UNSUPERVISED DOMAIN SHIFT DETECTION WITH INTERPRETABLE SUBSPACE ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We developed a tool for detecting domain shifts, namely subtle differences in the probability distributions of datasets. We identify these shifts using an algorithm designed to detect localised density differences in high-dimensional feature spaces. If an anomaly is present, we then identify the subset of features in which the anomaly is most pronounced. This allows to trace the domain shift to a small set of features, making the shift interpretable. Moreover, we provide a protocol for compensating domain shifts by extracting, from two unlabelled datasets, subsets of samples in which the anomaly is not present. We validate the framework on controlled 20-dimensional benchmarks with known ground truth, recovering both broad and localized shifts together with their supporting feature subspaces. We then apply it to healthy electrocardiogram (ECG) recordings represented by 782 features. In age- and sex-matched cohort comparisons differing in measurement-device composition, the method detects device-induced shifts, extracts representative subsets enriched in the imbalanced device components, and identifies ECG features associated with the anomaly.

1 INTRODUCTION

Comparing two collections of unlabeled samples is a fundamental task in machine learning. It arises in domain adaptation (Quiñonero-Candela et al., 2009; Sugiyama & Kawanabe, 2012), dataset auditing (Rabanser et al., 2019), covariate-shift correction (Gretton et al., 2009; Sugiyama et al., 2008), and the validation of learned representations (Lopez-Paz & Oquab, 2017). The central statistical question is whether two finite samples are consistent with a common distribution and, when they are not, how the discrepancy is organized across samples and features. Since only the feature samples are observed, those discrepancies are *covariate shifts* in the terminology of Quiñonero-Candela et al. (2009); discrepancies due to the relationship between features and labels are outside the scope of this work. In machine-learning pipelines, undetected distributional discrepancies between training and deployment data, or between two reference cohorts, can distort learned representations, invalidate benchmark comparisons, and introduce systematic biases into predictions (Quiñonero-Candela et al., 2009; Rabanser et al., 2019). Identifying such discrepancies before fitting or evaluating models is therefore an important component of reliable data-driven analysis.

Global two-sample tests, such as maximum mean discrepancy (Gretton et al., 2012) and energy-distance tests (Székely & Rizzo, 2013), provide evidence of distributional change at the dataset level. Classifier two-sample tests (C2ST) (Lopez-Paz & Oquab, 2017) cast the same problem as binary discrimination and can capture complex differences through flexible decision functions. These methods are powerful tools for testing or quantifying whether two datasets differ, but they do not identify sets of samples associated with the shift, nor the feature subspaces that support the discrepancy. Distributional discrepancies can involve a minority subpopulation, a restricted region of the sample space, or a feature subset in which the contrast is most pronounced, while the bulk of both datasets is comparable (Koh et al., 2021; Yang et al., 2023; Rabanser et al., 2019). In such settings, dataset-level tests may still detect a discrepancy, but they provide limited information about its location or feature-level organization; the question they answer is whether the datasets differ, whereas here we ask where, and in which features. Nearest-neighbor two-sample tests (Schilling, 1986; Friedman & Rafsky, 1979) exploit local geometry and can provide local evidence of distributional change, but also those methods do not produce equalized subsets and the features which determine

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the discrepancy. Simultaneously localizing the responsible samples, attributing the discrepancy to an active feature support, and constructing equalized subsets therefore requires a more targeted pipeline.

We here introduce a framework for *unsupervised detection, localization, and feature attribution* of distributional discrepancies between two unlabeled datasets. The framework proceeds in three stages. First, we use a recently introduced local two-sample score (Springer et al., 2025) as a pointwise discrepancy backbone, assigning to each sample an anomaly score for local over- or under-representation relative to the other dataset. Second, we introduce a bidirectional equalization procedure that identifies statistically significant excess-mass samples and constructs pruned subsets with undetectable discrepancy. Third, we introduce Neighbor-Enriched Subspace Localization, a procedure that identifies the active feature support by maximizing the anomaly score. This last ingredient is, to our knowledge, fully novel and, as we will show, is essential to boost the detection power of the algorithm.

We validate the framework on controlled 20-dimensional benchmarks with known ground truth, where it reliably recovers *localized shifts* affecting structured minority populations that pointwise scoring baselines fail to isolate. We then apply it to healthy electrocardiogram (ECG) recordings from PTB-XL+ (Wagner et al., 2020; Strodthoff et al., 2023), represented by 782 clinically interpretable features, in age- and sex-matched comparisons constructed to differ primarily in measurement-device composition. The framework detects device-associated distributional shifts, extracts representative subsets enriched in the device components driving the mismatch, and identifies coherent feature supports across controlled comparisons, illustrating how it can be used to preprocess dataset composition before downstream modelling.

The main contributions of this paper are:

- **A bidirectional iterative equalization algorithm** that prunes excess-mass points from the two datasets (Section 2.2).
- A **Neighbor-Enriched Subspace Localization (NESL)** procedure that attributes the pruned set to the features which are primarily responsible for the domain shift (Section 2.3).
- **A complete domain shift detection and attribution pipeline**, requiring no class labels or supervised training on the datasets under comparison (Section 2).
- **Controlled synthetic experiments** with known ground truth demonstrating recovery of global and localized shifts with strongly improved performance with respect to similar methods (Section 3.1).
- **A real-world ECG application** demonstrating detection of device-induced acquisition bias in age- and sex-matched healthy cohorts, with coherent feature attribution across controlled device-composition contrasts (Section 3.2).

2 METHOD

2.1 BACKBONE: EAGLEEYE LOCAL TWO-SAMPLE SCORING

We adopt EagleEye (Springer et al., 2025) as a deterministic, unsupervised, training-free two-sample procedure that localizes where two datasets differ in density. Given two samples $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_{n_\mathcal{X}}\}$ and $\mathcal{Y} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_{n_\mathcal{Y}}\}$, EagleEye aims to identify *localized density anomalies*: regions of feature space where $\mathcal{Y}$ is significantly over- or under-represented relative to $\mathcal{X}$. We present the procedure for identifying anomalous points in $\mathcal{Y}$, but the procedure is run in both directions to capture localized excesses and deficits in each sample. Let $\mathcal{U} = \mathcal{X} \cup \mathcal{Y}$ be the pooled set. For each test point $\mathcal{Y}_i \in \mathcal{Y}$, EagleEye computes its nearest neighbours in $\mathcal{U}$ up to a maximal neighbourhood rank K_M and encodes neighbour memberships as a binary sequence

$$b_i = (b_i^1, \ldots, b_i^{K_M}), \qquad b_i^k = \begin{cases} 0, & \text{if the } k\text{-th neighbour of } \mathcal{Y}_i \text{ lies in } \mathcal{X}, \\ 1, & \text{if the } k\text{-th neighbour of } \mathcal{Y}_i \text{ lies in } \mathcal{Y}. \end{cases}$$

Under the null hypothesis that the two samples are *locally exchangeable*, neighbour labels follow the global mixture proportions, so the baseline probability that a neighbour belongs to $\mathcal{Y}$ is $\hat{p} = \frac{n_\mathcal{Y}}{n_\mathcal{X}+n_\mathcal{Y}}$. For $K \in \{1, \ldots, K_M\}$, define the cumulative number of test-sample neighbours among the first K

neighbours as

$$B(i,K) = \sum_{k=1}^{K} b_i^k.$$

In the regime $K \ll n_{\mathcal{X}} + n_{\mathcal{Y}}$, we approximate $B(i,K)$ under the null as $\mathrm{Binomial}(K,\hat{p})$ and define an evidence score

$$\Upsilon(i,K) := -\log(\Pr[B(i,K) \geq B_{\mathrm{obs}}(i,K)]) .$$

The pointwise anomaly score is then obtained by maximizing across spatial scales,

$$\Upsilon_i := \max_{1 \leq K \leq K_M} \Upsilon(i,K), \tag{1}$$

yielding sensitivity to discrepancies at unknown spatial scales.

