Variational Autoencoder Domain Adaptation for Cross-System Generalization in ML-Based SOP Monitoring

Citation for the original published paper (version of record):
Sadighi, L., Karlsson, S., Natalino Da Silva, C. et al (2026). Variational Autoencoder Domain Adaptation for Cross-System Generalization in ML-Based SOP Monitoring. Journal of Optical Communications and Networking, 18(8): C149-C159.
http://dx.doi.org/10.1364/JOCN.598996

Variational Autoencoder Domain Adaptation for Cross-System Generalization in ML-Based SOP Monitoring

LEYLA SADIGHI[1,2,*], STEFAN KARLSSON[3], CARLOS NATALINO[2], MOJTABA ESHGHIE[4], FEHMIDA USMANI[2], EOIN KENNY[5], LENA WOSINSKA[2], PAOLO MONTI[2], MARIJA FURDEK[2], AND MARCO RUFFINI[1]

[1] *School of Computer Science and Statistics, IRIS research group, ADAPT Centre, Trinity College Dublin (TCD), Dublin, Ireland*
[2] *Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden*
[3] *Micropol Fiberoptics AB, Stockholm, Sweden*
[4] *Department of Computing Science, Umeå University, Umeå, Sweden*
[5] *Asiera, Dublin, Ireland*
[*] *sadighil@tcd.ie*

Compiled July 1, 2026

Machine learning (ML) models trained to detect physical-layer threats on one optical fiber system often fail catastrophically when applied to a different system, due to variations in operating wavelength, fiber properties, and network architecture. To overcome this, we propose a Domain Adaptation (DA) framework based on a Variational Autoencoder (VAE) that learns a shared representation capturing event signatures common to both systems while suppressing system-specific differences. The shared encoder is first trained on the combined data from two distinct optical systems: a 21 km O-band dark-fiber testbed (System 1) and a 63.4 km C-band live metro ring (System 2). The encoder is then frozen, and a classifier is trained using labels from an individual system. The proposed approach achieves 95.3% and 73.5% cross-system accuracy when moving from System 1 to System 2 and vice versa, respectively. This corresponds to gains of 83.4% and 51% over a fully supervised Deep Neural Network (DNN) baseline trained on a single system, while preserving intra-system performance.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Optical fiber networks are the backbone of modern telecommunications infrastructure, carrying the vast majority of global data traffic. To keep pace with ever-increasing traffic demands, these networks are evolving rapidly: transmission is being extended into new spectral bands beyond the conventional C-band, and network operation is transitioning towards open, disaggregated and autonomous networks that combine various types of equipment from different vendors. Modern optical networks are, therefore, highly heterogeneous, and feature diverse deployments characterized by varying link lengths, topologies, and environmental conditions.

Beyond their primary role in communications, optical fibers are increasingly used for environment sensing and physical-layer threat monitoring applications [1–3]. This enables a new role of optical fibers as ubiquitously available sensors of external disturbances, which helps identify potential threats, prevent service disruptions, and protect optical network integrity.

Optical fiber sensors enable high-precision environmental monitoring due to their sensitivity to various physical phenomena, including Rayleigh, Raman, and Brillouin scattering, as well as interferometric effects and polarization variations induced by external perturbations [4]. Among these sensing modalities, the State of Polarization (SOP)is a particularly sensitive metric: external events such as vibrations or eavesdropping attempts alter the birefringence of the fiber, inducing characteristic, measurable variations in the SOP of the propagating light. The advantages of SOP-based monitoring are bolstered by the widespread deployment of coherent optical transceivers in modern optical networks. These devices inherently capture polarization-resolved optical field information through their dual-polarization receivers, enabling software-based SOP estimation without requiring additional dedicated hardware [3, 5]. This capability makes SOP-based network monitoring cost-effective and readily deployable in existing infrastructure.

Effectively interpreting SOP signatures of different events remains a significant challenge. Traditional monitoring approaches that rely on predetermined rules and fixed thresholds often fail to capture the complexity and variability of real-world fiber events, particularly those that introduce subtle, transient, or overlapping SOP variations [1]. Recent work has demonstrated that Machine Learning (ML)-based classifiers applied to SOP signatures can detect and categorize physical-layer events with high accuracy when trained and evaluated within the same optical system [6]. However, the growing heterogeneity of optical networks, spanning different spectral bands, fiber types, link lengths, and environmental conditions, introduces significant variability in the physical-layer characteristics observed over different network segments, posing a fundamental challenge for data-driven monitoring approaches. This raises an important research question: *can ML models trained on one optical system maintain their accuracy when deployed on a system with fundamentally different physical characteristics (e.g. spectral bands, fiber links, and network topologies)?*

In our previous study [7], we demonstrated that an eXtreme Gradient Boosting (XGBoost) classifier trained on one system, which transmits a signal in the O-band over a 21 km dark fiber (referred to as System 1), achieved only 9.08% accuracy when tested on another system, transmitting in C-band over a 63.4 km link in the Asiera metro network (referred to as System 2), which is worse than the theoretical random guessing. As SOP monitoring is characterized by systematic differences in spectral bands, fiber links, network topologies, and environmental conditions between systems, the distributions of the learned features diverge substantially, and the resulting domain shift causes unacceptable performance degradation [8].

Domain Adaptation (DA) techniques provide a framework for overcoming domain shift by learning representations that are invariant over source and target domains, that is, features that capture the essential characteristics of physical-layer events regardless of which optical system produced the data [8]. Several DA strategies exist, each with distinct trade-offs. Direct Transfer Learning (TL) fine-tunes a source-pretrained model on the target domain, but requires labeled target data that may be scarce or costly in operational networks. Domain-adversarial methods encourage domain-invariant features by training against a domain discriminator, yet suffer from adversarial instability and risk discarding class-relevant information. Purely supervised approaches, regardless of architectural complexity, inherently overfit to source-domain statistics and offer no mechanism to handle unseen domain variation. Among generative approaches, i.e., methods that model the underlying data distribution rather than directly mapping inputs to class labels, Variational Autoencoders (VAEs) offer a compelling alternative [9]. Their probabilistic, regularized latent space naturally separates domain-specific variation from shared structure, and when trained jointly on multiple systems, encourages geometrically similar representations for semantically similar events regardless of their system of origin. This makes VAEs particularly well suited to bridging domain shift between heterogeneous optical systems, yet their benefits for DA in optical fiber monitoring have not been investigated.

In this paper, we extend our preliminary study [7] by proposing a VAE-based DA framework for cross-system generalization in ML-based SOP signature classification. To the best of our knowledge, this is the first application of VAE-based DA to optical fiber monitoring. In this work, we focus on classifying three physical-layer events from their SOP signatures: the nominal undisturbed fiber state (*relaxed*), unauthorized fiber tapping attempts (*eavesdropping*), and gradual mechanical perturbations caused by fiber bending (*soft bending*). We evaluate classification performance under four scenarios: two intra-system scenarios, S_1: System 1→System 1 and S_2: System 2→System 2, where the model is trained and tested on the same optical system, and two cross-system scenarios, S_3: System 1→System 2 and S_4: System 2→System 1, where the model is trained on one system and tested on a different one operating in a different spectral band, over a different fiber link, and within a different network topology. The cross-system scenarios are the primary focus of this work, as they directly probe whether ML models can generalize across heterogeneous optical infrastructure. To isolate the source of cross-system generalization gains, we establish a supervised Deep Neural Network (DNN) baseline and a single-system VAE (VAE_{sgl}) as reference points. Both achieve strong intra-system accuracy yet collapse to below-chance under cross-system transfer, demonstrating that neither architectural sophistication nor probabilistic latent structure alone are sufficient to bridge the domain shift. This motivates our proposed combined-system VAE (VAE_{cmb}), which trains a shared encoder jointly on both systems through a two-phase design, allowing the model to first learn the common patterns present in data from both systems before being taught to classify events. Our model achieves cross-system accuracy of 95.3% for System 1→System 2 and 73.5% for System 2→System 1, corresponding to gains of 83.4% and 51% over the DNN baseline.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the experimental testbed, problem formulation, and domain shift characterization. Section 4 describes the proposed VAE-based DA framework and the DNN benchmark. Section 5 details the data preprocessing, Hyperparameter (HP) search spaces, and Hyperparameter Optimization (HPO) setup. Section 6 presents and discusses the results, while Section 7 concludes the paper.

