Assessing Robustness of EEG Representations under Data-shifts via Latent Space and Uncertainty Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The recent availability of large datasets in bio-medicine has inspired the development of representation learning methods for multiple healthcare applications. Despite advances in predictive performance, the clinical utility of such methods is limited when exposed to real-world data. Here we develop model diagnostic measures to detect potential pitfalls during deployment without assuming access to external data. Specifically, we focus on modeling realistic data shifts in electrophysiological signals (EEGs) via data transforms, and extend the conventional task-based evaluations with analyses of a) model's latent space and b) predictive uncertainty, under these transforms. We conduct experiments on multiple EEG feature encoders and two clinically relevant downstream tasks using publicly available large-scale clinical EEGs. Within this experimental setting, our results suggest that measures of latent space integrity and model uncertainty under the proposed data shifts may help anticipate performance degradation during deployment.

1 Introduction

In recent years, the availability of large datasets in bio-medicine has inspired the development of deep representation learning methods for diagnostic applications in healthcare domains such as cardiology, neurology, and radiology, to name a few. However, there are increasing examples of failures of such systems during real-world deployment, predominantly due to dataset/distribution shifts in the deployment setting (1; 2; 3; 4; 5; 6). Currently, the most common approach undertaken to evaluate healthcare machine learning (ML) models is analyzing task performance in a held-out test dataset. The recent advances in covariate-shift detection and adaptation approaches (7; 8; 9; 10; 11) have not been widely adopted in the healthcare ML community because they do not reflect the unique health-data-related challenges. As such, the practical value of many healthcare ML models remains unclear at present without additional assessments of model resilience in deployment settings (12). This gap reflects a major hurdle in translating healthcare ML models as they are required to perform consistently under varying noise characteristics, diverse data acquisition protocols, and multiple sites and populations, during deployment. Hence, there is a critical need to evaluate the robustness of healthcare ML models under realistic dataset shifts prior to deployment.

In this paper, we focus on a particular area of healthcare ML pertaining to learning representations of physiological signals such as electroencephalograms (EEGs). EEGs are commonly used to diagnose various neurological, psychiatric, and sleep disorders, as well as in applications involving brain-machine interfaces. Deep representation learning of raw EEG signals has recently gained popularity because of the availability of large-scale EEG datasets (13) and has shown promise in improving the labor-intensive and error-prone manual process undertaken in clinical EEG reviews. A variety of approaches have been proposed including, fully-supervised deep feature encoders (14; 15; 16), self-supervised encoders (17; 18; 19), and traditional feature-based encoders using power spectral features (20). However, many of those approaches struggle to generalize to previously unseen real-world data, i.e., data acquired from different populations, institutions, or devices (21; 22).

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We believe that a multi-pronged approach is necessary to holistically evaluate the robustness of healthcare ML models prior to deployment. First, models of realistic data shifts are needed to assess robustness during model development itself. For instance, realistic EEG data shifts may include differences in behavioral states during EEG acquisition (awake or asleep), the number of EEG sensors used, the EEG reference used, noise during acquisition, analog-to-digital conversion, hardware filter settings, and impedance resulting from electrodes. Second, we argue that robustness evaluations may also benefit from analyzing the properties of the latent embedding space (7). Specifically, quantifying the topological changes in the latent space under perturbations can, 1) provide early clues into test set performance when ground truth labels are not available; and 2) indicate failure modes of model interpretation techniques that rely on analyzing similarity properties in the latent space. Third, we highlight the need for estimating the uncertainty in model predictions at the output layer (23). Uncertainty modeling is critical in ensuring the safe and reliable use of healthcare models since accurate communication of uncertainty allows human experts to selectively ignore the model.

In this study, we evaluated the robustness of several previously proposed EEG feature encoders using the above multi-pronged approach. Specifically, we introduced four EEG signal transformations that model the effects of instrumentation-related real-world EEG variability, namely: hardware band-pass filters, quantization precision, narrow-band impedance noise, and broad-band Gaussian noise. Next, we developed a topological measure of latent space integrity based on Delaunay neighborhood graphs (adapted from (24)) and a Monte Carlo dropout-based approach to estimate predictive uncertainty. In this setting, we then evaluated fully-supervised, self-supervised, and traditional feature-based EEG encoders in both classification and regression tasks. Our major contributions are the following:

1. We developed four domain-guided EEG data shifts (EEG-DS) reflecting instrumentation-related variability observable in real-world deployment.
2. Using EEG-DS, we devised a multi-pronged approach to evaluate the robustness of multiple ML models including assessments of latent space integrity and predictive uncertainty.
3. Using two large-scale EEG datasets (TUH (13) and NMT (25)), we show that the proposed approach can help anticipate in-the-wild performance during model development itself and that popular EEG encoders exhibit variable robustness profiles.

2 Related Work

Generalization to unseen data: The importance of generalization in healthcare ML is widely acknowledged (1; 5; 6; 26). Prior work in general ML has proposed, a) methods to detect covariate-shift or out-of-distribution samples (7; 27; 28); and b) novel learning approaches that generalize well or adapt to new domains (29; 30; 31; 8; 9; 10). However, the majority of the adaptive approaches require examples from target distributions and therefore have limited utility. Some methods to detect covariate-shift focus on specific properties such as non-stationarity (27; 28), which are simplistic and do not reflect the complexity of healthcare data and shifts. In this paper, we focus on detecting failure modes prior to deployment via topological analysis of the latent space and evaluation of model uncertainty under realistic data shifts. In this context, the closest to our work is (7), where authors used measures such as Wasserstein's distance for detecting data shifts in the latent space for vision and language tasks. Here we develop a more comprehensive topological evaluation of the latent space using Delaunay neighborhood graphs and use it to analyze the robustness of scalp EEG representations. In addition, some recent studies have highlighted the benefits of uncertainty quantification in out-of-distribution settings, in knowing when not to trust a model's predictions (32). Some studies have also indicated that certain types of representation learning approaches exhibit better robustness in terms of predictive uncertainty than others (33). Building on these insights, we utilized a Monte Carlo dropout (MCD)-based uncertainty quantification to analyze the predictive uncertainty of different EEG feature encoders under specific data shifts.