We flag putative anomalous points in the test sample as

$$\mathcal{Y}^+ := \{\, \mathcal{Y}_i \in \mathcal{Y} \mid \Upsilon_i \geq \Upsilon_+^* \,\},$$

where the critical threshold Υ_+^* is calibrated so that the *per-point* null exceedance probability equals a user-chosen level p_{ext} (default 10^{-5}). Because Υ_i is a maximum over $K \leq K_M$, its null distribution is calibrated by Monte Carlo using Bernoulli sequences of length K_M. The flagged set $\mathcal{Y}^+$ can be interpreted as an anomaly map comprising an anomalous core together with a surrounding "halo" of points whose neighborhoods intersect a true discrepancy region.

2.2 Tail-Based Bidirectional Equalization

The next step of the procedure aims at identifying the statistically significant excess mass relative to the corresponding backbone nulls of the observed anomaly-score distributions, and to remove such excesses jointly from the two samples whenever needed. This generalizes the density equalization procedure introduced in Springer et al. (2025) making the criterion more directly tied to the global tail behavior of the score distributions and better suited to identify domain shift, in which over- and under-densities may occur in nearby regions of space. For a fixed neighborhood depth K_M and class proportion $p = n_{\mathcal{Y}}/(n_{\mathcal{X}} + n_{\mathcal{Y}})$, the null distributions for over- and under-density scores are estimated by generating a large number of Bernoulli sequences of length K_M with parameters p and $1-p$, respectively, and computing the corresponding anomaly scores as in Section 2.1. For each direction, we then define a tail threshold τ_{tail} as a high null quantile (by default, the 0.97 quantile) and compare the observed and null score tails above τ_{tail} by means of a one-sided two-sample Kolmogorov–Smirnov test. We use a one-sided test because the goal of equalization is to reach a regime in which the empirical score tail is stochastically dominated by the corresponding null tail, meaning that extreme anomaly scores are no more frequent than expected under the backbone null. At each iteration, pruning is therefore activated on a given side only if the corresponding tail test rejects this condition.

When at least one of the two tail tests rejects the null hypothesis, we perform a bidirectional equalization step. On each active side, it select at both tails the most anomalous points (or, for computational efficiency, a small set of points from the corresponding score tail) and add them to the pruned representative sets $\hat{\mathcal{Y}}^+$ and/or $\hat{\mathcal{X}}^+$. The associated excess-mass neighborhoods are then removed from the current equalized samples, stopping at the first point belonging to the opposite sample, as in the implementation in Springer et al. (2025). In contrast to the procedure in Springer et al. (2025), however, this pruning is assessed jointly in the two directions, so that the neighborhood structure is updated while accounting for the total mass already removed from both samples. This is especially useful when over- and under-density modes lie in nearby regions of space.

To keep the procedure computationally efficient, it alternates between local pruning and a global score recomputation. During the pruning phase, equalization is carried out only with respect to the tail candidates identified at its start. After each pruning step, anomaly scores are recomputed only for the points belonging to the two tails, since these are the only ones needed to determine the next candidates for pruning. Once both tails have been equalized, anomaly scores are recomputed on the full equalized samples $\mathcal{X}_{\mathrm{eq}}$ and $\mathcal{Y}_{\mathrm{eq}}$, and the whole procedure is repeated until both KS tests pass. This outer iteration is important because pruning alters the neighborhood structure, so points that were initially outside the tails may enter them after the equalization step, while recomputing all scores after every pruning step would be computationally inefficient. Pseudocode for the equalization procedure is provided in the Supplementary Material, Sec. A.1.

2.3 NEIGHBOR-ENRICHED SUBSPACE LOCALIZATION (NESL)

In the final stage, the points extracted as described above are used to identify a subset of coordinates which determines the domain shift (if this subset exists). This stage, which, as we will see, is essential in the analysis of realistic data sets, is the most important algorithmic novelty introduced in this contribution. Let $\mathcal{U} \in \mathbb{R}^{n\times d}$ denote the pooled feature matrix, whose rows $u_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, are the individual samples of $\mathcal{X} \cup \mathcal{Y}$, and let $\mathcal{Q} \subset \mathcal{U}$ be the query set, consisting of the high-evidence candidates associated with a domain shift identified as described above, namely the pruned representative set $\hat{\mathcal{Y}}^+$ (or $\hat{\mathcal{X}}^+$). Let $\mathcal{T} \subset \mathcal{U}$ be a target set, chosen as either $\mathcal{Y}$ (if $\mathcal{Q} \subset \mathcal{Y}$) or $\mathcal{X}$ (otherwise). We seek a feature weighting under which the K nearest neighbors of the queries are maximally enriched in samples from $\mathcal{T}$. We introduce one nonnegative weight per feature. We distinguish between *raw* weights, which are the directly learned nonnegative parameters, and *effective* weights, which define the distance metric. The effective weights are obtained by rescaling the raw weights to scale at a constant d at each iteration to fix their overall magnitude, thereby removing the otherwise irrelevant global scale of the metric and preventing drift toward uniformly large or small weights. Distances are computed as,

$$D_{w_{\text{eff}}}(u_i, u_j) = \| w_{\text{eff}} \odot u_i - w_{\text{eff}} \odot u_j \|_2^2,$$

whereas the sparsity-inducing penalty is applied to the raw weights. This decouples scale control from sparsity promotion: normalization fixes the overall weight budget of the metric, while the penalty remains free to favor solutions concentrated on a relatively small number of features. The weights are learned through a differentiable soft KNN construction on stratified mini-batches $\mathcal{B} \subset \mathcal{U}$, sampled to maintain a fixed proportion of query and non-query samples. For each batch, weighted distances are converted into similarity logits,

$$\ell_{ij} = -\frac{D_{w_{\text{eff}}}(u_i, u_j)}{\tau},$$

and neighbor probabilities are given by a row-wise softmax restricted to the K largest logits:

$$P_{ij} = \frac{\exp(\ell_{ij})}{\sum_{j' \in \text{TopK}(i)} \exp(\ell_{ij'})} \text{ for } j \in \text{TopK}(i), \qquad P_{ij} = 0 \text{ otherwise.}$$

For each query $i \in \mathcal{Q} \cap \mathcal{B}$, we define the soft target mass $m_i = \sum_j P_{ij}\, \mathbb{1}\{j \in \mathcal{T} \cap \mathcal{B}\}$. The weights are then learned by maximizing the average soft target mass over the batch queries together with an ℓ_1 penalty on the raw weights:

$$\max \frac{1}{|\mathcal{Q} \cap \mathcal{B}|} \sum_{i \in \mathcal{Q} \cap \mathcal{B}} m_i \; - \; \lambda \|w_{\text{raw}}\|_1.$$

After optimization, features are ranked by their effective weights. This yields a compact set of coordinates that most strongly supports the localized discrepancy identified by the procedure in Section 2.2. For further details see Supp. A.2.

The overall domain-shift identification protocol proceeds by alternating anomaly localization in the current feature space with feature-subspace refinement around the detected modes. Once a discrepancy has been localized, the associated pruned points are partitioned into *modes*, that is, clusters of pruned points corresponding to distinct localized discrepancy regions, obtained by density-peak clustering of the pruned set as in (Springer et al., 2025; d'Errico et al., 2021); Neighbor-Enriched subspace localization is then applied to each mode separately. This mode-wise refinement is crucial because distinct modes may be supported by distinct subsets of coordinates. By isolating them individually, the procedure can reveal lower-dimensional subspaces in which the corresponding shift is sharper, more interpretable, and therefore easier to equalize selectively.