2. RELATED WORK

The application of ML to SOP-based fiber monitoring has been proliferating over the recent years [3, 10]. Supervised SOP signature classification has progressed from detection of harmful events on C-band laboratory data [6, 11] to robustness evaluation on real-world O-band deployments [12]. Representative results from Deep Learning (DL)-based approaches include accuracies of up to 98.57% on a short-link (0.3 km round-trip) and 92.26% on a long-link (21 km round-trip) metropolitan deployment [13]. Other relevant Supervised Learning (SL) approaches include physical-layer threat monitoring via simple SOP analyzers [14], vision-transformer-based anomaly classification from SOP spectrograms [15], abnormal activity detection on Single-Mode (SM) and Multi-Mode (MM) fibers [16], and Convolutional Neural Network (CNN)-based fiber-bending eavesdropping detection [17]. Beyond SL methods, Semi-supervised Learning (SSL) and Unsupervised Learning (USL) anomaly detection using One-Class Support Vector Machine (OCSVM) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has been demonstrated [18]. Cluster-based USL eavesdropping detection has been proposed for Wavelength Division Multiplexing (WDM) systems [19], and Digital Signal Processing (DSP)-blind SOP-based anomaly detection has been introduced for metropolitan fibers [20].

Despite the high intra-system accuracy reported across these works, cross-system generalization remains a largely unresolved challenge. The first systematic investigation of this problem for SOP-based monitoring revealed that classifiers trained on one optical link experience catastrophic accuracy degradation when applied to a different link operating in a different spectral band [7]. This domain shift problem is well recognized in other optical networking problems solved through the application of ML. For example, Quality of Transmission (QoT) estimation models have been shown to fail to generalize across network topologies [21], and analogous failures have been reported for cross-lightpath failure-cause identification [22]. Furthermore, transfer learning for QoT prediction depends on source–target domain similarity [23]. Several DA and TL strategies have been proposed to mitigate domain shift in optical networks, including model transfer for QoT prediction [24], domain adversarial adaptation with few labeled samples [25], and TL with image-encoded SOP data for event classification under data scarcity [26]. However, these approaches have primarily targeted QoT estimation or general fiber fault management, while none has addressed the cross-system generalization problem for SOP-based classification.

VAEs [9] learn compact, structured representations of high-dimensional data by training a model to reconstruct its inputs through a constrained bottleneck. This bottleneck forces the model to capture only the most essential characteristics of the data, while the probabilistic regularization encourages similar inputs to be mapped to nearby points in the learned representation space. In the context of optical fiber monitoring, this means that SOP signatures caused by the same physical event, such as eavesdropping, are represented similarly regardless of which optical system produced them, while system-specific differences, such as those arising from different spectral bands or fiber lengths, are naturally suppressed. Despite their demonstrated effectiveness in computer vision [27, 28] and signal processing [29], VAE-based DA methods have not been applied to optical network monitoring. The core unresolved research challenge in SOP-based monitoring tackled in this work, i.e., learning representations that capture event-specific signatures while being invariant to the optical system that produced them, aligns directly with the strengths of VAEs: their regularized latent space naturally encourages inputs that share the same underlying cause to be mapped to similar representations, even when the raw measurements differ substantially due to system-specific factors. This paper leverages this property by introducing a VAE-based DA framework for cross-system SOP signature classification, demonstrating that it can effectively bridge the severe domain shift between fundamentally different optical systems.

3. EXPERIMENTAL METHODOLOGY AND PROBLEM FORMULATION

This section describes the two optical systems used for data collection, formalizes the cross-system generalization problem, and provides empirical evidence of the domain shift that motivates the proposed framework.

A. Optical Testbeds and SOP Signature Collection

The experimental setup used in this study, illustrated in Fig. 1, comprises two independent fiber-optic measurement systems operating in distinct spectral bands. The first, System 1, operates in the O-band over a 21 km round-trip dark fiber link accessed via the OpenIreland testbed [30]. It employs a Continuous Wave Distributed Feedback (CW-DFB) laser source and standard SM G.652 fiber. As a field-deployed fiber, this link introduces realistic impairments representative of production infrastructure. The second, System 2, operates in the C-band over a live production metro ring network in Dublin city, operated by Ireland's National Education and Research Network, Asiera (formerly HEAnet) [31]. This network spans 63.4 km of fiber across six Reconfigurable Optical Add-Drop Multiplexer (ROADM) nodes and uses an External Cavity Laser (ECL) source within a 400 GHz spectral window (192.8–193.2 THz), allocated as a Spectrum-as-a-Service slice. Both systems share a common optical analyzer, connected via a switch that selects one transmission link at a time. The selected link is subjected to controlled physical perturbations designed to replicate real-world tampering and eavesdropping attempts, each inducing distinct variations in the SOP of the propagating signal. Both systems employ an unmodulated Continuous Wave (CW) source, which is representative of an operator performing SOP analysis on a dedicated wavelength without disturbing the existing data-bearing channels. The use of a non-data-bearing source does not affect the results. The source of a standard 10 GE transceiver in the C- or O-band is itself a polarized laser, and our polarization sampling rate of approximately 1 kHz is roughly seven orders of magnitude slower than the baud rate of 10 GBaud, so each polarization sample averages over millions of bits and is therefore insensitive to the modulation. We have separately verified that a 10 Gb/s on-off-keying channel yields equivalent SOP-classification performance [32].

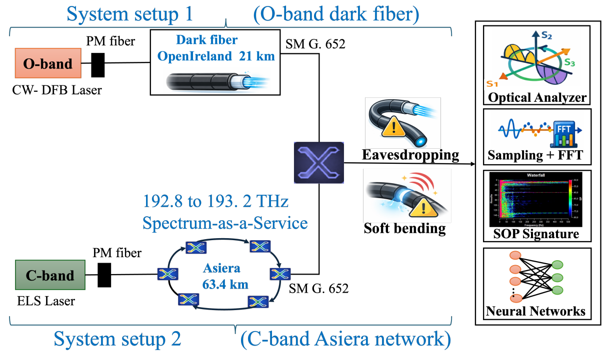


Fig. 1. Schematic of experimental testbed comprising System 1 and System 2.

We capture the SOP traces at a 0.5 ms sampling interval over a 20-minute window for each event, yielding 2.4 million samples per event and per spectral band. From these traces, we derive a spectral feature representation suitable for ML analysis. We begin by computing the Numerical Polarization State Variation (NPSV), a scalar metric that quantifies the spatial variation of the SOP between successive sampling instants [33]. The measurement is based on two orthogonally polarized detection channels, where the intensities of the horizontal (I_1) and vertical (I_2) polarization components are recorded independently. At each time instant t, we calculate $\Delta S(t)$ as $\Delta S(t) = I_1(t) - I_2(t)$, which reflects the phase-dependent interference between polarization modes and is sensitive to external perturbations on the fiber. The NPSV at time slot t is defined as the spatial difference between consecutive samples: $\text{NPSV}_t = \Delta S(t) - \Delta S(t-1)$. A value of $\text{NPSV}_t \approx 0$ indicates a stable polarization state, while positive values reflect a shift of the polarization state toward horizontal dominance, and negative values reflect a shift toward vertical dominance. The magnitude of NPSV_t thus captures the speed and direction of SOP transitions induced by physical perturbations on the fiber. The resulting NPSV time series is partitioned into non-overlapping segments of 500 samples. A 512-point Fast Fourier Transform (FFT) with a Hamming window [34] is applied to each segment, producing a power spectrum of 512 frequency bins. Aggregated over the full observation window, this yields a two-dimensional spectral representation of 4,800 time segments × 512 frequency bins per event, constituting a unique *SOP signature* that characterizes the polarization response of the fiber under that specific physical perturbation.