Data shifts in EEG data: Several previous studies have analyzed the performance of EEG ML models using datasets from multiple institutions (21; 22; 34). Apart from reporting differences in model performance, these studies did not investigate approaches to identify failure modes prior to deployment, model data shifts, or generate insights for future studies. Some studies have developed data augmentations for EEG signals to improve the diversity of training data and have demonstrated some success in achieving better task performance (18; 35; 36). While augmentations such as amplitude scaling, time shift, DC shift, and band-stop filtering are useful to increase training data

diversity (18), they do not meet the complexity of real-world EEG signal variability observed in deployment settings. Hence, in this paper, we focus on developing domain-guided data shifts observable in deployment settings as opposed to simpler training time data augmentations.

EEG feature encoders: A variety of EEG feature encoders have been proposed including, fully-supervised deep encoders (14; 15; 16), self-supervised encoders (17; 18; 19), and traditional domain-derived feature encoders using power spectral features (20; 37). While the generalization capabilities of the majority of these encoders are unclear, an encoder derived based on self-supervised learning has shown promising trends in generalization when specific domain-derived self-supervision tasks were utilized (17). However, a comprehensive robustness evaluation of the different types of EEG feature encoders under realistic EEG data shifts is currently lacking.

3 Background & Methods

In this section, we describe the formulations of multichannel EEG data, feature encoders, and the proposed robustness evaluation approach including realistic EEG data shifts (EEG-DS), latent space analysis, and uncertainty quantification. The proposed approach does not require additional data or labels obtained in deployment settings and therefore is applicable during model development.

3.1 Learning EEG data representations

We use $x_i \in \mathbb{R}^{M \times N}$ to denote the EEG data of a single epoch recorded using M sensors with N samples (i.e., $N =$ time duration $\times$ sampling frequency). Typically, a full EEG recording consists of multiple such non-overlapping epochs. Consider a feature encoder $f_\theta : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^d$ that maps a multi-channel EEG epoch x_i into a d-dimensional latent representation $z_i \in \mathbb{R}^d$, where $z_i = f_\theta(x_i)$. In traditional supervised learning, f_θ is trained jointly with a classification or regression head $g_\phi(.)$ to predict the desired target y_i with the following learning objective.

$$\phi^*, \theta^* = \underset{\phi,\theta}{\text{argmin}} \sum_i \mathcal{L}(y_i, g_\phi(f_\theta(x_i)))$$

where $\mathcal{L}$ is a loss function. However, in representation learning (e.g., self-supervised learning), f_θ is first pretrained using a different objective to obtain a feature extractor f_{θ^*}, which produces a latent embedding for a given input. Classification or regression heads $g_\phi(.)$ can then be trained using the latent embeddings keeping f_{θ^*} fixed, where the learning objective takes the following form.

$$\phi^* = \underset{\phi}{\text{argmin}} \sum_i \mathcal{L}(y_i, g_\phi(f_{\theta^*}(x_i)))$$

3.2 Realistic data shifts in EEG data

In this study, we focus on modeling data shifts in EEG datasets caused by real-world variability. Consider the training (i.e., source) data domain $p_s(x)$ and the real-world data domain (i.e., target) $p_t(x)$, where x is input data. Covariate shifts are defined when the concept mapping $g_\phi \circ f_\theta : x \rightarrow y$ remains unchanged, resulting in the same conditional distribution of labels i.e., $p_s(y|x) = p_t(y|x)$. However, the marginal distributions of data are assumed to change i.e., $p_s(x) \neq p_t(x)$. In this study, we assume no access to external labels or additional data from the real-world. Instead, we synthetically construct the target domain $p_t(x)$ based on an expert understanding of the EEG signal acquisition process via a set of EEG data transformations $\mathcal{T}$. Therefore, within the scope of the current study, $p_t(x) = p_s(\mathcal{T}(x))$, and the proposed robustness assessment approach can be viewed as quantifying the differences between $p_s(x)$ and $p_s(\mathcal{T}(x))$ through a set of metrics $\mathcal{M}$.

The effects of real-world variability on EEG segments can be modeled using a data transformation function $t : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{M \times N}$, where $\tilde{x} = t(x)$ represents a shifted EEG sample. Real-world EEG data shifts may occur due to differences in populations, behavioral states (e.g., awake, asleep), the number of EEG sensors used, the EEG reference used, noise during acquisition, analog-to-digital conversion, hardware filter settings, and impedance resulting from electrodes. However, here we focus on data shifts induced by acquisition conditions and assume that other variables are controlled. In that context, we develop four EEG transformations: 1) band-pass filter t_{BP}, 2) quantization precision t_{QP}, 3) narrowband impedance noise t_{IN}, and 4) broadband noise t_{BN} (summarized in Table 1).

Table 1: Realistic EEG data shifts represented as transformations.

EEG Shift	Definition	Parameters	Description
Band-pass filter	$t_{\text{BP}} := \Psi(x, f_L, f_H)$	$f_L \in \{0.5, 1\}$ $f_H \in \{25, 30\}$	Ψ: band-pass filter function f_L, f_H: low and high cut-off filter frequencies
Quantization precision	$t_{\text{QP}} := k(x, D)$	$k : \mathbb{R} \to \mathbb{R}$ $D \in \{6, 8, 12\}$	k: restricts x to D decimal digits
Impedance noise	$t_{\text{IN}} := x + \epsilon$	$\sigma \in \{0.001, 0.01, 0.1\}$ $z_\sigma \sim \mathcal{N}(0, \sigma)$ $\epsilon = \Psi(z_\sigma, 0, 1)$	σ: noise strength z_σ: broadband noise ϵ: 0–1 Hz noise
Broadband noise	$t_{\text{BN}} := x + \epsilon$	$\sigma \in \{0.001, 0.01, 0.1\}$ $\epsilon \sim \mathcal{N}(0, \sigma)$	σ: noise strength ϵ: broadband noise

Band-pass filter: EEG device manufacturers commonly implement hardware-level band-pass filters in order to restrict the spectral content of the acquired signal. However, those filter settings may vary between different manufacturers and devices. Hence, we define a data shift t_{BP} to reflect the effects of different band-pass filter settings such as, 0.5–30 Hz, 1–30 Hz, and 1–25 Hz.

Quantization precision: In commercial EEG devices, resource constraints limit the precision of the recorded EEG data and the analog signal is digitized with variable number of bits. We implement this effect as a truncation of decimal digits to either 6, 8, or 12 points and denote it using t_{QP}.