3 EXPERIMENTS

3.1 CONTROLLED 20D BENCHMARK FOR GLOBAL AND LOCAL DOMAIN-SHIFT RECOVERY

We first evaluate the proposed workflow in a controlled 20-dimensional benchmark with known ground truth. The reference distribution is a four-component Gaussian mixture, and each cohort contains $50{,}000$ background samples. On top of this common background, we consider two complementary

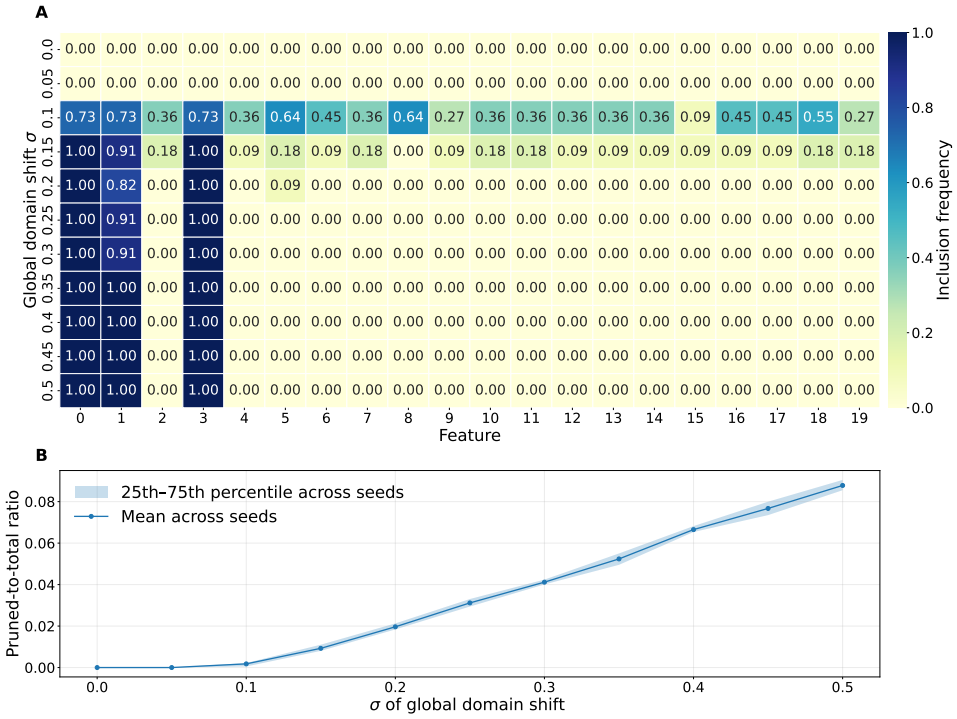

Figure 1: **Robustness of the identified shift set across global domain-shift amplitudes and random seeds. (A)** Inclusion frequency of each feature in the identified shift set, aggregated across random seeds, as a function of the global domain-shift amplitude σ. **(B)** Pruned-to-total ratio as a function of σ. The solid line denotes the mean across seeds, and the shaded region denotes the 25th–75th percentile range across seeds.

types of domain shift. In the first experiment, we introduce a *global* shift by displacing one Gaussian-mixture component along coordinates $(0, 1, 3)$, with amplitude σ. In the second experiment, we introduce a *localized* shift by injecting into one cohort a structured minority population supported on coordinates $(2, 4, 6, 8, 9)$. Full construction details are provided in Supp. A.4. This benchmark is designed to test two distinct capabilities of the method: first, whether it can recover a broad domain shift affecting an extended portion of the data and correctly identify the coordinates supporting it; and second, whether it can recover a localized minority shift embedded in a much larger background population and localize the relevant coordinates. In both experiments, we repeat the analysis across multiple random seeds to assess the stability of both recovery and subspace identification.

Global domain-shift recovery. The result of the analysis for the global shift is illustrated in Fig. 1. The shift amplitude is varied from $\sigma = 0$ to $\sigma = 0.5$ in increments of 0.05, and each configuration is repeated across 11 random seeds. For $\sigma = 0$ and $\sigma = 0.05$, the identified shift set is empty across seeds, indicating that the procedure does not produce spurious detections in the near-null regime. At $\sigma = 0.1$, the true supporting coordinates $(0, 1, 3)$ already begin to emerge, although some non-supporting coordinates are still selected intermittently. By $\sigma = 0.15$, the recovered set becomes visibly cleaner, and from $\sigma = 0.2$ onward the method enters a stable regime: panel A shows that coordinates $(0, 1, 3)$ dominate the identified shift set with high inclusion frequency, while spurious coordinates are almost entirely eliminated. Consistently, panel B shows that the pruned-to-total ratio increases steadily with σ, indicating that broader global mismatches are increasingly easy to isolate through pruning. These results show that the workflow can reliably detect a global domain shift and recover its true supporting coordinates once the shift exceeds a minimal detectable amplitude.

Localized domain-shift recovery We next consider a localized discrepancy supported on a small subset of the data, a condition which would be challenging for other approaches. As shown in Fig. 2, recovery improves sharply as the cardinality of the localized domain shift increases. For small

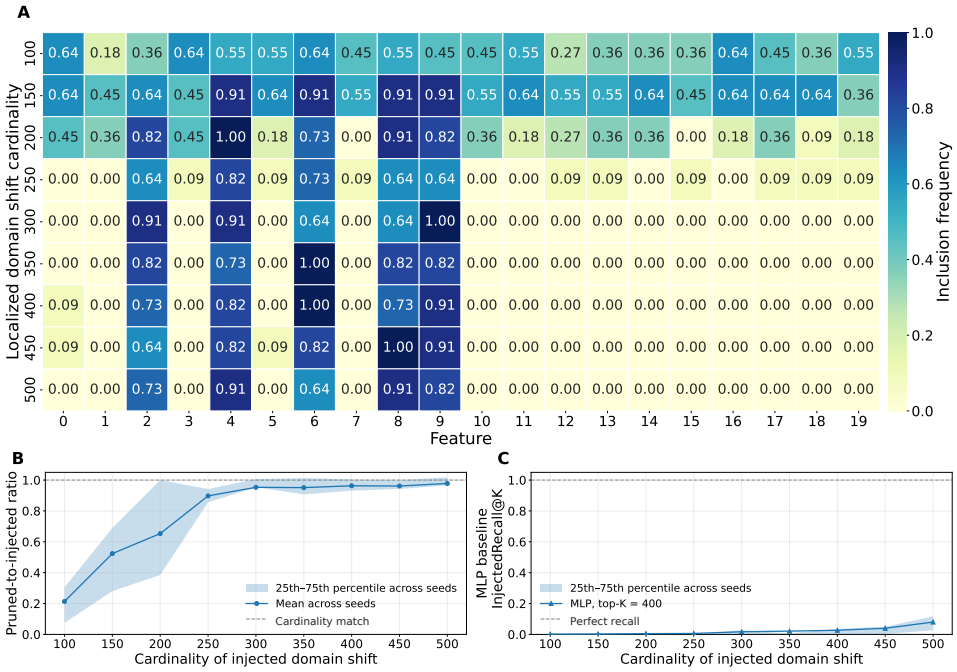

Figure 2: **Robustness of localized domain-shift recovery across injected shift cardinalities and random seeds. (A)** Inclusion frequency of each feature in the identified shift set, aggregated across random seeds, as a function of the localized domain-shift cardinality. **(B)** Pruned-to-injected ratio, i.e., the number of pruned samples divided by the number of injected samples, as a function of the cardinality of the injected localized shift. This ratio compares cardinalities, not set overlap; overlap-based recovery metrics are reported in Supp. A.5. The solid line denotes the mean across seeds, the shaded region denotes the 25th–75th percentile range across seeds, and the dashed horizontal line marks the cardinality-matched level (ratio equal to one). **(C)** MLP two-sample classifier baseline in the full 20-dimensional space. Target samples are ranked by the MLP density-ratio score, and the top-400 ranked samples are used as candidate anchors. $\mathrm{InjectedRecall@400}$ reports the fraction of injected localized-shift samples contained among these top-400 candidate anchors. The solid line and shaded region denote the mean and 25th–75th percentile range across seeds, respectively. A matched comparison of all methods at equal candidate-set size is given in Table 1.