This procedure is repeated independently for both spectral systems, producing two datasets that are used as inputs to the ML pipeline. Each dataset comprises SOP signatures from three physical-layer event classes, measured independently on both systems: relaxed (*rlx*), representing the nominal undisturbed fiber state whose SOP signature reflects only background environmental noise; eavesdropping (*eav*), simulating unauthorized fiber tapping via optical couplers and producing characteristic SOP perturbations from the induced bending stress; and soft bending (*sbd*), representing gradual mechanical perturbations introduced by controlled fiber bending, which manifest as slower SOP variations. The *sbd* event is not intended to reproduce a fixed, calibrated attenuation, but to emulate the everyday handling a patch cable experiences in service, such as being repositioned or lightly flexed. In our setup it was produced by manually flexing a fiber segment to a curvature radius of approximately 2 cm and releasing it, repeating the action about every 10 seconds to reproduce the cable manipulation common during routine maintenance. The *eav* event reproduces an actual tapping attempt, prepared according to the procedure in our earlier work [35]. The exposed fiber was bent over a 10 mm-diameter rod to a curvature radius of 4 mm at a 25° angle, extracting about 3% of the guided light.In both classes, our framework recognizes the dynamic change in polarization as the fiber moves from the relaxed reference into the perturbed states, not any fixed polarization level. The tapping procedure is the same in both bands. As detection relies on the variation of the polarization state rather than the optical power, the higher fiber loss at shorter wavelengths is not a limiting factor as long as the received power stays above the detector sensitivity. Wavelength does, however, shape the polarization response. Over the same bent length with the same induced birefringence, a shorter wavelength accumulates a larger phase difference between its polarization components, so the O-band signal rotates further and faster than the C-band signal. The two wavelengths, therefore, produce genuinely distinct dynamic signatures, and reconciling this difference is exactly the cross-system generalization problem our framework targets, consistent with the domain-shift assumption of Eq. (1)–Eq. (2). The resulting System 1 and System 2 datasets comprise signatures $\{rlx_1, eav_1, sbd_1\}$ and $\{rlx_2, eav_2, sbd_2\}$, respectively, each containing 4,800 samples represented by 512-Dimensional (D) frequency feature vectors.

B. Cross-System Generalization Problem Formulation

Although both systems observe the same physical events and share the set of event classes $\mathcal{Y} = \{rlx, eav, sbd\}$, their SOP signatures are collected using distinct spectral bands, fiber lengths, and network topologies, inducing a *domain shift* between the two systems. Formally, let $\mathcal{X}_S = \{(\mathbf{x}_i^S, y_i^S)\}_{i=1}^{n_S}$ and $\mathcal{X}_T = \{(\mathbf{x}_j^T, y_j^T)\}_{j=1}^{n_T}$ denote the annotated SOP feature datasets from System 1 (source) and System 2 (target), with n_S and n_T denoting the number of labeled samples in each system, where $\mathbf{x} \in \mathbb{R}^{512}$ is a spectral feature vector and $y \in \mathcal{Y}$ its event class. The domain shift is characterized by:

$$P_S(\mathbf{x}) \neq P_T(\mathbf{x}), \tag{1}$$

$$P_S(y \mid \mathbf{x}) \approx P_T(y \mid \mathbf{x}), \tag{2}$$

where Eq. (1) states that the marginal input distributions differ due to system-specific polarization dynamics, and Eq. (2) reflects the physical expectation that identical events produce qualitatively consistent SOP perturbations in both systems. This is the key assumption that makes generalization theoretically feasible.

The core problem is to learn a classifier $f : \mathbb{R}^{512} \to \mathcal{Y}$ *that transfers from* $\mathcal{X}_S$ *to* $\mathcal{X}_T$ *despite the marginal mismatch in* Eq. (1). The model must bridge this mismatch by learning a domain-invariant representation that inherently preserves the class consistency of Eq. (2).

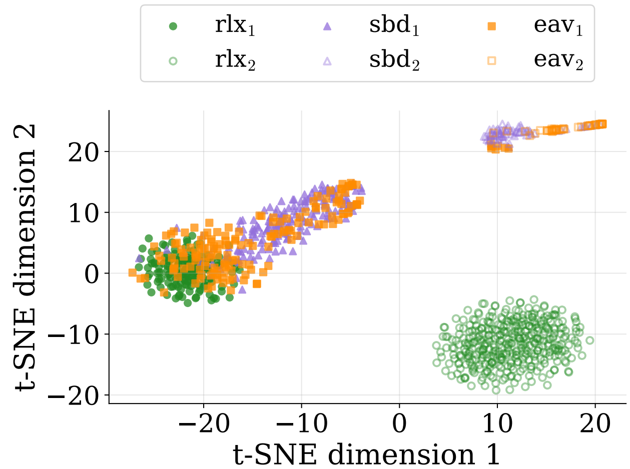


Fig. 2. t-SNE projection of 500 samples per system, colored by event class. Filled markers denote System 1 and open markers denote System 2.

C. Empirical Characterization of Domain Shift

To confirm the presence of domain shift, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) [36] to 500 training samples per system, projecting the 512-D SOP feature vectors onto a 2-D embedding. Fig. 2 colors each point by event class, using filled markers for System 1 and open markers for System 2. The figure reveals two key observations. First, the filled and open markers occupy completely disjoint regions, confirming the marginal distribution mismatch of Eq. (1). Second, instances of the same physical event from different systems, e.g., rlx_1 and rlx_2, are spatially separated to a degree comparable to inter-class distances, directly explaining the near-chance cross-system accuracy of 9.08% in scenario S_3 [7].

4. PROPOSED METHODOLOGY

The disjoint feature-space distributions revealed in Section 3.C indicate that direct cross-system transfer is highly challenging, motivating a representation learning approach that suppresses system-specific variation while preserving class discriminability. We first establish a supervised DNN baseline to quantify the intra-system performance ceiling and the cross-system generalization gap, and then present the proposed VAE-based framework designed to bridge this gap.

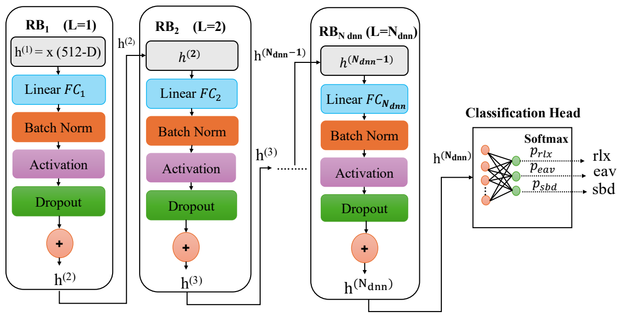


Fig. 3. Architecture of the Fully Connected (FC) DNN baseline with sequential Residual Blocks (RBs).

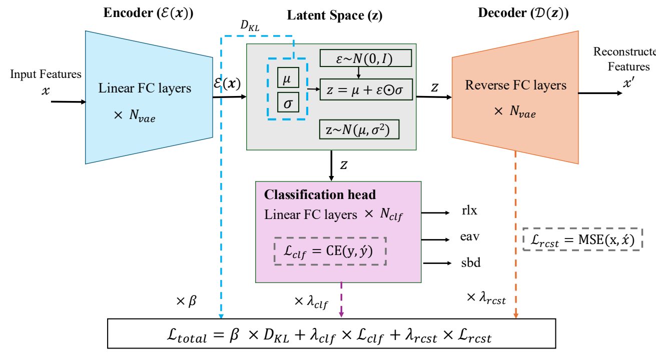


Fig. 4. General architecture of the proposed VAE framework comprises three components: an encoder, a decoder, and a classification head.