Signal noise: The acquisition of EEG signals is highly susceptible to multiple sources of noise and artifacts: subject movement-related sources in the form of eye blinks, head movement, jaw clenching, etc., environmental sources such as AC power line interference, and instrumentation-related sources such as electrical impedance due to poor sensor contact with the scalp. We use t_{IN} to model electrical impedance noise due to poor sensor contact or choice of dry/wet electrodes that manifests as narrow-band low-frequency (0–1 Hz) random noise in the EEG signal (38) with different strengths ($\sigma \in \{0.001, 0.01, 0.1\}$). Next, we use t_{BN} to capture random physiological noise that may manifest as broadband (0–45 Hz) noise with different strengths ($\sigma \in \{0.001, 0.01, 0.1\}$).

3.3 Topological analysis of the latent space

Here we aim to quantify the distortion in latent space caused by input data transforms, i.e., differences between $z = f_\theta(x)$ and $\tilde{z} = f_\theta(\mathcal{T}(x))$. Some previous studies have utilized distance metrics such as Wasserstein's distance to quantify the differences (7). Here we employ a more comprehensive approach adapted from (39) and (24), where we estimate a manifold that contains point clouds comprising the latent representations $\{z_i\}$ and $\{\tilde{z}_i\}$ and conduct a geometric/topological analysis on this estimated manifold. We estimate the manifold based on Delaunay neighborhood graphs, which connect spatial neighbors in a way that captures local point density as well as global outliers (24).

Suppose we use $Z \in \mathbb{R}^d$ and $\tilde{Z} \in \mathbb{R}^d$ to denote the sets including all embeddings of raw and transformed input data, respectively, and define a set of vertices $\mathcal{V} = \{Z \cup \tilde{Z}\}$. Then, the Voronoi cell of $v \in \mathcal{V}$ is the set of points in $\mathbb{R}^d$ for which v is the closest among $\mathcal{V}$, i.e., $\text{Cell}(v) = \left\{x \in \mathbb{R}^d \mid \|x - v\| \leq \|x - v_i\| \, \forall v_i \in \mathcal{V}\right\}$. Then, we define edges $\mathcal{E}$ by connecting points v_i and v_j whose Voronoi cells intersect, i.e., $\mathcal{E} = \{(v_i, v_j) \in \mathcal{V} \times \mathcal{V} \mid \text{Cell}\,(v_i) \cap \text{Cell}\,(v_j) \neq \emptyset, v_i \neq v_j\}$. This set of vertices $\mathcal{V}$ and edges $\mathcal{E}$ form the neighborhoods in the Delaunay graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Intuitively, the points in $\mathcal{V}$ that share a Voronoi cell boundary are considered natural spatial neighbors irrespective of the distance between them. This property of Delaunay neighborhoods enables robust manifold estimation even when point density varies locally and outlier points exist. Notably, these spatial neighbors or edges can emerge in a way that connects points in Z to points in $\tilde{Z}$ (denoted by R and E respectively in (24)). Such edges can be considered 'heterogeneous', as opposed to 'homogeneous' edges which connect two points belonging to the same set. The degree of edge heterogeneity of $\mathcal{G}$ can be measured by a simple *quality* measure (39) $q(\mathcal{G})$ defined in Equation 1.

In Equation 1, $\left|\mathcal{G}^Z\right|_\mathcal{E}$ and $\left|\mathcal{G}^{\tilde{Z}}\right|_\mathcal{E}$ denote the number of homogeneous edges within $\mathcal{G}$ belonging to Z and $\tilde{Z}$, respectively, and $|\mathcal{G}|_\mathcal{E}$ denotes the total edges in $\mathcal{G}$. Therefore, if $\mathcal{G}$ exclusively contains

homogeneous edges, $q(\mathcal{G}) = 0$, and if all edges are heterogeneous, then $q(\mathcal{G}) = 1$. In this work, we use the degree of edge heterogeneity, $q(\mathcal{G})$, as a measure of latent space integrity between un-shifted and shifted samples under perturbations $\mathcal{T}$. A visual intuition of $q(\mathcal{G})$ is provided in appendix.

$$q(\mathcal{G}) = 1 - \frac{\left(\left|\mathcal{G}^{Z}\right|_{\mathcal{E}} + \left|\mathcal{G}^{\tilde{Z}}\right|_{\mathcal{E}}\right)}{\left|\mathcal{G}\right|_{\mathcal{E}}} \tag{1}$$

3.4 Quantifying model uncertainty using Monte Carlo dropout

In this work, we use a Monte Carlo dropout (MCD) strategy to examine per-sample variability in predictions, as a measure of model 'uncertainty' for a new sample (23). Specifically, we directly compare the changes in model uncertainty (measured using the standard deviation) under the aforementioned EEG-DS. Here we briefly summarize the theoretical background and our implementation.

In the Bayesian setting, the predictive distribution of a model f with parameters θ for an input x and output $\hat{y}$ is expressed as $p(\hat{y} \mid x, D) = \int p(\hat{y} \mid x, \theta)\, p(\theta \mid D) d\theta$, where D is training data. Because exact computation of this expression is intractable, we perform approximate inference using a variational distribution. A previous study showed that training neural networks with Bernoulli dropout variables is equivalent to approximate inference of the posterior of model parameters (23). With T sets of Bernoulli realizations of θ $\{\theta_t\}_{t=1}^{T}$, the Monte Carlo estimate of the mean prediction and its variance are computed as $\mathbb{E}[\hat{y}|x] \approx \frac{1}{T}\sum_{t=1}^{T} f(x, \theta_t)$ and $Var(\hat{y}|x) \approx \frac{1}{T}\sum_{t=1}^{T} \left(f(x, \theta_t) - \mathbb{E}[\hat{y}|x]\right)^2$. In our experiments, we used a dropout probability of $c = 0.5$ and repeated forward passes $T = 20$ times to estimate the mean and variance of predictions $\hat{y}$. We then compare the standard deviation and the agreement index of the predictions to make conclusions about predictive uncertainty of regression and classification tasks, respectively. The agreement index between multiple Monte Carlo dropouts for a threshold τ_t is defined as, $\phi = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{f(x, \theta_t) \geq \tau_t\}$.

4 Experimental Setup

Our overall experiment setup is illustrated in Figure 1. Using raw EEG signals and their shifted versions under transforms EEG-DS introduced in Section 3.2, we evaluate the robustness of EEG encoders in both in-sample and out-of-sample data. Through our experiments, we hope to answer the following questions: a) do latent space analysis and predictive uncertainty help anticipate model behavior in deployment settings?, b) what types of real-world EEG data shifts impact model robustness the most?, and c) do some encoders exhibit better robustness than others? In the following, we briefly describe the components of our evaluation setup: the EEG datasets, preprocessing steps, EEG feature encoders, learning tasks, data splits, and evaluation metrics.