injected populations (100–150 points), both the recovered feature sets and the pruned-to-injected ratio remain variable across random seeds, indicating that the signal is still weak relative to the background. From roughly 250 points onward, corresponding to only 0.5% of the data, however, the method enters a stable regime: panel A shows that the true supporting coordinates $(2, 4, 6, 8, 9)$ dominate the identified shift set with high inclusion frequency, while spurious coordinates are progressively eliminated. Consistently, panel B shows that the mean pruned-to-injected ratio rises from well below one at small cardinalities to about 0.9 at 250 points and above 0.95 for cardinalities of 300 and higher. Inter-seed variability follows the same trend, being largest at low cardinalities and markedly reduced once the localized shift is sufficiently represented. Since this ratio compares only cardinalities, Supp. A.5 also reports the injected-set recall and the fraction of pruned samples lying within the 95.45% probability-mass region of the injected Gaussian: from 250 points onward, recall in the final subspace reaches 1.00 and 77–94% of the pruned samples fall inside that region. As a representative classifier-based baseline, we also train an MLP two-sample discriminator (Lopez-Paz & Oquab, 2017). Panel C reports $\mathrm{InjectedRecall@400}$, the fraction of injected samples among the 400 target samples with the highest MLP density-ratio score, which remains below 0.1 at all cardinalities.

Table 1 compares the performance of this approach with EagleEye (Springer et al., 2025), the MLP discriminator, a KNN cohort-enrichment score, and uLSIF (Kanamori et al., 2009), by the fraction of injected samples among the top-n_{inj} candidates of each method.

Table 1: Matched comparison on the localized-shift benchmark: fraction of injected samples among the n_{inj} highest-scoring target samples for each method (EagleEye: initial score in the full feature space; KNN enrich.: KNN cohort-enrichment score). Cells report the median and the 25th–75th percentile range across 11 random seeds; best value per row in bold.

n_{inj}	Proposed	EagleEye	MLP	KNN enrich.	uLSIF
100	0.07 [0.02, 0.11]	0.10 [0.10, 0.17]	0.00 [0.00, 0.00]	0.07 [0.03, 0.12]	**0.11** [0.07, 0.12]
150	0.22 [0.10, 0.64]	**0.25** [0.24, 0.30]	0.00 [0.00, 0.00]	0.18 [0.13, 0.28]	0.14 [0.11, 0.16]
200	**0.64** [0.58, 0.93]	0.43 [0.41, 0.47]	0.00 [0.00, 0.00]	0.38 [0.33, 0.43]	0.16 [0.13, 0.18]
250	**0.90** [0.83, 0.96]	0.54 [0.52, 0.57]	0.00 [0.00, 0.00]	0.50 [0.45, 0.52]	0.16 [0.15, 0.22]
300	**0.97** [0.96, 0.99]	0.62 [0.58, 0.63]	0.00 [0.00, 0.01]	0.57 [0.54, 0.58]	0.23 [0.21, 0.26]
350	**0.97** [0.94, 0.99]	0.66 [0.63, 0.67]	0.00 [0.00, 0.00]	0.62 [0.60, 0.63]	0.20 [0.18, 0.22]
400	**0.96** [0.95, 0.98]	0.68 [0.66, 0.69]	0.01 [0.00, 0.04]	0.66 [0.64, 0.67]	0.27 [0.23, 0.28]
450	**0.98** [0.95, 0.99]	0.72 [0.69, 0.72]	0.02 [0.00, 0.06]	0.69 [0.67, 0.70]	0.28 [0.26, 0.33]
500	**0.98** [0.96, 0.99]	0.73 [0.70, 0.74]	0.06 [0.04, 0.15]	0.71 [0.69, 0.72]	0.32 [0.30, 0.34]

With 250 injected points the proposed workflow recovers 0.90 of the injected samples, whereas all other methods remain below 0.54; from 300 points onward it exceeds 0.96. The workflow thus stably recovers both the feature support of a localized shift and a representative set of the associated samples, with an accuracy that no other method achieves.

3.2 ECG DATASET: PTB-XL+ WITH 12SL FEATURES

We next evaluate the proposed workflow in a real clinical feature space, focusing on domain shifts among nominally healthy ECG cohorts. We use PTB-XL+, an extension of the publicly available PTB-XL dataset of 12-lead ECG recordings (Wagner et al., 2020; Strodthoff et al., 2023), and focus on its Marquette 12SL feature set. Each recording is represented by a 782-dimensional vector of clinically interpretable ECG features, including global and lead-specific amplitudes, intervals, and morphology-related measurements. Prior to all analyses, features are standardized to zero mean and unit variance so that differences in physical units and dynamic ranges do not influence the results. We use diagnostic statements and metadata to restrict the analysis to nominally healthy recordings and to construct cohort pairs matched for age and sex composition while varying measurement-device composition. Only the standardized 12SL feature vectors are provided to the method; diagnostic statements and metadata are used exclusively for cohort construction and post hoc interpretation. This experiment complements the synthetic benchmark by replacing known feature-level ground truth with a controlled real-world source of variation.

We analyzed recordings on patients which are all classified as healthy acquired with three measurement devices: `AT-6 C 5.5`, `CS-12 E`, and `CS-12`. For readability, we refer to these devices as `Device 1`, `Device 2`, and `Device 3`, respectively. We then constructed controlled pairs of cohorts differing in their measurement-device composition. For each comparison, two cohorts $\mathcal{X}$ and $\mathcal{Y}$ were sampled from these device-specific subsets according to prescribed mixture proportions and then matched for age and sex. Device identity was used only to construct the cohorts and to interpret the results post hoc. We considered four device-composition comparisons (see Figure 3). Cases A–C correspond to distinct primary device-mixture contrasts. In each of these designs, one device is fully assigned to one cohort, a second device is split approximately equally between the two cohorts, and the third device is absent from that cohort but fully assigned to the opposite cohort. Thus, each comparison contains one shared device component and two device-specific components, creating a controlled device-composition contrast while preserving a partially common acquisition background. Case D is a perturbed version of case C, in which the same dominant contrast is preserved but each cohort contains a small fraction of the device that is dominant in the opposite cohort. This design allows us to ask whether the method responds coherently to acquisition-related distributional changes: if the detected discrepancy is device-related, the pruned representative samples should be enriched in the device components that are imbalanced between the two cohorts.

Figure 3 summarizes the results. Across comparisons, the pruned representative sets $\hat{\mathcal{X}}^+$ and $\hat{\mathcal{Y}}^+$ contain roughly 3%–6% of the corresponding healthy cohort. The pruned sets are systematically enriched in the device components that determine the mismatch. Importantly, the procedure does not simply report a generic difference between healthy cohorts, but extracts representative samples