A. Fully Connected (FC) Residual DNN Baseline

As a supervised benchmark, we employ an FC residual DNN model that operates directly on the 512-D SOP feature vectors. The network architecture, as illustrated in Fig. 3, follows a sequential *encode-then-classify* structure: a backbone of N_{dnn} stacked Residual Blocks (RBs) transforms the input representation, followed by a linear classification head that maps the final hidden state to class logits. Each RB at layer ℓ applies the transformation:

$$\mathbf{h}^{(\ell+1)} = f\Big(\mathrm{BN}\Big(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)}\Big)\Big) + \mathbf{W}_{\text{skip}}^{(\ell)}\,\mathbf{h}^{(\ell)}, \tag{3}$$

where $\mathbf{h}^{(1)} = \mathbf{x}$ is the original 512-D SOP feature vector, and $\mathbf{h}^{(\ell)} \in \mathbb{R}^{d_\ell}$ for $\ell > 1$ is the hidden representation entering block ℓ; $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell}$ and $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}}$ are the learnable weight matrix and bias of the dense layer inside block ℓ; $\mathrm{BN}(\cdot)$ denotes Batch Normalization (BN) [37]; and $f(\cdot)$ is a non-linear activation function. To prevent over-fitting to the relatively small per-system training sets, each block further applies Dual-Polarization (DP) [38] immediately after the activation, stochastically zeroing a fraction p of the activations during training:

$$\tilde{\mathbf{a}}^{(\ell)} = \mathrm{Drop}\Big(f\Big(\mathrm{BN}\Big(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)}\Big)\Big),\, p\Big), \tag{4}$$

where p is treated as an HP and should be selected by HPO. The complete output of block ℓ, incorporating both the regularized transformation and the skip connection ($\mathbf{W}_{\text{skip}}$), is therefore:

$$\mathbf{h}^{(\ell+1)} = \tilde{\mathbf{a}}^{(\ell)} + \mathbf{W}_{\text{skip}}^{(\ell)}\,\mathbf{h}^{(\ell)}, \tag{5}$$

The skip connection weight $\mathbf{W}_{\text{skip}}^{(\ell)}$ takes one of two forms depending on the layer dimensions:

$$\mathbf{W}_{\text{skip}}^{(\ell)} = \begin{cases} \mathbf{I} & \text{if } d_{\ell+1} = d_\ell \quad \text{(identity)}, \\ \mathbf{W}_p^{(\ell)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell} & \text{if } d_{\ell+1} \neq d_\ell \quad \text{(learned projection)}. \end{cases} \tag{6}$$

The outcome $\mathbf{h}^{(\ell+1)}$ is then passed directly as the input to the next block, so that the representation is refined sequentially across all N_{dnn} RBs. After the final RB, the output $\mathbf{h}^{(N_{\text{dnn}}+1)} \in \mathbb{R}^{d_n}$ constitutes a compact, task-relevant embedding of the original SOP feature vector, where d_n denotes the width of the last RB. This embedding is passed to a linear classification head, which projects it onto three output class scores:

$$\mathbf{o} = \mathbf{W}_h\,\mathbf{h}^{(N_{\text{dnn}}+1)} + \mathbf{b}_h, \qquad \mathbf{W}_h \in \mathbb{R}^{3\times d_n}, \tag{7}$$

where $\mathbf{W}_h$ and $\mathbf{b}_h$ are learnable parameters and no activation function is applied. The class scores $\mathbf{o} \in \mathbb{R}^3$ are converted to class probabilities via the softmax function to obtain three output class scores corresponding to the event classes. All model parameters of the residual backbone and the linear head are optimized jointly by minimizing a classification loss $\mathcal{L}_{\text{dnn}}$ via gradient descent over the training set. The specific loss function (cross-entropy with optional label smoothing or focal loss) and the optimizer are treated as HPs and determined by HPO, as detailed in Section 5.B.

B. Proposed VAE Framework for DA

The DNN baseline maps each input directly to a deterministic embedding optimized solely for classification on the source domain. As a consequence, the learned representation is shaped by system-specific characteristics and carries no mechanism to suppress domain-induced variation. To address this limitation, we propose a VAE-based framework with a classification head that extends the deterministic encode-then-classify structure with a probabilistic encoder and a reconstruction decoder. Unlike a standard Autoencoder (AE), which maps each input to a single-point representation in latent space $\mathbf{z}$, the VAE encoder learns a conditional distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$ over latent variables. The Kullback–Leibler (KL) regularization term constrains this distribution toward a structured prior, encouraging a continuous and well-regularized latent space. Consequently, geometrically similar latent representations correspond to semantically similar inputs irrespective of their domain of origin, provided the encoder is exposed to both distributions during training. The jointly trained classification head further ensures that the learned representation remains directly actionable for event detection, making the framework well-suited for cross-system transfer.

The proposed framework, illustrated in Fig. 4, comprises three components: a probabilistic encoder, a decoder, and a classification head. Given an input SOP feature vector $\mathbf{x} \in \mathbb{R}^{512}$, the encoder $\mathcal{E}(\mathbf{x})$, comprising N_{vae} stacked FC layers with BN and DP regularization, maps $\mathbf{x}$ to the parameters of a posterior Gaussian distribution over a low-D latent space:

$$\boldsymbol{\mu},\, \log\sigma^2 = \mathcal{E}(\mathbf{x}), \qquad q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\Big(\boldsymbol{\mu},\, \mathrm{diag}(\boldsymbol{\sigma}^2)\Big), \tag{8}$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \in \mathbb{R}^{d_z}$ are the latent mean and variance vectors, and d_z is the latent dimensionality selected by HPO. A latent sample is drawn via the reparameterization trick,

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \odot \boldsymbol{\sigma}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

which preserves differentiability through the sampling operation and enables end-to-end training. The decoder $\mathcal{D}(\mathbf{z})$, a symmetric FC network of N_{vae} layers, reconstructs the input as $\mathbf{x}' \in \mathbb{R}^{512}$, while the classification head maps $\mathbf{z}$ through N_{clf} FC layers to class probabilities over $\mathcal{Y} = \{rlx, eav, sbd\}$ via softmax. The activation function and all layer dimensions are treated as HPs and will be determined by HPO. All three loss components are optimized jointly through:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \beta \cdot D_{\text{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) \\ &\quad + \lambda_{\text{rcst}} \cdot \mathcal{L}_{\text{rcst}}(\mathbf{x}, \mathbf{x}') + \lambda_{\text{clf}} \cdot \mathcal{L}_{\text{clf}}(\hat{y}, y),\end{aligned} \tag{10}$$

where D_{KL} measures the divergence between the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$, inferred by the encoder, and the standard Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, regularizing the latent space toward a compact and well-structured distribution; $\mathcal{L}_{\text{rcst}} = \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{x}_i - \mathbf{x}_i'\|^2$ is the Mean Squared Error (MSE) over a mini-batch of B samples between the input and its reconstruction, ensuring that $\mathbf{z}$ retains sufficient information to faithfully recover the original SOP features; and $\mathcal{L}_{\text{clf}} = -\frac{1}{B}\sum_{i=1}^{B} \log \hat{p}_{i,y_i}$ is the unweighted Cross Entropy (CE) loss over the same mini-batch, where $\hat{p}_{i,y_i}$ is the predicted probability assigned to the true class y_i of sample i, enforcing class discriminability directly within the latent space. The weighting coefficient β controls the strength of the KL regularization relative to the other objectives, λ_{rcst} balances reconstruction fidelity against the classification pressure, and λ_{clf} governs the contribution of the supervised signal in shaping the latent representation. All three weighting coefficients, together with all remaining HPs, are determined through Bayesian HPO using the Optuna framework, as described in Section 5.C.

The framework is instantiated in two operational modes, each addressing a distinct generalization objective. Both modes share the same architectural template and loss formulation; their distinction lies solely in the scope of the input distribution the encoder is designed to handle.

B.1. Single-System VAE (VAE$_{sgl}$): Per-System Reference Mode

In the VAE$_{\text{sgl}}$ mode, two independent VAE models are constructed, one per system, each with its own encoder, decoder, and classification head. Each model produces a latent space conditioned exclusively on the statistics of its source domain, yielding strong intra-system discriminability under S_1 and S_2 and serving as a per-system performance reference. Since the two models are structurally isolated, their latent spaces are inherently domain-specific and carry no mechanism to bridge the marginal distribution mismatch of Eq. (1), making cross-system scenarios S_3 and S_4 a direct probe of how much class-discriminative structure is incidentally preserved under distribution shift.