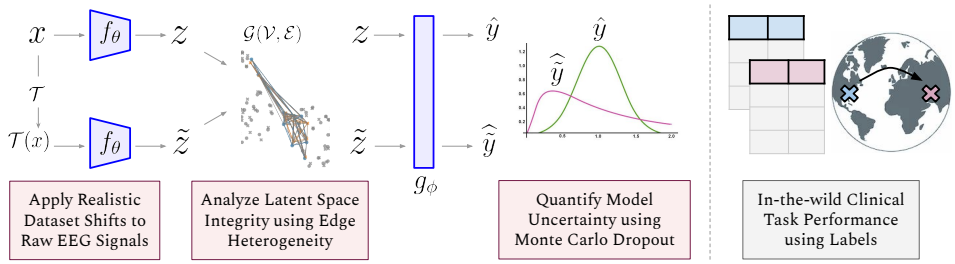

Figure 1: **The overall experiment setup.** First, we trained several EEG feature encoders (f_θ) using unmodified training data in fully-supervised, self-supervised, and traditional PSD feature-based settings. Second, we generated latent representations for the unmodified test data and the modified test data (using data shifts $\mathcal{T}$), and compared the topological similarities in the latent space (z and $\tilde{z}$). Next, we fine-tuned the feature encoders on classification and regression tasks (g_ϕ) and generated predictions y and $\tilde{y}$ for both z and $\tilde{z}$. We then quantified the uncertainty in predictions using a MCD strategy. Finally, we evaluated encoder performances in the in-sample test data (for modified and unmodified versions) and in the out-of-sample data.

EEG dataset & preprocessing: We utilize EEG data from two large-scale databases, TUH-EEG database from Temple University Hospital (13) and NMT database from Pak-Emirates Military Hospital (25). For our experiments, we utilize 2,993 EEGs from TUH-EEG (a subset known as

TUAB) and 1,369 EEGs from the NMT database, both of which contain normal and abnormal labels. All EEGs were recorded according to the standard 10-20 montage (40). We extracted age labels from text reports accompanying the EEG recordings using a previously published tool (41). Next, we preprocessed the EEGs as follows: a) we ordered the EEG channels as 'Fp1', 'Fp2', 'F3', 'F4', 'C3', 'C4', 'P3', 'P4', 'O1', 'O2', 'F7', 'F8', 'T7', 'T8', 'P7', 'P8', 'Fz', 'Cz', and 'Pz', b) resampled the recordings to 128Hz, c) applied a bandpass filter between 0.5Hz and 45.0Hz, d) divided the recordings into contiguous non-overlapping epochs of 10-seconds each, e) identified and removed bad epochs when the total power in their 'Cz' channel exceeded 2 standard deviations as calculated from statistics of each recording, f) clipped amplitude values to not exceed $\pm$ 800μV, and finally g) we normalized the recording per-channel using the identified valid epochs. Note that whenever transforms $\mathcal{T}$ were utilized, they were applied directly on the raw data before preprocessing.

EEG encoders: We train three diverse types of raw EEG encoders: fully-supervised, self-supervised, and traditional feature-based. The fully supervised encoders (FSE) were based on the widely used 'ShallowNet' architecture (14) with additional classification and regression heads described below. The self-supervised encoder (SSE) also utilized the 'ShallowNet' architecture with pre-training performed using a proxy measure of the patient's behavioral state (17). Finally, we utilized a power-spectral-density-based encoder (PSDE) with EEG spectral power calculated within 7 frequency bands: δ (2-4Hz), θ (4-8Hz), low α (8-10Hz), high α (10-13Hz), low β (13-16Hz), high β (16-25Hz), and γ (25-40Hz) for each of the 19 channels. Both FSE and SSE took raw EEG traces as inputs and produced 128-dimensional embeddings, while PSDE generated $19\times7=133$ features.

We utilized additional classification and regression heads to develop task-specific predictors using the above encoders. Note that we trained FSE entirely using the task labels, whereas we fine-tuned SSE by training the classification/regression layer while keeping the encoder network frozen. In addition, we employed dropout with $p = 0.5$ in all heads and used a binary cross-entropy loss and a smooth variant of the mean absolute error loss for training the classification and regression heads, respectively. We trained the models up to a maximum of 500 epochs and while monitoring convergence using a validation set of 373 EEGs. We used either Adam or SGD with a cyclic learning rate scheduler (42) optimizer. The full set of hyper-parameter values are provided in the appendix and are also documented within code released as a URL therein.

Classification and regression tasks: We utilized meaningful classification and regression tasks to evaluate the robustness of above encoders. EEG grade classification as 'normal' or 'abnormal' is a common task performed by clinical experts via visual review. Performing such classification using ML approaches in an automated fashion is of great interest since it has the potential to reduce the workload of clinical experts. Next, the potential to predict an individual's age using EEGs, a regression task commonly known as brain age prediction, is useful because it can indicate deviations from healthy brain aging and therefore help with early diagnosis of neurodegenerative diseases. Hence, we utilized two tasks: a binary classification task classifying EEG grades as 'normal' or 'abnormal' and a regression task predicting brain age from EEGs, to conduct our experiments.

Training, validation, and testing: We split TUAB into fixed training (1,490 EEGs; 258,631 epochs), validation (373 EEGs; 68,493 epochs), and held-out test (466 EEGs; 82,331 epochs) sets. Furthermore, we performed all model training on TUAB at the epoch level without any shifts. Since both EEG grade and age are recording-level measures, during evaluation, we aggregate epoch-level predictions by averaging the predictions across all epochs in a recording (17; 20).

In-sample and out-of-sample evaluations: We evaluated the robustness of EEG encoders in the in-sample setting by comparing performances between the modified and unmodified TUAB test set. We evaluated the out-of-sample performance using the external NMT dataset, while restricting the evaluations to data from adults (18+) to avoid any physiological variability unseen during training.

Evaluation metrics: We perform all evaluations on the held-out test set. The key metrics utilized to perform the proposed encoder assessments are described in Section 3. Based on the graph construction outlined in Section 3.3, we interpret the measure of edge heterogeneity obtained using Equation 1 as a score of integrity of the latent space under perturbation. For each encoder, we compare the representations of raw EEG samples against their shifted versions in the latent space. Next, we use predictive uncertainty to analyze per-sample variability of encoders under dataset shifts. Specifically, we analyze the median agreement index (ϕ) and standard deviation in model predictions obtained via MCD as measures of uncertainty. Finally, we quantify task performance using the ground truth labels. To analyze task performance, we use the area under the receiver operating characteristic curve (AUC)

and the mean absolute error (MAE) metrics for EEG grade and brain age tasks, respectively. Note that we performed the analyses of the latent space using embeddings of all the epochs in the test set and analyzed task performance and uncertainty using aggregated EEG-recording level metrics.