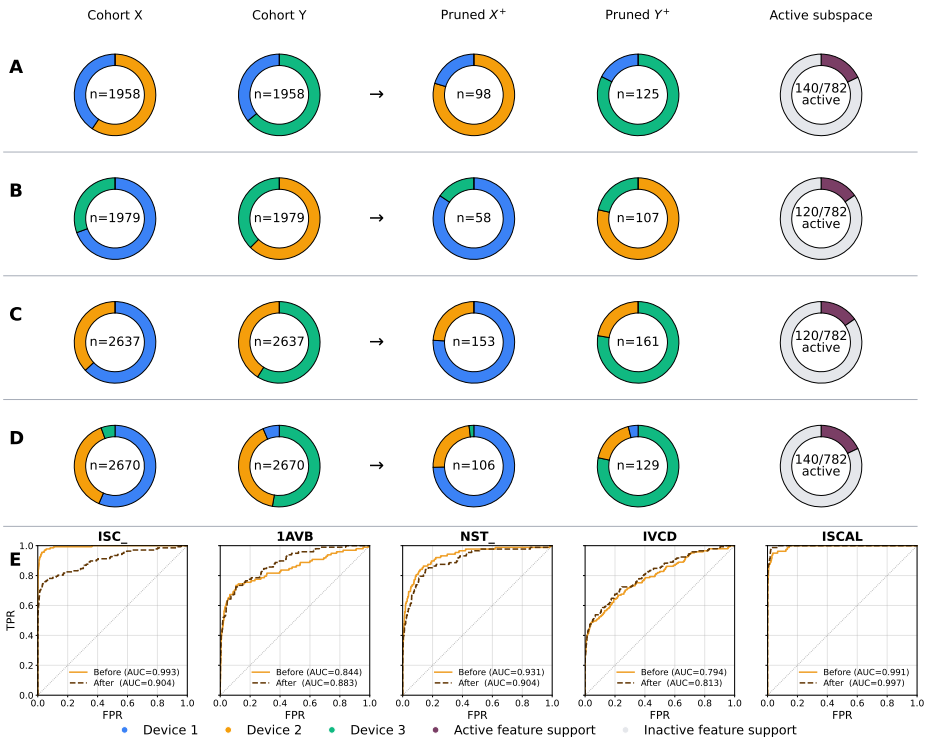

Figure 3: **Device-induced domain shifts in healthy ECG cohorts.** Each row corresponds to one controlled comparison between two age- and sex-matched healthy ECG cohorts, $\mathcal{X}$ and $\mathcal{Y}$, constructed from three measurement-device-specific subsets. Cases A–C are primary device-composition contrasts, each with one device specific to $\mathcal{X}$, one device approximately shared between cohorts, and one device specific to $\mathcal{Y}$. Case D is a perturbed version of Case C, preserving the same dominant contrast while introducing a small fraction of the opposite device into each cohort. The first two columns show the device composition of the input cohorts. The next two columns show the device composition of the pruned representative sets $\hat{\mathcal{X}}^+$ and $\hat{\mathcal{Y}}^+$ extracted by our procedure. The active-subspace column reports the size of the feature support selected by Neighbor-Enriched subspace localization using $\hat{\mathcal{Y}}^+$ as query samples. Row E shows AUCROC curves for binary NORM-vs-PATH classification on five illustrative pathologies, evaluated before and after equalization under the Case C device-composition split. The classifier shown in Row E was selected as the one with the largest mean absolute relative AUCROC change across the five tasks; all classifiers and the corresponding confusion matrices are reported in Supp. A.7 and Figures 4–5.

and the features associated with the device-induced domain shift. The last column of Fig. 3 reports the size of the active feature support selected in this $\hat{\mathcal{Y}}^+$-query subspace. The selected supports are relatively broad, ranging from $120/782$ to $140/782$ features across comparisons. Within this attribution, the discrepancy is therefore represented by a distributed active support rather than by a handful of isolated ECG descriptors. This is an important advantage of our approach: it allows identifying domain shifts when they are supported in very high-dimensional subspaces. To quantify the stability of the selected feature supports, we computed pairwise Jaccard overlaps between the active supports obtained from the $\hat{\mathcal{Y}}^+$-query subspaces across the four test cases. The perturbed case D showed strong overlap with case C ($J = 0.86$), consistent with its construction as a moderate perturbation of the same dominant device contrast. Cases A, C, and D also showed substantial overlap ($J = 0.79$–0.80), whereas case B had markedly lower overlap with the other comparisons ($J = 0.24$–0.26). Taken together, these results show that controlled changes in measurement-device composition induce detectable, directionally localized domain shifts in healthy ECG cohorts, and that the corresponding feature supports reflect the similarity of the underlying device-composition contrasts.

To assess whether device-associated equalization can affect downstream prediction, we performed a diagnostic classification check under the Case C device-composition split. Five PTB-XL+ pathology labels were used as illustrative NORM-vs-PATH tasks: 1AVB, NST_, IVCD, ISC_, and ISCAL. EagleEye equalization was estimated on the healthy cohorts only, without access to pathology labels or device identity, and fixed before pathology-positive recordings were added a posteriori according to the same device-composition design, with training samples matched to $\mathcal{X}$ and test samples matched to $\mathcal{Y}$; the classification step itself is supervised and serves only to assess the downstream effect of this unsupervised equalization. Three classifiers spanning linear and nonlinear models were evaluated before and after equalization (LR-L2, LR-EN, HGBT; see Supp. A.7 for details); row E of Figure 3 reports ROC curves for the classifier with the largest mean absolute relative AUCROC change across the five tasks, with all conditions provided in Figures 4–5. The variation in the reference set of healthy subjects leads in some cases to better predictions (1AVB), in other cases to worse predictions (ISC), but in all cases to *different* predictions. This indicates that, even in a highly curated dataset such as the one considered here, small domain shifts involving only a small percentage of the sample can affect the classification results. Our approach allows identifying those small shifts.

4 Discussion and limitations

The experiments presented show that the proposed framework can detect small domain shifts between unlabeled datasets, attribute them to interpretable feature subspaces, and construct equalized subsets, both in controlled benchmarks and in a real clinical data. Several limitations should be considered when interpreting the results presented above.

The equalized subsets $\mathcal{X}_{\mathrm{eq}}$ and $\mathcal{Y}_{\mathrm{eq}}$ produced by our procedure satisfy a statistical stopping criterion: the one-sided KS test on the anomaly-score tail no longer rejects at the prescribed significance level α. This means that the detected excess mass has been reduced under the local scoring backbone, but it does not imply that the two equalized subsets are drawn from identical distributions. Residual discrepancies below the sensitivity threshold may remain. Equalization should therefore be interpreted as reduction of the detected distributional discrepancy, not as a certificate of distributional equivalence. A related limitation is that the pipeline is triggered by the local anomaly score estimated in the initial feature space. If no discrepancy is detected in the full feature space, the Neighbor-Enriched Subspace Localization has no anchor set from which to start. This can occur in very high-dimensional settings when a weak discrepancy is confined to a small feature subset. The detection limit we find in our experiments is of approximately 0.5% for anomalies involving five features out of 20, but providing formal convergence guarantees is left to future work. The hyperparameters introduced beyond those of EagleEye (which are fixed according to the procedure in ref(Springer et al., 2025)) are kept at the same defaults in all experiments. Supp. A.6 shows stability under moderate variations of α, q_{tail}, and the main NESL parameters.

The ECG cohorts in Section 3.2 are matched for age and sex, but residual confounding may remain from variables not recorded or not controlled in PTB-XL+, such as body habitus, electrode placement protocol, or other acquisition-site effects. The detected discrepancies are consistent with device-associated acquisition effects, and the coherence of the feature supports across controlled device-composition contrasts supports this interpretation. However, the results should be read as identifying distributional discrepancies associated with measurement-device composition, not as proving device effect as their unique cause.

AI USE STATEMENT

Generative AI tools were not used to develop the method, generate or process data, or produce the reported results and analyses. An AI assistant was used to edit the manuscript for readability and concision and to assist with literature searches. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified ECG data (PTB-XL and PTB-XL+) together with synthetic data; no new human-subject data were collected. Detected discrepancies are statistical associations and should not be interpreted as causal effects or used to justify the exclusion of subpopulations without domain review; see Section 4.

REPRODUCIBILITY STATEMENT

The two procedures introduced in this work are specified in Section 2, with pseudocode for the equalization procedure in Appendix A.1 and implementation details of Neighbor-Enriched Subspace Localization in Appendix A.2. All hyperparameters, kept fixed across experiments, are listed in Appendix A.3. The synthetic benchmark is fully specified in Appendix A.4, including the mixture parameters and the injection protocol; the ECG cohort construction is described in Section 3.2 and the classification protocol in Appendix A.7. PTB-XL and PTB-XL+ are publicly available. Code and reproducibility materials will be released upon acceptance.

