future work will extend the framework to broader soft-failure taxonomies and investigate adaptive cost models that dynamically reflect operator priorities.

REFERENCES

- [1] F. Musumeci *et al.*, “Failure management in optical networks with ML: A tutorial on applications, challenges, and pitfalls [Invited],” *J. Opt. Commun. Netw.*, vol. 17, no. 8, pp. C144–C155, Aug. 2025.
- [2] B. Shariati *et al.*, “Learning from the optical spectrum: Failure detection and identification,” *J. Lightw. Technol.*, vol. 37, no. 2, pp. 433–440, Jan. 2019.
- [3] S. Shahkarami *et al.*, “Machine-learning-based soft-failure detection and identification in optical networks,” in *Proc. Optical Fiber Communication Conference (OFC)*, 2018, pp. 1–3, paper M3A.5.
- [4] H. Lun *et al.*, “Soft failure identification for long-haul optical communication systems based on one-dimensional convolutional neural network,” *J. Lightw. Technol.*, vol. 38, no. 11, pp. 2992–3002, Jun. 2020.
- [5] C. Zhang *et al.*, “Potential failure cause identification for optical networks using deep learning with an attention mechanism,” *J. Opt. Commun. Netw.*, vol. 14, no. 2, pp. A122–A133, Feb. 2022.
- [6] S. Ghosh *et al.*, “Soft failure detection and identification in optical networks using cascaded deep learning model,” *Comput. Netw.*, vol. 262, p. 111217, 2025.
- [7] S. Ghosh *et al.*, “Optical network soft failure dataset,” Feb. 2025. [Online]. Available: <https://data.mendeley.com/datasets/y3pspy7j83/1>
- [8] X. Chen *et al.*, “Automating optical network fault management with machine learning,” *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 88–94, Oct. 2022.
- [9] H. Lun *et al.*, “A GAN based soft failure detection and identification framework for long-haul coherent optical communication systems,” *J. Lightw. Technol.*, vol. 41, no. 5, pp. 1398–1408, Mar. 2023.
- [10] L. E. Kruse *et al.*, “Experimental investigation of machine-learning-based soft-failure management using the optical spectrum,” *J. Opt. Commun. Netw.*, vol. 16, no. 2, pp. 94–103, Feb. 2024.
- [11] M. C. Mohamed *et al.*, “Machine learning agents leveraging digital twins for failure prediction in optical networks,” in *Proc. Int. Conf. Machine Learning for Communication Networks (ICMLCN)*, 2025.
- [12] S. Behera *et al.*, “Modeling soft-failure evolution for triggering timely repair with low QoT margins,” in *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2022, pp. 2140–2145.
- [13] C. Zhang *et al.*, “Cause-aware failure detection using an interpretable XGBoost for optical networks,” *Opt. Express*, vol. 29, no. 20, pp. 31974–31992, Sep. 2021.
- [14] O. M. Ali *et al.*, “Proactive soft-failure prediction in optical transport networks via physics-inspired features and infrastructure-as-code orchestration,” *Sci. Rep.*, vol. 16, p. 52186, 2026.
- [15] S. M. Lundberg *et al.*, “A unified approach to interpreting model predictions,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [16] I. T. Jolliffe *et al.*, “Principal component analysis: a review and recent developments,” *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
- [17] H. Jammal *et al.*, “Risk-aware machine learning-based approach for lighthouse QoT estimation in optical networks,” in *Proc. International Conference on Optical Network Design and Modeling (ONDM)*, 2025.
- [18] T. Akiba *et al.*, “Optuna: A next-generation hyperparameter optimization framework,” in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2019, pp. 2623–2631.