Software & hardware: The preprocessing, baseline features, and model training were implemented using a combination of the following Python libraries: 1) MNE-python (43), 3) PyTorch (44), and 3) NumPy (45). Delaunay graph construction and analysis were performed using code released by (24). The experiments were performed using 2 Nvidia RTX 3090 GPUs.

5 Results

Our results are summarized in Figures 2 and 3, and Table 2. We performed latent space analysis and uncertainty quantification without using the ground truth labels and evaluated whether the changes in those measures resulting from test-time data-shifts indicate performance degradation in the in-sample and out-of-sample test sets. In the following, we use the following acronyms to refer to the data shifts and encoders: BP - bandpass filter, QP - quantization precision, IN - impedance noise, BN - broadband noise, SSE - self-supervised encoder, FSE-grade - fully-supervised encoder fine-tuned on grade classification, FSE-age - fully-supervised encoder fine-tuned on age prediction, PSDE - traditional power-spectral density-based encoder.

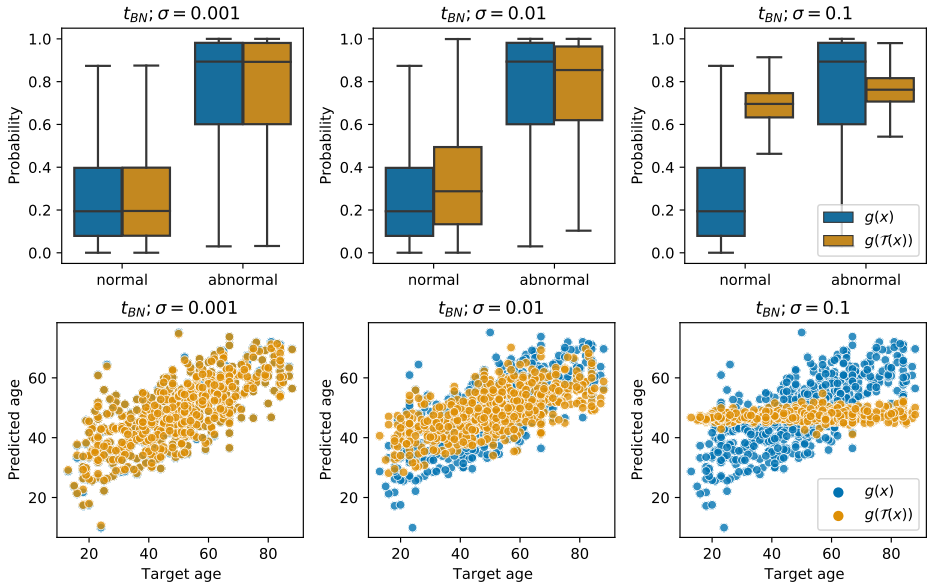

Figure 2: The impact of broadband noise on the predictions of FSE encoder in EEG grade and brain age tasks.

Impact of data-shifts on predictions: Figure 2 shows an example of how model predictions are affected when the proposed transformations are applied to the input data. In the top row, we show a comparison of the predicted probability of EEG grade by the FSE encoder on held-out data under additive broadband noise t_{BN} of varying strength σ. The probabilities are grouped along the x-axis by expert labels. We observe that the predicted probability distributions for the two ground truth classes overlap more under stronger shifts. In the bottom row, we show scatter plots of brain age predicted by FSE against the ground truth under similar comparison settings as above. With stronger shifts, we observe a collapse in the range of FSE predictions.

Evaluation of the latent space: Figure 3 shows the latent space integrity scores (↑) in the in-sample (lines) and out-of-sample ('×'s) settings for multiple EEG encoders. The same results are provided in a tabular form in the supplement including standard deviations. First, we established an empirical baseline for ideal integrity scores by comparing un-modified data with itself for each encoder (labeled as 'No shift' in Figure 3). Then we compared the in-sample integrity scores between the encoders by increasing the strength of each data shift (left → right). We find that all proposed data shifts affect the integrity scores, albeit to varying degrees, with BN and QP being the most and least effective, respectively. Interestingly, we also find that PSDE and SSE encoders produce better overall integrity in the latent space compared to others. We then computed these scores between unmodified in-sample

and out-of-sample test sets for 100 randomly chosen EEGs. We find that there are notable differences between the two test sets, and that the representations generated by SSE preserve integrity the most.

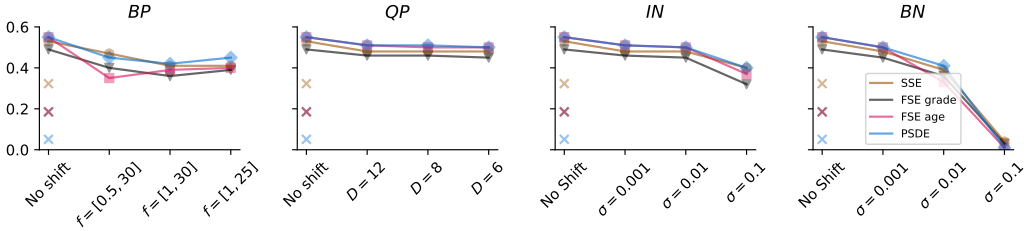

Figure 3: Integrity measures ($q(\mathcal{G})$) in the in-sample (lines) and out-of-sample ('$\times$'s) settings for different encoders. Measures were computed at different strengths of shifts in the in-sample setting, and between unmodified TUH and NMT tests for the out-of-sample setting.