REFERENCES

Maria d'Errico, Elena Facco, Alessandro Laio, and Alex Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.01.010. URL https://www.sciencedirect.com/science/article/pii/S0020025521000116.

Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979. doi: 10.1214/aos/1176344722.

Arthur Gretton, Alexander J. Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M. Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. In Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.), *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009. doi: 10.7551/mitpress/9780262170055.003.0008.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002. doi: 10.1145/582415.582418.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press, Cambridge, MA, 2009. ISBN 9780262170055.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1394–1406, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

Mark F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986. doi: 10.1080/01621459.1986.10478337.

Sebastian Springer, Andre Scaffidi, Maximilian Autenrieth, Gabriella Contardo, Alessandro Laio, Roberto Trotta, and Heikki Haario. Detecting localized density anomalies in multivariate data via coin-flip statistics, 2025. URL https://arxiv.org/abs/2503.23927.

Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J. Aston, Ashish Sundar, Claus Graff, Jørgen K. Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, Markus Bär, and Tobias Schaeffter. Ptb-xl+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, 10(1):279, 2023. doi: 10.1038/s41597-023-02153-8. URL https://doi.org/10.1038/s41597-023-02153-8.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, MA, 2012. ISBN 9780262017091. doi: 10.7551/mitpress/9780262017091.001.0001.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008. doi: 10.1007/s10463-008-0197-x.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

A SUPPLEMENTARY INFORMATION

A.1 SUPPLEMENT:TAIL-BASED BIDIRECTIONAL EQUALIZATION

Algorithm 1 Tail-Based Bidirectional Equalization

Require: Samples $\mathcal{X}, \mathcal{Y}$, neighborhood depth K_M, tail quantile level q_{tail}, KS significance level α

Initialize equalized samples $\mathcal{X}_{\text{eq}} \leftarrow \mathcal{X}$, $\mathcal{Y}_{\text{eq}} \leftarrow \mathcal{Y}$

Initialize pruned representative sets $\hat{\mathcal{X}}^+ \leftarrow \emptyset$, $\hat{\mathcal{Y}}^+ \leftarrow \emptyset$

Estimate the two backbone null score distributions for $\mathcal{Y}$- and $\mathcal{X}$-overdensities

repeat

  Compute anomaly scores on all points in $\mathcal{X}_{\text{eq}}$ and $\mathcal{Y}_{\text{eq}}$ as in Section 2.1

  Set the tail thresholds $\tau_{\text{tail}}^{(\mathcal{Y}^+)}$ and $\tau_{\text{tail}}^{(\mathcal{X}^+)}$ from the corresponding null quantiles at level q_{tail}

  Define the current tail candidates

$$\mathcal{C}^{(\mathcal{Y}^+)} \leftarrow \{\mathcal{Y}_i \in \mathcal{Y}_{\text{eq}} : \Upsilon_i^{(\mathcal{Y}^+)} \geq \tau_{\text{tail}}^{(\mathcal{Y}^+)}\}, \quad \mathcal{C}^{(\mathcal{X}^+)} \leftarrow \{\mathcal{X}_j \in \mathcal{X}_{\text{eq}} : \Upsilon_j^{(\mathcal{X}^+)} \geq \tau_{\text{tail}}^{(\mathcal{X}^+)}\}$$

  Run one-sided two-sample KS tests comparing the empirical tails of $\mathcal{C}^{(\mathcal{Y}^+)}$ and $\mathcal{C}^{(\mathcal{X}^+)}$ against the corresponding null tails

  while at least one of the two KS tests rejects at level α **do**

    if $\mathcal{C}^{(\mathcal{Y}^+)}$ is active **then**

      Select the most extreme point(s) in $\mathcal{C}^{(\mathcal{Y}^+)}$

      Let $\mathcal{N}^{(\mathcal{Y}^+)} \subset \mathcal{Y}_{\text{eq}}$ denote the selected point(s) together with all neighbors up to the first point belonging to $\mathcal{X}_{\text{eq}}$

      Update $\hat{\mathcal{Y}}^+ \leftarrow \hat{\mathcal{Y}}^+ \cup \mathcal{N}^{(\mathcal{Y}^+)}$ and $\mathcal{Y}_{\text{eq}} \leftarrow \mathcal{Y}_{\text{eq}} \setminus \mathcal{N}^{(\mathcal{Y}^+)}$

    end if

    if $\mathcal{C}^{(\mathcal{X}^+)}$ is active **then**

      Select the most extreme point(s) in $\mathcal{C}^{(\mathcal{X}^+)}$

      Let $\mathcal{N}^{(\mathcal{X}^+)} \subset \mathcal{X}_{\text{eq}}$ denote the selected point(s) together with all neighbors up to the first point belonging to $\mathcal{Y}_{\text{eq}}$

      Update $\hat{\mathcal{X}}^+ \leftarrow \hat{\mathcal{X}}^+ \cup \mathcal{N}^{(\mathcal{X}^+)}$ and $\mathcal{X}_{\text{eq}} \leftarrow \mathcal{X}_{\text{eq}} \setminus \mathcal{N}^{(\mathcal{X}^+)}$

    end if

    Update $\mathcal{C}^{(\mathcal{Y}^+)} \leftarrow \mathcal{C}^{(\mathcal{Y}^+)} \cap \mathcal{Y}_{\text{eq}}$ and $\mathcal{C}^{(\mathcal{X}^+)} \leftarrow \mathcal{C}^{(\mathcal{X}^+)} \cap \mathcal{X}_{\text{eq}}$

    Recompute anomaly scores only for points currently belonging to $\mathcal{C}^{(\mathcal{Y}^+)}$ and $\mathcal{C}^{(\mathcal{X}^+)}$, then re-evaluate the two KS tests on the updated tails

  end while

until both KS tests do not reject after global recomputation

return $\hat{\mathcal{X}}^+, \hat{\mathcal{Y}}^+, \mathcal{X}_{\text{eq}}, \mathcal{Y}_{\text{eq}}$

A.2 SUPPLEMENT: NEIGHBOR-ENRICHED FEATURE SELECTION (IMPLEMENTATION DETAILS)

S1: Inputs, target/query definition, and defaults. C1 operates on a feature matrix $\mathcal{U} \in \mathbb{R}^{n \times d}$ (rows are samples, columns are features), which in our experiments is obtained by pooling the two cohorts used by EagleEye. We assume $\mathcal{U}$ is standardised per feature (zero mean and unit variance) before running feature selection (*in our experiments we perform this preprocessing step once and reuse it throughout*).

The neighbour-enrichment objective is defined through two boolean masks of length n:

- `target_mask`$\in \{0,1\}^n$: `target_mask`[j]=1 if sample j is a desired neighbour (a "hit").
- `query_mask`$\in \{0,1\}^n$: `query_mask`[i]=1 if sample i is a query/anchor whose neighbourhood we aim to enrich in target samples.

In the $\mathcal{X}/\mathcal{Y}$ setting used throughout the paper, we typically set the target set to the test cohort ($\mathcal{T} = \mathcal{Y}$), i.e. `target_mask` marks the rows of the pooled matrix that belong to $\mathcal{Y}$, and set the query set $\mathcal{Q} \subseteq \mathcal{Y}$ to the pruned returned by EagleEye (so `query_mask` marks only those selected $\mathcal{Y}$ rows). If `query_mask` is not provided, the implementation defaults to `query_mask=target_mask`.

Unless stated otherwise we use the following defaults: $K = 100$ neighbours, $\beta = 0.2$ (rank tie-break strength), $T = 3000$ epochs, mini-batch size $B = 200$, stratified query fraction `pos_frac`$= 0.5$, temperature annealing $\tau_{\text{start}} = 1.0 \to \tau_{\text{end}} = 0.1$ (geometric), Adam learning rate `lr`$= 10^{-2}$, ℓ_1 penalty `l1`$= 10^{-2}$, 5-fold cross-validation (`n_splits`$= 5$), and random seed 0.