B.2. Combined-System VAE (VAE$_{cmb}$): Combined Cross-System Training Mode

In the VAE$_{\text{cmb}}$ mode, a single shared VAE is constructed whose encoder simultaneously processes inputs from both systems, producing a unified latent space $\mathbf{z}$ that must account for the full cross-domain variability. The joint pressure from the reconstruction and KL terms in Eq. (10) encourages the encoder to suppress system-specific variation and organize $\mathbf{z}$ around the class-discriminative structure that is consistent over both systems, in accordance with Eq. (2). A separate classification head is trained and attached for each evaluation scenario, allowing the transferability of the shared latent representation to be probed independently under all four intra- and cross-system conditions.

Table 1. Dataset and sample count summary per class and their allocation for training, validation, and test splits.

System	Class	Total	Train	Val	Test
System 1 (O-band)	rlx_1	4,800	3,360	720	720
	eav_1	4,800	3,360	720	720
	sbd_1	4,800	3,360	720	720
System 2 (C-band)	rlx_2	4,800	3,360	720	720
	eav_2	644	450	97	97
	sbd_2	773	541	116	116
Combined	O-band	14,400	10,080	2,160	2,160
	C-band	6,217	4,351	933	933
	All	**20,617**	**14,431**	**3,093**	**3,093**

5. ML ANALYSIS

This section describes the data preprocessing pipeline, the HP search spaces for the DNN and VAE frameworks, and the Bayesian HPO setup used to identify optimal configurations.

A. Data Preprocessing

Since the quality of the raw data directly affects the reliability of the subsequent ML analysis, the following preprocessing steps are applied on the collected data in Section 3.A.

A.1. Step 1: System 2 Activity Filtering

Since physical perturbations are applied manually and intermittently during data collection, the eav_2 and sbd_2 recordings naturally contain intervals in which the fiber remains relaxed between consecutive events. These segments are removed to retain only samples corresponding to active physical disturbances. This filtering step enhances the separability of each class's event-specific spectral characteristics from the baseline rlx_2 state. After filtering, the System 2 dataset becomes imbalanced, and contains 644 eav_2 samples and 773 sbd_2 samples, while all 4,800 rlx_2 samples are retained unchanged.

A.2. Step 2: Domain-Aware Stratified Dataset Partitioning

Following preprocessing, the System 1 and System 2 datasets are merged into a unified dataset of 20,617 samples, each represented by a 512-D spectral feature vector with two associated labels: an event class label $y \in \mathcal{Y} = \{0, 1, 2\}$ and a domain label $d \in \{0, 1\}$ identifying the source system. The dataset is then partitioned using two-stage stratified sampling on the joint key (y, d), yielding 14,431 training (70%), 3,093 validation (15%), and 3,093 test (15%) samples, with identical class and domain proportions preserved across all splits, including the System 2 class imbalance. Per-system views are extracted from each split to support all four evaluation scenarios (S_1–S_4) without re-splitting or data duplication. Since all models share the same splits, cross-model comparisons remain directly comparable. Table 1 summarizes the resulting sample counts.

A.3. Step 3: Feature Normalization

Finally, z-score standardization is applied to each of the spectral features to ensure stable and unbiased model training. For a feature value x, the normalized value $\hat{x}$ is computed as $\hat{x} = \frac{x-\mu_{\text{train}}}{\sigma_{\text{train}}}$, where μ_{train} and σ_{train} are the feature-wise mean and standard deviation computed exclusively from the training set. The same statistics are then applied to normalize the validation and test sets.

B. ML Analysis of DNN Model for HPO

The DNN model exposes two categories of HPs: architectural and optimization. The architectural HPs govern the depth and width of the residual backbone. The number of RBs, N_{dnn}, is varied over $\{2, 3, 4, 5\}$. The width of the first hidden layer is searched over $\{256, 320, \ldots, 768\}$ (step 64). The activation function $f(\cdot)$ in each RB is selected from {GELU, ReLU, ELU, SiLU}, the per-block DP probability p is searched over [0.10, 0.50], and BN is applied with fixed default parameters (momentum = 0.1, $\varepsilon = 10^{-5}$, learnable affine transform).

The optimization HPs include the loss function, optimizer, learning rate η, weight decay λ, Learning Rate (LR) schedule, and mini-batch size. The loss function is selected from two candidates: CE with optional Label Smoothing (LS) (smoothing parameter $\alpha \in [0.0, 0.15]$, where $\alpha = 0$ recovers standard CE), which controls the degree of prediction confidence; and Focal Loss (FL) (focusing parameter $\gamma \in [0.5\,5.0]$) [39], which down-weights easy examples to focus training on hard or misclassified samples. The optimizer is chosen from {AdamW, SGD-Nesterov, RMSprop}: Adaptive Moment Estimation with Decoupled Weight Decay (AdamW) decouples weight decay from the gradient update and adapts per-parameter LRs, making it well-suited to sparse or heterogeneous gradients; Stochastic Gradient Descent (SGD) with Nesterov momentum offers strong generalization when carefully tuned; and Root Mean Square Propagation (RMSprop) adapts step sizes via a running average of squared gradients, which can stabilize training under noisy or non-stationary objectives. The LR, i.e. η is searched over $[10^{-4}, 5 \times 10^{-3}]$ (log-uniform) and weight decay λ over $[10^{-6}, 10^{-3}]$ (log-uniform); momentum is additionally searched over [0.80, 0.99] when SGD or RMSprop is selected. The LR schedule is chosen between a linear warm-up followed by cosine annealing (warm-up duration $\in [5, 20]$ epochs) and a plateau-based reduction (reduce on plateau, factor 0.5). Finally, the mini-batch size is selected from $\{32, 64, 128, 256\}$.

C. ML Analysis of VAE Model for HPO

The VAE search space is organized into three categories: architectural, loss-weighting, and optimization HPs.

The *architectural* HPs cover the encoder depth (number of hidden layers), width (number of neurons in each layer), the latent dimensionality, activation function, and regularization strategy. The number of hidden layers N_{vae} is searched over $\{1, 2, 3, 4\}$, and each hidden layer width is sampled independently from $\{64, 128, 192, \ldots, 768\}$ (step 64, 12 values). The latent dimensionality d_z is selected from $\{8, 16, 32, 64, 128, 256\}$, the activation function is selected from {LeakyReLU, ReLU, GELU, ELU, SiLU}, BN is treated as a binary HP (enabled or disabled), and the DP probability p is searched over [0.0, 0.6] (uniform).

The *loss-weighting* HPs are specific to the VAE objective of Eq. (10): the KL weight β is searched over $[10^{-4}, 5.0]$ (log-uniform), the reconstruction weight λ_{rcst} over [0.01, 5.0] (uniform), and the classification weight λ_{clf} over [0.1, 5.0] (uniform). The β warm-up duration is searched over [0, 60] epochs, where a value of zero disables cosine annealing entirely and allows HPO to determine whether gradual KL introduction is beneficial for a given configuration.

The *optimization* HPs govern the the optimizer, learning rate η, weight decay λ, mini-batch size, and LR schedule. The optimizer is selected from {AdamW, Adam, SGD-Nesterov, RMSprop}, with momentum additionally searched over [0.80, 0.99] when SGD or RMSprop is selected. The learning rate η is searched over $[10^{-5}, 10^{-2}]$ (log-uniform), weight decay λ over $[10^{-7}, 10^{-2}]$ (log-uniform), mini-batch size from $\{32, 64, 128, 256\}$. The LR schedule is selected from three candidates: plateau-based reduction (factor 0.5, patience 10 epochs), cosine annealing, and one-cycle LR.