Task performance & uncertainty: Table 2 lists the task performance measures and the uncertainty measures for EEG grade classification and brain age regression tasks evaluated on the in-sample and out-of-sample test sets under different data shifts. For additional context, the same metrics computed on the train and validation sets are provided as supplement. The task performance metrics reported are the means of the recording-level AUCs and MAEs computed based on the 20 MCD repeats (standard deviations are provided in the supplement). Regardless of shift, we find that FSE outperforms both SSE and PSDE in both tasks significantly. However, SSE and PSDE exhibit consistent performance under perturbations in the age prediction task although their absolute performance is worse compared to FSE. Data shifts IN and BN affect all encoders in EEG grade classification, and BP, IN, and BN affect FSE encoder in age prediction. As measures of uncertainty, we report the average agreement index in EEG grade prediction and the standard deviation of the predicted age values across 20 MCD repeats. We find that both SSE and PSDE exhibit overconfidence in model predictions despite achieving worse task performance than FSE in both tasks. Next, all EEG-DS cause only marginal changes on uncertainty measures and only the strongest BN ($\sigma = 0.1$) causes a noticeable change. The out-of-sample evaluation shows a universal degradation in both task performance and uncertainty measures across all encoders compared to the baseline.

Table 2: Task performance and uncertainty measures for classification (EEG Grade) and regression (Age) tasks in in-sample and out-of-sample data under different shifts. Abbreviations: AUC - area under curve (↑), MAE - mean absolute error (↓), ϕ - agreement index (↑), and SD - standard deviation in predicted values (↓). Up/down arrows indicate direction of desirable trends. Bold highlights indicate the scenarios when the proposed EEG shifts significantly degraded the performance.

EEG Shifts ($\mathcal{T}$)	**EEG Grade (AUC)**			**Age (MAE)**			**EEG Grade (ϕ)**			**Age (SD)**		
	SSE	FSE	PSDE	SSE	FSE	PSDE	SSE	FSE	PSDE	SSE	FSE	PSDE
No shift (baseline)	0.77	0.92	0.77	13.74	9.47	15.56	0.99	0.99	0.96	0.10	0.50	0.32
No shift - NMT	0.72	0.72	0.64	17.46	12.41	17.74	0.97	0.97	0.85	0.14	0.73	0.52
t_{BP} ($f = [0.5, 30]$)	0.77	0.92	0.77	13.74	11.24	15.52	0.99	1.00	0.96	0.10	0.47	0.32
t_{BP} ($f = [1, 30]$)	0.78	0.92	0.78	13.70	11.14	15.44	0.99	1.00	0.95	0.10	0.46	0.33
t_{BP} ($f = [1, 25]$)	0.78	0.92	0.77	13.70	11.10	15.47	0.99	0.99	0.95	0.10	0.46	0.33
t_{QP} ($D = 12$)	0.77	0.92	0.77	13.74	9.47	15.56	0.99	0.99	0.96	0.10	0.50	0.32
t_{QP} ($D = 8$)	0.77	0.92	0.77	13.74	9.47	15.56	0.99	0.99	0.96	0.10	0.50	0.32
t_{QP} ($D = 6$)	0.77	0.92	0.77	13.75	9.51	15.56	0.99	0.99	0.96	0.10	0.49	0.32
t_{IN} ($\sigma = 0.001$)	0.77	0.92	0.77	13.74	9.47	15.56	0.99	0.99	0.96	0.10	0.50	0.32
t_{IN} ($\sigma = 0.01$)	0.77	0.92	0.76	13.75	9.49	15.58	0.99	0.99	0.96	0.10	0.49	0.32
t_{IN} ($\sigma = 0.1$)	**0.71**	**0.91**	**0.67**	**13.77**	**11.17**	**16.07**	**0.99**	**0.99**	**0.98**	**0.10**	**0.44**	**0.27**
t_{BN} ($\sigma = 0.001$)	0.77	0.92	0.77	13.74	9.50	15.56	0.99	0.99	0.96	0.10	0.49	0.32
t_{BN} ($\sigma = 0.01$)	0.78	0.91	0.78	13.84	11.06	15.41	0.99	0.99	0.95	0.08	0.47	0.31
t_{BN} ($\sigma = 0.1$)	**0.72**	**0.86**	**0.71**	**14.08**	**14.05**	**14.66**	**0.83**	**0.94**	**0.83**	**0.11**	**0.76**	**0.35**

6 Discussion

The aim of this study was to develop model diagnostic measures for EEG feature encoders that help anticipate deployment failures during model development itself. We modeled real-world variability in EEG signals by designing four data shifts, i.e., EEG-DS (Table 1), that capture the effects of band-pass filter settings, analog-to-digital quantization precision, and manifestations of noise. Then, we trained diverse types of EEG encoders (self-supervised, fully-supervised, and conventional) on un-shifted

data for two clinically important tasks (Section 4), and assessed their robustness under these shifts during model development. Notably, we analyzed the encoders' latent space (Figure 3) and predictive uncertainty (Table 2) in a label-agnostic fashion, in addition to analyzing task performance (Table 2). Finally, an evaluation using an external dataset indicated that the performance drop observed in the out-of-sample setting is proportional to what we observed in the above experiments.

The value of latent space analysis: If we consider task performance as gold-standard metric in deployment settings, the latent space analysis allows us to see 'early' signs that predict degraded task performance (Figure 3 and Table 2). The results of aggressive shifts (e.g., σ=0.1) in the in-sample experiments strongly correlated with the out-of-sample performance, highlighting the potential use of the latent space analysis. Intuitively, a strong distortion in the latent space is likely to push samples from one side of the decision boundary to the other, leading to worse performance.

The value of uncertainty quantification: Our uncertainty studies identify a major pitfall that is possible to occur during deployment. We found that both SSE and PSDE encoders were highly confident in their predictions although their task performances were much worse compared to FSE. This might indicate that the SSE and PSDE encoders were not sufficiently trained to achieve desirable task performances (i.e., classifier bias) and hence showed less variability in predictions. On the other hand, the uncertainty measures of the FSE encoder provided meaningful interpretations although the changes in uncertainty under the proposed data shifts were marginal. We believe that analyzing measures of uncertainty in conjunction with task performance can help assess real-world robustness.

Data shifts: Trends in Figure 3 indicate that a (1-25Hz) band-pass filter and strong additive broadband or narrowband noise ($\sigma = 0.1$) impact latent space the most. Table 2 shows the same trend in task performance, although the effects are more pronounced in age prediction compared to EEG grade classification. From an EEG perspective, sensitivity to band-pass filtering and noise is expected since they affect the spectral content of the signal, which is important for the tasks of EEG grade classification and brain age regression. We found that the data shifts did not impact uncertainty measures all that much except for the strongest broadband noise ($\sigma = 0.1$). While we do not consider all sources of EEG shift, the fact that the proposed instrumentation-related shifts can anticipate out-of-sample performance suggests that they represent a good fraction of real-world diversity, and are indeed valuable in anticipating deployment failures.