Mini-batches are stratified by `query_mask`: `pos_frac` is the fraction of *query* samples included in each batch (default 0.5).

S2: Nonnegative weights and scale-normalised effective weights. We learn one nonnegative weight per feature via an unconstrained parameter vector $\ell \in \mathbb{R}^d$ (stored as `logw` in the implementation), initialised to zeros. The optimiser updates ℓ; the weights w_{raw} and w_{eff} are deterministic functions of ℓ. At each optimisation step we map ℓ to *raw* nonnegative weights through the Softplus transform,

$$w_{\text{raw},j} = \text{softplus}(\ell_j) = \log\bigl(1 + e^{\ell_j}\bigr) \quad \Rightarrow \quad w_{\text{raw}} \in \mathbb{R}^d_{\geq 0}.$$

This guarantees nonnegativity at every iteration while remaining smooth for gradient-based optimisation; weights can become arbitrarily small (effectively zero in practice) as $\ell_j \to -\infty$.

Because multiplying all weights by a common constant does not change neighbour identities up to a global rescaling of distances, unconstrained optimisation can drift toward trivially large (or small) weights. We therefore define *effective* weights used in the distance metric by fixing the global scale to have sum approximately d:

$$w_{\text{eff}} \;=\; w_{\text{raw}} \cdot \frac{d}{\sum_{j=1}^{d} w_{\text{raw},j}}.$$

Importantly, the implementation *stops gradients* through the denominator (i.e., it treats the normalising sum as a constant when computing parameter updates) and adds a small stabiliser:

$$w_{\text{eff}} \;=\; w_{\text{raw}} \cdot \frac{d}{\text{stopgrad}\Bigl(\sum_{j=1}^{d} w_{\text{raw},j}\Bigr) + 10^{-8}}.$$

The resulting weights satisfy $\sum_j w_{\text{eff},j} \approx d$, up to the 10^{-8} stabiliser.

Stopping gradients through the normalisation prevents the optimiser from compensating the sparsity penalty (by shrinking the denominator) instead of genuinely driving many feature weights toward zero; it also makes the effective weights comparable across runs because their overall "budget" is fixed. Distances and neighbour probabilities are computed using w_{eff}, while sparsity is enforced via an ℓ_1 penalty on w_{raw} (see S4).

S3: In-batch soft KNN neighbourhood probabilities. Given a mini-batch feature matrix $Z_b \in \mathbb{R}^{B \times d}$ (rows are samples), we compute neighbour probabilities $P \in [0,1]^{B \times B}$ using the current effective weights w_{eff} (from S2).

We first rescale features elementwise and compute all pairwise squared Euclidean distances within the batch:

$$\tilde{Z}_b = Z_b \odot w_{\text{eff}}, \qquad d^2_{ij} = \|\tilde{z}_i - \tilde{z}_j\|_2^2, \qquad i, j \in \{1, \ldots, B\}.$$

We convert distances into similarity logits (unnormalised similarity scores) via

$$\ell_{ij} = -\frac{d^2_{ij}}{\max(10^{-6}, \tau)},$$

so larger logits mean "more similar". The $\max(10^{-6}, \tau)$ guard prevents numerical issues when τ is very small, and self-matches are excluded by subtracting a large constant from the diagonal:

$$\ell_{ii} \leftarrow \ell_{ii} - 10^9.$$

Neighbour probabilities are obtained by applying a row-wise softmax restricted to the K largest logits in each row. Concretely, for each i we keep only the indices of the top-K logits $\{\ell_{ij}\}_j$, set the remaining entries to $-\infty$, and apply a softmax:

$$P_{ij} = \frac{\exp(\ell_{ij})}{\sum_{j' \in \text{TopK}(i)} \exp(\ell_{ij'})} \;\text{ for }\; j \in \text{TopK}(i), \qquad P_{ij} = 0 \;\text{ otherwise.}$$

If $K \geq B$, we default to a full row-wise softmax over all non-diagonal entries.

The resulting matrix P satisfies $\sum_j P_{ij} = 1$ for each row i and can be interpreted as a soft assignment distribution over the K nearest neighbours of sample i within the current mini-batch.

S4: Training objective and optimisation procedure. We learn feature weights by maximising *soft target enrichment* of the query samples' neighbourhoods, using mini-batch stochastic optimisation with Adam.

Let `query_mask` mark the query set $\mathcal{Q}$ and let the remaining samples be *non-queries*. Each mini-batch corresponds to an index set $\mathcal{B} \subseteq \mathcal{U}$ of size B and is constructed by sampling, without replacement, approximately $(B \cdot \texttt{pos_frac})$ query indices and $(B - B \cdot \texttt{pos_frac})$ non-query indices, then shuffling the concatenated indices. Across epochs we anneal the softmax temperature τ *geometrically* from τ_{start} to τ_{end}.

Within a batch, we compute neighbour probabilities P via the soft KNN construction in S3 and define, for each batch sample i, the *soft target mass*

$$m_i \;=\; \sum_{j=1}^{B} P_{ij}\, \mathbb{1}\{j \in \mathcal{T} \cap \mathcal{B}\},$$

where the target indicator corresponds to `target_mask` restricted to the current batch. Only batch samples that are queries contribute to the objective. The batch objective is

$$\text{obj} \;=\; \frac{1}{|\mathcal{Q} \cap \mathcal{B}|} \sum_{i \in \mathcal{Q} \cap \mathcal{B}} m_i \;,$$

(with the update skipped if $\mathcal{Q} \cap \mathcal{B} = \emptyset$). We maximise the objective while promoting sparsity via an ℓ_1 penalty on raw weights:

$$\mathcal{L} \;=\; -\,\text{obj} \;+\; \lambda \sum_{j=1}^{d} w_{\text{raw},j},$$

where $\lambda = 10^{-2}$ and w_{raw} is the Softplus-transformed nonnegative weight vector from S2. We use Adam on the underlying unconstrained parameters with learning rate $\texttt{lr} = 10^{-2}$ and default Adam hyperparameters.

S5: From weights to a discrete feature subset. After training, we use the effective weights w_{eff} to rank features and to select the subset size m via a leakage-safe hard KNN cross-validation procedure (with the same K used throughout the paper).

For each candidate subset size m, we take the indices of the m largest weights, $\text{TopM}(w_{\text{eff}}, m)$. To speed up the search, the candidate values of m are restricted to a predefined grid. In the current implementation, the grid starts from d and then sweeps down to 150 in steps of 50, then continues down to 50 in steps of 5, then down to 16 in steps of 2, and finally considers each integer from 15 down to 1.

To choose m, we evaluate each candidate subset $\text{TopM}(w_{\text{eff}}, m)$ using K-fold cross-validation. Here, "training" refers to the portion of the pooled dataset used to *define neighbourhoods* for that fold, while "validation" refers to held-out query samples used only to *test* how target-enriched their neighbourhoods are under the chosen subset.

Concretely, in each fold we split the pooled indices into a training set and a validation set. We then compute hard K nearest neighbours for each validation query sample using distances in the selected feature subset, but we restrict the neighbour search to the training set only (validation samples are never used as neighbours of one another). This avoids optimistic bias when assessing enrichment.

We split the data into K folds in a stratified manner that preserves the query/non-query proportion across folds and ensures that each fold contains both groups whenever possible.

For a validation query sample i, let c_i be the number of target samples among its K nearest neighbours, and let $\phi_i = c_i/K$ denote the corresponding neighbourhood purity.