D. HPO Setup and HP Search Space

HPs for the DNN and VAE frameworks are tuned using *Optuna* with a Tree-structured Parzen Estimator (TPE) sampler [40], and a Median Pruner (10 startup trials, 15 warm-up steps) to terminate unpromising trials early. For the DNN and VAE_{sgl}, two independent optimization rounds are conducted per model: one optimized on the System 1 training and validation sets (used for S_1 and S_3) and one on the System 2 counterparts (used for S_2 and S_4), to maximize source-domain validation accuracy. For VAE_{cmb}, a single optimization round is conducted on the combined system training and validation sets, to maximize the mean of the System 1 and System 2 validation accuracies, ensuring that the model generalizes well to both domains simultaneously rather than a single one. Each study runs for 100 trials, with each trial training for at most 150 epochs, subject to early stopping with patience of 20 epochs. Following HPO, the best configuration identified for each study is used to retrain a final model from scratch on the full training set, with a maximum of 500 epochs and early-stopping patience of 40 epochs. All experiments were implemented in Python using the PyTorch DL framework (v≥2.0) [41] and executed on a system equipped with NVIDIA H100 80GB GPUs.

6. RESULTS AND DISCUSSION

This section presents and discusses the experimental results for four evaluation scenarios: intra-system S_1 and S_2, and cross-system S_3 and S_4. Results are reported in two stages: HPO-selected architectures and training dynamics are analyzed first, followed by a comprehensive performance evaluation.

A. Best Selected HPs

A.1. Best-Found DNN Architectures

Table 2 summarizes the HPO-selected architecture and training configuration for the System 1 and System 2 of DNN models.

The two models differ in both their depth and width, reflecting the distinct characteristics of each domain. The System 1 model adopts a deep, narrow configuration in which the representation is aggressively compressed after the first block and then refined at a fixed width across four subsequent blocks. The System 2 model, by contrast, selects a shallower but substantially wider architecture. Both models use a SiLU activation function and a plateau-based LR schedule, while differing in their optimizers and batch sizes. The validation accuracy gap between the two models, 85.79% for System 1 vs. 97.64% for System 2, indicates that System 1 presents a more challenging classification task, consistent with the greater spectral overlap

Table 2. HPO-selected configuration for the best-found DNN models.

Hyperparameter	System 1 (S_1, S_3)	System 2 (S_2, S_4)
Architecture		
N_{dnn}	5	3
Layer widths	$512 \to 256 \to 64^{\times 4} \to 3$	$512 \to 704 \to 368 \to 92 \to 3$
Parameters	161,475	657,147
Activation $f(\cdot)$	SiLU	SiLU
Dropout p	0.233	0.267
Optimization		
Loss	CE+LS	CE+LS
LS α	0.085	0.082
Optimiser	SGD (Nesterov)	AdamW
Learning rate η	3.36×10^{-3}	3.84×10^{-4}
Weight decay λ	1.54×10^{-6}	3.99×10^{-6}
Momentum	0.803	—
LR schedule	Plateau ($\times 0.5$)	Plateau ($\times 0.5$)
Batch size	32	128
Result		
Best val. accuracy	85.79%	97.64%

Table 3. HPO-selected configuration for the best-found VAE_{sgl} models.

Hyperparameter	System 1 (S_1, S_3)	System 2 (S_2, S_4)
Architecture		
N_{vae}	2	4
Layer widths	$512 \to 704 \to 640 \to 32$	$512 \to 384 \to 320^{\times 2} \to 192 \to 32$
Latent dim d_z	32	32
Activation $f(\cdot)$	GELU	SiLU
Batch norm	Yes	Yes
Dropout p	0.445	0.167
Loss Weighting		
KL weight β	0.100	0.294
Recon. weight λ_{rcst}	3.886	2.530
Classif. weight λ_{clf}	3.248	1.414
β warm-up (epochs)	43	23
Optimization		
Optimizer	AdamW	Adam
Learning rate η	2.96×10^{-4}	1.62×10^{-3}
Weight decay λ	3.27×10^{-6}	9.10×10^{-5}
LR schedule	Plateau ($\times 0.5$)	Plateau ($\times 0.5$)
Batch size	256	256
Result		
Best val. accuracy	84.75%	97.50%

between the *eav* and *sbd* signatures observed in the balanced O-band dataset, whereas the filtered and imbalanced System 2 dataset retains more spectrally distinct event signatures.

Both final DNN models converge without a notable generalization gap between training and validation sets, with System 1 converging faster (approximately 55 epochs) than System 2 (approximately 115 epochs), consistent with its smaller architecture.

A.2. Best-Found Single-System VAE Models

Table 3 summarizes the HPO-selected configurations for the two VAE_{sgl} models, which reflect the distinct statistical properties of each system.

Several contrasts are noteworthy. The lower DP for System 2 reflects the need to preserve the scarce minority-class signal under class imbalance, where aggressive regularization would suppress the few *eav* and *sbd* samples. The higher β for System 2 indicates that the more complex, imbalanced distribution requires stronger KL regularization to maintain a well-structured posterior. The β warm-up in both cases confirms that gradual KL introduction is beneficial in both systems, consistent with the known instability of jointly optimizing reconstruction and classification objectives from the first epoch. The best validation accuracies of 84.75% and 97.50% closely match the corresponding DNN baselines shown in Table 2, confirming competitive intra-system discriminability for both VAE_{sgl} models. Both models converge within approximately 100 epochs without a notable generalization gap between training and validation sets.

A.3. Best-Found Combined-System VAE Model

In the combined-system VAE models, training follows a two-phase design. In Phase 1, the complete model is trained on combined data from both systems, optimizing the combined objective in Eq. (10) so that the encoder learns a latent space aligned with the class-discriminative structure consistent with both systems, in accordance with Eq. (2). In Phase 2, the encoder is frozen, and a randomly initialized classification head is trained per scenario using only source-domain (single system) labels, with no access to target-domain (cross-system) class information. This design bridges the distribution mismatch in Eq. (1) before any classifier is trained, without requiring target-domain supervision, an advantage structurally unavailable to the DNN baseline and previous work in the literature.

Table 4 summarizes the HPO-selected configuration for VAE_{cmb}, whose single optimization round targets the combined System 1 and System 2 training and validation sets. The selected 3-layer encoder $512 \to 128 \to 768 \to 768 \to 32$ contains 3,156,099 parameters. The initial compression to 128 units acts as an early domain-agnostic filter before the wider layers learn a shared cross-domain representation. The KL weight $\beta = 0.279$, consistent with the System 2 VAE_{sgl} (0.294) and higher than the System 1 counterpart (0.100), confirms that combined training demands KL regularization comparable to that of the more complex System 2 distribution alone. The classification weight $\lambda_{\text{clf}} = 2.372$ exceeding $\lambda_{\text{rcst}} = 1.161$ reflects deliberate emphasis on class-discriminative pressure. Cosine annealing is selected over the plateau-based schedule of both VAE_{sgl} models, providing smoother LR decay under the noisier joint-domain loss landscape. The weight decay of 1.42×10^{-4}, one order of magnitude larger than either VAE_{sgl} model, reflects the stronger L_2 regularization required for the larger parameter count and more diverse training distribution.

The model converges at epoch 30, achieving a best combined

Table 4. HPO-selected configuration for the best-found VAE_{cmb} model.

Hyperparameter	**Combined (S_1–S_4)**
Architecture	
N_{vae}	3
Layer widths	$512 \to 128 \to 768 \to 768 \to 32$
Latent dim d_z	32
Activation $f(\cdot)$	ReLU
Batch norm	Yes
Dropout p	0.114
Loss Weighting	
KL weight β	0.279
Recon. weight λ_{rcst}	1.161
Classif. weight λ_{clf}	2.372
β warm-up (epochs)	34
Optimization	
Optimizer	Adam
Learning rate η	1.87×10^{-3}
Weight decay λ	1.42×10^{-4}
LR schedule	cosine
Batch size	64
Result	
Best val. accuracy	91.79%

validation accuracy of 91.79%, with no notable generalization gap between training and validation sets despite the 3.15M-parameter capacity.

B. Experimental Results and Performance Analysis

We first compare the overall accuracy of the frameworks across the four scenarios. We then analyze its classification behavior in depth through per-scenario confusion matrices and per-class comparisons against the DNN baseline.

B.1. Overall Accuracy Comparison over the Four Scenarios

Fig. 5 compares the test-set accuracy of all three models over all four scenarios. Under intra-system conditions, all three approaches perform comparably. VAE_{cmb} achieves 85.3% (S_1) and 97.7% (S_2) accuracy, marginally surpassing the DNN (82.9% for (S_1), 97.3% for (S_2)) and VAE_{sgl} (84.5% for (S_1), 97.4% for (S_2)). This confirms that combined-domain training does not compromise source-domain discriminability.