EEG feature encoders: A general trend seen in Table 2 across all shifts is that FSE provides better absolute performance and that SSE and PSDE encoder performances are more stable. This apparent trade-off between absolute performance and robustness can be explained by considering the effective model sizes (i.e., number of learnable parameters) in each setting. Recall from Section 4 that only the last linear layer $g_\phi(.)$ is trainable for SSE and PSDE while the entire network $g_\phi(f_\theta(.))$ is trained in case of FSE. Due to disproportionately higher number of parameters, FSE can apriori be expected to perform better than SSE and FSE. However, this also implies that dropout in FSE would inactivate a disproportionately higher number of units, leading to less stable performance.

Limitations and future directions: Besides instrumentation-related variability, data shifts originating from physiological changes in brain activity (e.g., sleep) may also play a significant role in real-world robustness. In future work, we will implement additional data shifts representing the physiological variability in brain activity and investigate whether adversarial training and training with data shifts could improve real-world robustness.

7 Conclusion

In this paper, we developed an approach to assess real-world robustness of EEG-ML models during model development itself. Our approach included implementation of realistic data shifts and evaluation of model robustness using a latent space topological analysis and quantification of predictive uncertainty. We developed four data shifts for EEG signals reflecting potential instrumentation-related variability observable in the real-world. We then performed experiments to evaluate differences in latent space integrity, predictive uncertainty, and task performance, under the above data shifts and correlated the results with out-of-sample performance. Within our experimental setting, including three EEG feature encoders, two downstream tasks, and two large-scale EEG datasets, our results empirically demonstrate that the proposed approach can help anticipate real-world performance during model development itself. In future work, we will consider other plausible data shifts, adversarial training, and training models explicitly on realistic data shifts.

References

[1] E. A. AlBadawy, A. Saha, and M. A. Mazurowski, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," *Medical physics*, vol. 45, no. 3, pp. 1150–1158, 2018.

[2] S. K. Lynch, A. Shah, J. C. Folk, X. Wu, and M. D. Abramoff, "Catastrophic failure in image-based convolutional neural network algorithms for detecting diabetic retinopathy," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 8, pp. 3776–3776, 2017.

[3] E. Strickland, "Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care," *IEEE Spectrum*, vol. 56, no. 4, pp. 24–31, 2019.

[4] D. Wang, L. Wang, Z. Zhang, D. Wang, H. Zhu, Y. Gao, X. Fan, and F. Tian, ""brilliant ai doctor" in rural clinics: Challenges in ai-powered clinical decision support system deployment," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–18.

[5] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[6] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger *et al.*, "The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, 2020.

[7] L. Betthauser, U. Chajewska, M. Diesendruck, and R. Pesala, "Discovering distribution shifts using latent space representations," *arXiv preprint arXiv:2202.02339*, 2022.

[8] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[9] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, "Correcting sample selection bias by unlabeled data," *Advances in neural information processing systems*, vol. 19, 2006.

[10] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 539–11 551, 2020.

[11] O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, T. Cemgil *et al.*, "A fine-grained analysis on distribution shift," *arXiv preprint arXiv:2110.11328*, 2021.

[12] S. Saria and A. Subbaswamy, "Tutorial: safe and reliable machine learning," *arXiv preprint arXiv:1904.07204*, 2019.

[13] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in neuroscience*, vol. 10, p. 196, 2016.

[14] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] S. Roy, I. Kiral-Kornek, and S. Harrer, "Deep learning enabled automatic abnormal eeg identification," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 2756–2759.

[16] M. Alhussein, G. Muhammad, and M. S. Hossain, "Eeg pathology detection based on deep learning," *IEEE Access*, vol. 7, pp. 27 781–27 788, 2019.

[17] N. Wagh, J. Wei, S. Rawal, B. Berry, L. Barnard, B. Brinkmann, G. Worrell, D. Jones, and Y. Varatharajah, "Domain-guided self-supervision of eeg data improves downstream classification performance and generalizability," in *Machine Learning for Health*. PMLR, 2021, pp. 130–142.

[18] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.

[19] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046020, 2021.

[20] N. Wagh and Y. Varatharajah, "Eeg-gcnn: Augmenting electroencephalogram-based neurological disease diagnosis using a domain-guided graph convolutional neural network," in *Machine Learning for Health*. PMLR, 2020, pp. 367–378.

[21] Y. Varatharajah, B. Joseph, B. Brinkmann, M. Morita-Sherman, Z. Fitzgerald, D. Vegh, D. Nair, R. Burgess, F. Cendes, L. Jehi *et al.*, "Quantitative analysis of visually reviewed normal scalp eeg predicts seizure freedom following anterior temporal lobectomy," *Epilepsia*.

[22] P. Thangavel, J. Thomas, W. Y. Peh, J. Jing, R. Yuvaraj, S. S. Cash, R. Chaudhari, S. Karia, R. Rathakrishnan, V. Saini *et al.*, "Time–frequency decomposition of scalp electroencephalograms improves deep learning-based epilepsy diagnosis," *International journal of neural systems*, vol. 31, no. 08, p. 2150032, 2021.

[23] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[24] P. Poklukar, V. Polianskii, A. Varava, F. Pokorny, and D. Kragic, "Delaunay component analysis for evaluation of data representations," *arXiv preprint arXiv:2202.06866*, 2022.

[25] H. A. Khan, R. U. Ain, A. M. Kamboh, H. T. Butt, S. Shafait, W. Alamgir, D. Stricker, and F. Shafait, "The nmt scalp eeg dataset: An open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling," *Frontiers in Neuroscience*, vol. 15, 2021.

[26] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi, "An empirical framework for domain generalization in clinical settings," in *Proceedings of the Conference on Health, Inference, and Learning*, 2021, pp. 279–290.

[27] H. Raza, G. Prasad, and Y. Li, "Adaptive learning with covariate shift-detection for non-stationary environments," in *2014 14th UK Workshop on Computational Intelligence (UKCI)*. IEEE, 2014, pp. 1–8.

[28] ——, "Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments," *Pattern Recognition*, vol. 48, no. 3, pp. 659–669, 2015.

[29] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[30] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.

[31] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7728–7738.

[32] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[33] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2712–2721.

[34] P. Nejedly, V. Kremen, V. Sladky, M. Nasseri, H. Guragain, P. Klimes, J. Cimbalnik, Y. Varatharajah, B. H. Brinkmann, and G. A. Worrell, "Deep-learning for seizure forecasting in canines with epilepsy," *Journal of neural engineering*, vol. 16, no. 3, p. 036031, 2019.