The central finding emerges under domain shift. As visible in Fig. 5, both the DNN and VAE_{sgl} collapse well below the 33.3% random-chance baseline in the cross-system scenarios, reaching only 11.7% and 11.9% on S_3, and 22.5% and 34.0% on S_4, respectively. This failure is structural: both models produce domain-specific latent spaces that lack a mechanism to bridge the marginal distribution mismatch in Eq. (1). VAE_{cmb}, by contrast, achieves 95.3% on S_3 and 73.5% on S_4, which are absolute gains of 83.4 and 51 percentage points over the DNN baseline. This demonstrates that combined VAE training enables learning a domain-invariant latent representation that supports accurate cross-system classification. The remainder of the analysis focuses on the VAE_{cmb} performance in detail.

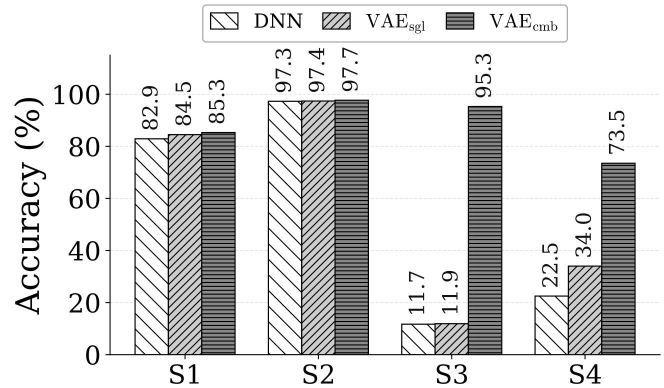


Fig. 5. Test-set accuracy comparison across all four evaluation scenarios for the DNN baseline, VAE_{sgl}, and VAE_{cmb}.

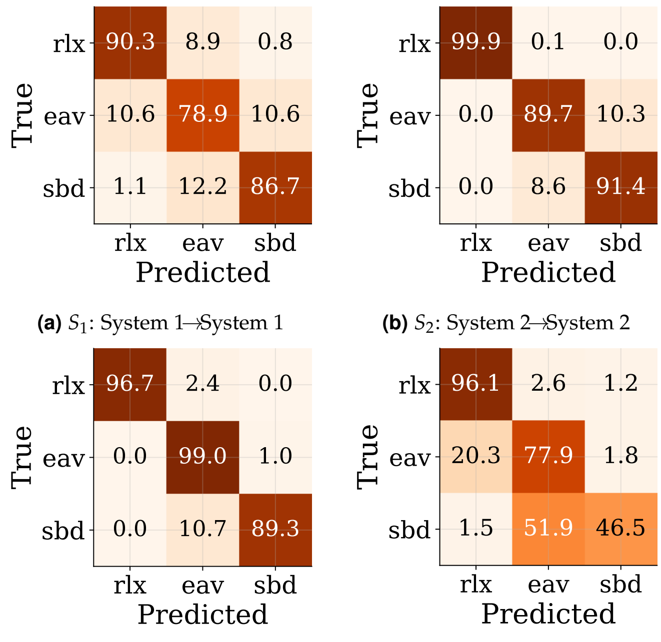
B.2. VAE_{cmb} Performance Analysis

Scenario-Level Analysis via Confusion Matrices. Fig. 6 reports the confusion matrices of VAE_{cmb} for all four scenarios. In the intra-system scenarios, accuracy for S_1 is high across all classes, with the modest *eav* prediction rate drop.In the intra-system scenario S_2, the model achieves near-perfect performance in detecting the normal state, but we still observe misclassifications between *eav* and *sbd*.

For the highly-challenging cross-system scenario S_3, VAE_{cmb} classifies *rlx* and *eav* with near-perfect accuracy (97.6% and 99.0%). *sbd* is correctly detected in 89.3% of cases at a residual confusion with *eav* (10.7%), likely reflecting the inherent spectral similarity between gradual bending and eavesdropping. Nonetheless, this represents a complete recovery from the baseline's low performance. In S_4, *rlx* is classified with high accuracy of 96.1% and *eav* with 77.9%. Here, *sbd* at only 46.5% accuracy remains the most challenging class, with 51.9% of samples misclassified as *eav*.

Per-Class Comparison with the DNN Baseline. Fig. 7 compares the per-class accuracy of VAE_{cmb} and the DNN baseline across all four scenarios. In the intra-system settings (S_1, S_2), both models achieve comparable per-class rates, with VAE_{cmb} providing slight improvements, particularly for the minority classes *eav* and *sbd*, suggesting that the regularized latent space marginally improves generalization even within a single domain. However, the per-class breakdown under cross-system conditions (S_3, S_4) highlights the fundamental research gap addressed by this work: the DNN baseline achieves near-zero accuracy on *rlx* and *sbd* in S_3, and severely degraded rates across all classes in S_4. VAE_{cmb} recovers all three classes to operationally meaningful levels of accuracy in both cases, with the *sbd* class in S_4 remaining the primary challenge for future improvement. This result establishes VAE_{cmb} as an effective solution to the cross-system generalization problem for SOP-based monitoring of physical-layer security breaches.

The primary remaining challenge is the *sbd*–*eav* confusion in the System 2→System 1 direction, reflecting the greater representational complexity of the live metro ring relative to the dark fiber testbed.



(a) S_1: System 1→System 1

(b) S_2: System 2→System 2

(c) S_3: System 1→System 2

(d) S_4: System 2→System 1

Fig. 6. Confusion matrices of VAE$_{cmb}$ for all four evaluation scenarios. Values are per-class accuracy (%).

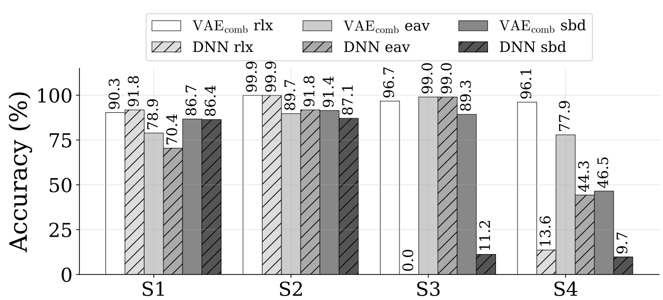


Fig. 7. Per-class accuracy of VAE$_{\text{cmb}}$ vs. the DNN baseline across all four scenarios.

7. CONCLUSION

This paper introduced a VAE-based DA framework for cross-system generalization in ML-driven SOP-based fiber monitoring, addressing the failure of existing supervised approaches to transfer across optical systems with different spectral bands, fiber links, and network topologies. Three models were evaluated over four intra- and cross-system scenarios using real-world SOP data from two fundamentally different optical links and topologies. While all models achieved comparable intra-system accuracy, both the supervised DNN baseline and VAE$_{\text{sgl}}$ collapsed to below random-chance accuracy under cross-system transfer. By contrast, the proposed VAE$_{\text{cmb}}$, trained on combined data from both systems through a two-phase design, achieved meaningful accuracy in both transfer directions.

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon Europe research and innovation programme (Grant No. 10113933) through the ICON project and the Digital Europe Programme (Grant No. 101127973) through the 5G-TACTIC project, as well as by the CELTIC-NEXT SUSTAINET-Advance project, Taighde Éireann – Research Ireland under Grant No. 18/RI/5721 (OpenIreland Research Infrastructure), 13/RC/2106_P2 (ADAPT centre) and the Swedish Research Council (Grant No. 2023–05249).

REFERENCES

1. D. Rafique, T. Szyrkowiec, H. Grießer, A. Autenrieth, and J.-P. Elbers, "Cognitive assurance architecture for optical network fault management," J. Light. Technol. **36**, 1443–1450 (2018).
2. G. Allwood, G. Wild, and S. Hinckley, "Optical fiber sensors in physical intrusion detection systems: A review," IEEE Sensors J. **16**, 5497–5509 (2016).
3. S. Pellegrini, L. Minelli, L. Andrenacci, G. Rizzelli, D. Pilori, G. Bosco, L. Della Chiesa, C. Crognale, S. Piciaccia, and R. Gaudino, "Overview on the state of polarization sensing: Application scenarios and anomaly detection algorithms," J. Opt. Commun. Netw. **17**, A196–A209 (2025).
4. P. Lu, N. Lalam, M. Badar *et al.*, "Distributed optical fiber sensing: Review and perspective," Appl. Phys. Rev. **6** (2019).
5. C. J. Carver and X. Zhou, "Polarization sensing of network health and seismic activity over a live terrestrial fiber-optic cable," Commun. Eng. **3**, 91 (2024).
6. L. Sadighi, S. Karlsson, C. Natalino, and M. Furdek, "Machine learning-based polarization signature analysis for detection and categorization of eavesdropping and harmful events," in *Optical Fiber Communications Conference and Exhibition (OFC),* (2024), p. M1H.1.
7. L. Sadighi, C. Natalino, S. Karlsson, M. Ruffini, E. Kenny, L. Wosinska, and M. Furdek, "Generalizability of ML-based classification of state of polarization signatures across different bands and links," in *2025 European Conference on Optical Communications (ECOC),* (2025), pp. 1–4.
8. S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowl. Data Eng. **22**, 1345–1359 (2010).
9. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *2nd International Conference on Learning Representations (ICLR),* (Banff, AB, Canada, 2014).
10. A. Rode, M. Farsi, V. Lauinger, M. Karlsson, E. Agrell, L. Schmalen, and C. Häger, "Machine learning opportunities for integrated polarization sensing and communication in optical fibers," Opt. Fiber Technol. **89**, 103924 (2025).
11. L. Sadighi, S. Karlsson, L. Wosinska, and M. Furdek, "Machine learning analysis of polarization signatures for distinguishing harmful from non-harmful fiber events," in *International Conference on Transparent Optical Networks (ICTON),* (2024).
12. L. Sadighi, S. Karlsson, C. Natalino, L. Wosinska, M. Ruffini, and M. Furdek, "Detection and classification of eavesdropping and mechanical vibrations in fiber optical networks by analyzing polarization signatures over a noisy environment," in *European Conference on Optical Communication (ECOC),* (2024), pp. 527–530.
13. L. Sadighi, S. Karlsson, C. Natalino, L. Wosinska, M. Ruffini, and M. Furdek, "Deep learning for detection of harmful events in real-world, noisy optical fiber deployments," J. Light. Technol. **43**, 6092–6101 (2025).
14. A. Tomasov, P. Dejdar, P. Munster, T. Horvath, P. Barcik, and F. Da Ros, "Enhancing fiber security using a simple state of polarization analyzer and machine learning," Opt. & Laser Technol. **167**, 109668 (2023).
15. K. Abdelli, M. Lonardi, J. Gripp, D. Correa, S. Olsson, F. Boitier, and P. Layec, "Vision transformers for anomaly classification and localization in optical networks using SOP spectrograms," J. Light. Technol. (2025).
16. S. Karlsson, M. Andersson, R. Lin, L. Wosinska, and P. Monti, "Detection of abnormal activities on a SM or MM fiber," in *Optical Fiber Communication Conference (OFC),* (Optica Publishing Group, San Diego, CA, USA, 2023), p. M3Z.6.
17. W. Qin, Q. Zhang, W. Hou, X. Zhang, and X. Gong, "Convolutional neural networks for fiber-bending eavesdropping attacks detection in

coherent optical communication systems," in *International Conference on Ubiquitous Communication (Ucom),* (Xi'an, China, 2024).
18. L. Sadighi, S. Karlsson, C. Natalino, and M. Furdek, "Ml-based state of polarization analysis to detect emerging threats to optical fiber security," IEEE Transactions on Netw. Serv. Manag. **23**, 432–442 (2026).
19. H. Song, R. Lin, L. Wosinska, P. Monti, M. Zhang, Y. Liang, Y. Li, and J. Zhang, "Cluster-based unsupervised method for eavesdropping detection and localization in WDM systems," J. Opt. Commun. Netw. **16**, F52–F61 (2024).
20. L. Minelli, S. Pellegrini, L. Andrenacci, D. Pilori, G. Bosco, L. Della Chiesa, A. Tanzi, C. Crognale, and R. Gaudino, "SOP-based DSP blind anomaly detection for sensing on deployed metropolitan fibers," in *European Conference on Optical Communications (ECOC),* (Glasgow, UK, 2023), pp. 519–522.
21. C. Rottondi, R. di Marino, M. Nava, A. Giusti, and A. Bianco, "On the benefits of domain adaptation techniques for quality of transmission estimation in optical networks," J. Opt. Commun. Netw. **13**, A34–A43 (2021).
22. F. Musumeci, V. Garbhapu Venkata, Y. Hirota, Y. Awaji, S. Xu, M. Shiraiwa, B. Mukherjee, and M. Tornatore, "Domain adaptation and transfer learning for failure detection and failure-cause identification in optical networks across different lightpaths [invited]," J. Opt. Commun. Netw. **14**, A91–A100 (2022).
23. W. Mo, Y.-K. Huang, S. Zhang, E. Ip, D. C. Kilper, Y. Aono, and T. Tajima, "Ann-based transfer learning for qot prediction in real-time mixed line-rate systems," in *2018 Optical Fiber Communications Conference and Exposition (OFC),* (2018), pp. 1–3.
24. J. Yu, W. Mo, Y.-K. Huang, E. Ip, and D. C. Kilper, "Model transfer of QoT prediction in optical networks based on artificial neural networks," J. Opt. Commun. Netw. **11**, C48–C57 (2019).
25. Z. Cai, Q. Wang, Y. Deng, P. Zhang, G. Zhou, Y. Li, and F. N. Khan, "Domain adversarial adaptation framework for few-shot QoT estimation in optical networks," J. Opt. Commun. Netw. **16**, 1133–1144 (2024).
26. K. Abdelli, M. Lonardi, J. Gripp, S. Olsson, F. Boitier, and P. Layec, "Risky event classification leveraging transfer learning for very limited datasets in optical networks," J. Opt. Commun. Netw. **16**, C51–C68 (2024).
27. M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoders," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning,* vol. 121 of *Proceedings of Machine Learning Research* (PMLR, 2020), pp. 322–348.
28. C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel, "The variational fair autoencoder," in *4th International Conference on Learning Representations (ICLR),* (2016).
29. W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* (IEEE, 2017), pp. 16–23.
30. CONNECT Centre for Future Networks and Communications, "Open Ireland Testbed," (2020). Available here.
31. ASIERA (formerly HEAnet), "Ireland's National Education and Research Network," (2025). Available here.
32. L. Sadighi, C. Natalino, S. Karlsson, L. Wosinska, E. Kenny, M. Ruffini, and M. Furdek, "Modulated vs. unmodulated polarization signatures for ml-based fiber sensing," J. Light. Technol. pp. 1–10 (2026).
33. S. Karlsson, "A method for detecting external influence on an optical cable," (1994). Filed Mar. 9, 1989; granted Oct. 19, 1994.
34. P. Podder, T. Z. Khan, M. H. Khan, and M. M. Rahman, "Comparative performance analysis of hamming, hanning and blackman window," Int. J. Comput. Appl. **96**, 1–7 (2014).
35. S. Karlsson, R. Lin, L. Wosinska, and P. Monti, "Eavesdropping G.652 vs. G.657 fibres: A performance comparison," in *2022 International Conference on Optical Network Design and Modelling (ONDM),* (IEEE, 2022), pp. 1–3.
36. L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." J. machine learning research **9** (2008).
37. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML),* (PMLR, 2015), pp. 448–456.
38. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. (JMLR) **15**, 1929–1958 (2014).
39. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* (IEEE, 2017), pp. 2980–2988.
40. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD),* (ACM, 2019), pp. 2623–2631.
41. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS),* vol. 32 (Curran Associates, Inc., 2019).