[35] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *Journal of Neuroscience Methods*, vol. 346, p. 108885, 2020.

[36] F. Wang, S.-h. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data augmentation for eeg-based emotion recognition with deep convolutional neural networks," in *International conference on multimedia modeling*. Springer, 2018, pp. 82–93.

[37] Y. Varatharajah, M. J. Chong, K. Saboo, B. Berry, B. Brinkmann, G. Worrell, and R. Iyer, "Eeg-graph: a factor-graph-based model for capturing spatial, temporal, and observational relationships in electroencephalograms," *Advances in neural information processing systems*, vol. 30, 2017.

[38] E. S. Kappenman and S. J. Luck, "The effects of electrode impedance on data quality and statistical significance in erp recordings," *Psychophysiology*, vol. 47, no. 5, pp. 888–904, 2010.

[39] P. Poklukar, A. Varava, and D. Kragic, "Geomca: Geometric evaluation of data representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8588–8598.

[40] G. E. Chatrian, E. Lettich, and P. L. Nelson, "Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities," *American Journal of EEG technology*, vol. 25, no. 2, pp. 83–92, 1985.

[41] S. Rawal and Y. Varatharajah, "Score-it: A machine learning-based tool for automatic standardization of eeg reports," *Machine Learning for Healthcare (Clinical Abstract)*, 2021.

[42] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[43] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen *et al.*, "Meg and eeg data analysis with mne-python," *Frontiers in neuroscience*, vol. 7, p. 267, 2013.

[44] 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," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[45] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The proposed data shifts for EEG signals are defined in Table 1, and results of latent space analysis and predictive uncertainty are summarized in Figure 3 and Table 2, respectively. We highlight that this study does not propose an approach to improve model robustness. See Section 6 for summary of significance and scope of study.
   (b) Did you describe the limitations of your work? [Yes] See Section 6 for discussion on limitations of study and caveats in results.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This study introduces realistic data shifts and approaches that can anticipate model failures and issues with model robustness before the model is deployed. Therefore, proposed approaches are strongly aimed at making a positive societal impact by attempting to detect costly real-world failures before they happen.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] This study uses publicly available EEG data and corresponding text reports (13) that have been de-identified before release. The dataset contains patients with a wide variety of neurological disorders and diverse age and gender groups. The training data used in our models reflects this diversity. While the focus of our study is evaluations of model robustness to EEG instrumentation-related shifts before deployment, we acknowledge that additional assessments are needed for clinical and protected sub-groups of interest.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have released a sanitized version of all code and instructions (i.e. anonymized for file paths and other sources of identifiable information about authors) needed to reproduce results in Figure 3, Figure 2, and Table 2 as a URL in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Training details including data splits (train, validation, test) and major hyperparameters are summarized in Section 4. The complete set of hyperparameters are enumerated in the appendix, and are also documented in code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Variability in model performance is a central theme of this study. As such, the variability in model performance (due to Monte Carlo dropout) on the two EEG tasks is summarized in Table 2. The supplement contains additional variability results that were omitted from the main text for sake of brevity.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4 for acknowledgement of software and internal cluster GPU hardware used for training the models.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4 for acknowledgement of authors whose code we built upon.
   (b) Did you mention the license of the assets? [No] The code we have used is either a publicly released Python library or a GitHub repository with their respective open-source licenses.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include links to anonymized code and instructions for replicating results as a URL in the appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The public datasets used in this study were de-identified before release.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

8 Appendix

8.1 Visualization of latent space edge heterogeneity under dataset shift

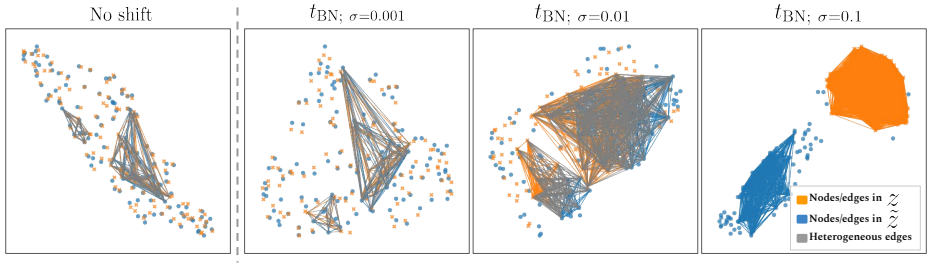

Figure 4: Visualization of changes in latent space integrity between unmodified (z; orange) and broadband noise shifted data ($\tilde{z}$; blue) as measured by the proportion of heterogeneous edges (grey) in the Delaunay neighborhood graph (Section 3.3). Latent representations were generated for a held-out recording using the fully-supervised encoder (FSE) and reduced to two dimensions for visualization. Note that as noise strength increases from left to right, there is a decrease in proportion of heterogeneous edges. This is interpreted as decreasing latent space integrity, and corresponds to lower $q(\mathcal{G})$ scores in Figure 3 of the main text. Figure best viewed in color.

Note that while the number of heterogeneous (grey) edges apparently increases, their proportion relative to total edges, in fact, decreases. This decrease in proportion lowers the proposed integrity measure. Out of the shifts shown in Figure 3 of the main text, broadband noise (BN; sigma=0.1) is aggressive enough to yield only homogeneous edges, and the measure does indeed go to 0. In the 'No shift (baseline)' scenario, the number of homogeneous and heterogeneous edges is roughly equal, leading to an integrity score of approximately 0.5.

8.2 List of training-related hyperparameters

Table 3: Hyperparameters related to neural network model training

Category	**Hyperparameter**	**Value**
Smooth L1 regression loss	Beta threshold	1.0
Adam optimizer &	Learning rate	0.001
Multi-step LR scheduler	Betas	(0.9, 0.999)
	Weight decay	1e-5
	Milestones	[30, 80, 150, 200, 250, 300, 350, 400, 450]
	Decay factor	0.5
SGD optimizer &	Base learning rate	1e-5
Cyclic LR scheduler	Maximum learning rate	0.01
	Momentum	0.9
	Weight decay	1e-5
	Step size up	2000
	Gamma	0.5
Layer initialization	Weight - Xavier uniform gain	1.0
(linear, conv1d, convtranspose1d)	Bias - constant fill	0.01

8.3 Link to code

Code and data required to reproduce results in Figure 2, Figure 3, and Table 2 is provided at `https://figshare.com/s/2a8eb98b47f050b99b16`.