When several subsets yield similar purity, we prefer the one where the target neighbours tend to appear *earlier* (i.e., at smaller ranks, hence closer). To quantify this, we use a standard rank-discounted score from information retrieval (discounted cumulative gain, DCG, and its normalized variant nDCG; (Järvelin & Kekäläinen, 2002; Manning et al., 2008)): each target neighbour contributes to the score, but contributions are down-weighted as the neighbour rank increases. We normalize by the best possible value given the same hit count c_i, yielding a normalized rank-sensitivity $q_i \in [0, 1]$ with $q_i = 1$ when the c_i target neighbours are the closest ones.

We then combine purity and rank sensitivity as

$$s_i = \phi_i + \beta(1 - \phi_i)q_i,$$

and define the subset score as the mean of s_i over validation queries, averaged across folds. In addition to this main score (with the user-specified β), we also report a "purity-only" curve obtained by setting $\beta = 0$, which corresponds to the target fraction among hard KNN neighbours.

Finally, we select $m^\star$ as the value in the grid that maximizes the cross-validated mean score, and return $\mathrm{TopM}(w_{\text{eff}}, m^\star)$ as the selected feature subset.

A.3 Default hyperparameters

Unless otherwise stated, all experiments use the default hyperparameters reported below. These values were kept fixed across the synthetic and ECG experiments, except when a parameter was explicitly varied as part of the experimental design.

EagleEye local two-sample scoring.

- Maximum neighbourhood depth: $K_M = 400$ for the synthetic benchmark and $K_M = 200$ for the ECG experiments.
- Null calibration: Monte Carlo calibration using Bernoulli sequences as in (Springer et al., 2025).

Clustering

- We use Density Peaks Advanced clustering with default parameters as in (Springer et al., 2025).

Bidirectional tail-KS equalization.

- Tail threshold: $q_{\text{tail}} = 0.97$ null quantile.
- Tail test: one-sided two-sample Kolmogorov–Smirnov test comparing observed and null anomaly-score tails.
- KS significance level: $\alpha = 0.05$, unless otherwise stated.
- Pruning rule: pruning is activated on a side only when the corresponding one-sided KS tail test rejects at level α.
- Pruned neighbourhood: for each selected tail point, samples are removed up to the first neighbour belonging to the opposite dataset.
- Recalculation strategy: anomaly scores are updated locally within the active tails during pruning and recomputed globally after both tail tests stop rejecting.
- Stopping rule: the procedure stops when both one-sided KS tail tests no longer reject after global recomputation.

Neighbor-Enriched Subspace Localization.

- Number of neighbours in the enrichment objective: $K = 100$.
- Feature weights: one nonnegative raw weight per feature, transformed through a softplus parameterization.

- Effective weight normalization: effective weights are rescaled to have total weight approximately equal to the ambient dimension d.
- Sparsity penalty: $\lambda = 10^{-2}$ on raw feature weights.
- Optimizer: Adam.
- Learning rate: 10^{-2}.
- Number of epochs: 3000.
- Mini-batch size: 200.
- Query fraction in stratified mini-batches: 0.5.
- Temperature schedule: geometric annealing from $\tau_{\text{start}} = 1.0$ to $\tau_{\text{end}} = 0.1$.
- Model selection: leakage-safe hard-KNN cross-validation over candidate subset sizes.
- Number of cross-validation folds: 5.
- Rank-discount strength for subset scoring: $\beta = 0.2$.
- Random seed: 0, unless otherwise stated.

MLP two-sample classifier baseline.

- Task: binary discrimination between samples from $\mathcal{X}$ and $\mathcal{Y}$.
- Input space: full 20-dimensional feature space for the synthetic localized-shift benchmark.
- Architecture: fully connected multilayer perceptron with hidden layers $(128, 64, 32)$.
- Activation: ReLU.
- Optimizer: Adam.
- Learning rate: 10^{-3}.
- L_2 regularization parameter: 10^{-4}.
- Batch size: 256.
- Maximum number of training iterations: 300.
- Density-ratio score: target samples are ranked by the classifier posterior-odds score,

$$\hat{r}(x) = \frac{\hat{p}(y = \mathcal{Y} \mid x)}{1 - \hat{p}(y = \mathcal{Y} \mid x)}.$$

- Anchor-recovery evaluation: $\mathrm{InjectedRecall@400}$, defined as the fraction of injected localized-shift samples contained among the top-400 ranked target samples.

A.4 Controlled 20D benchmark construction

We constructed a controlled 20-dimensional benchmark to evaluate recovery of both sample-level domain shifts and their supporting feature coordinates under known ground truth. The background distribution is a four-component Gaussian mixture in $\mathbb{R}^{20}$. Unless otherwise stated, each reference cohort contains $50{,}000$ background samples.

Background Gaussian mixture. The feature space is $d = 20$. The reference distribution is

$$P_0(x) = \sum_{m=1}^{4} \pi_m \mathcal{N}(x; \mu_m, \Sigma_m),$$

with mixture weights

$$\pi = (0.35, 0.30, 0.20, 0.15).$$

The component means are

$\mu_1 = (0.0, 0.0, 0.5, -0.5, 0.0, 0.3, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0),$
$\mu_2 = (2.5, -1.0, -0.5, 1.0, 0.5, -0.3, 0.0, 0.2, -0.2, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0),$
$\mu_3 = (-2.0, 1.5, 0.0, 0.5, -1.0, 0.0, 0.3, -0.2, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0),$
$\mu_4 = (1.0, 2.0, -1.0, -1.0, 0.0, 0.5, -0.5, 0.0, 0.0, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0).$

All covariance matrices are diagonal. Their diagonal entries are

$$\mathrm{diag}(\Sigma_1) = (1.2, 1.0, 0.8, 0.9, 0.7, 0.8, 1.0, 1.0, 0.9, 1.0, 1.0, 1.0, 1.0, 0.8, 0.8, 0.8, 0.9, 0.9, 0.9, 0.9),$$
$$\mathrm{diag}(\Sigma_2) = (0.9, 1.1, 0.7, 0.8, 0.8, 0.7, 1.0, 0.9, 1.0, 1.0, 1.0, 0.9, 1.0, 0.8, 0.8, 0.8, 1.0, 1.0, 0.9, 0.9),$$
$$\mathrm{diag}(\Sigma_3) = (1.0, 0.8, 1.0, 0.9, 0.7, 1.1, 0.8, 0.9, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9, 0.8, 0.8, 0.8, 0.9, 0.9, 1.0),$$
$$\mathrm{diag}(\Sigma_4) = (1.1, 0.9, 0.8, 1.0, 0.8, 0.8, 0.9, 1.0, 1.0, 0.9, 1.0, 1.0, 0.9, 0.8, 0.8, 0.8, 0.9, 0.9, 1.0, 1.0).$$

Global domain-shift experiment. The global domain-shift experiment introduces a systematic perturbation in one background mixture component. Specifically, component $m = 2$ is shifted in the second cohort along coordinates

$$\{0, 1, 3\}.$$

The displacement magnitude is controlled by the scalar parameter δ, denoted by σ in the main text. We vary

$$\sigma \in \{0, 0.05, 0.10, \ldots, 0.50\}.$$

For each value of σ, the analysis is repeated across 11 random seeds.

Localized domain-shift experiment. The localized domain-shift experiment injects a compact minority population into one cohort. The injected population is associated with component $m = 1$ and is supported on the ground-truth feature set

$$\{2, 4, 6, 8, 9\}.$$

The number of injected samples is varied as

$$\{100, 150, 200, 250, 300, 350, 400, 450, 500\}.$$

Injected samples are generated around an anchor defined relative to the standard deviations of the selected background component. Along the active dimensions S_{local}, the anchor offset in units of component standard deviation is

$$(1.1, -0.4, 0.4, -0.2, 0.3).$$

The injected cluster has anisotropic geometry with scales, again expressed in units of the component standard deviations,

$$(0.11, 0.08, 0.04, 0.04, 0.03),$$

and a random rotation is applied within the active subspace. In ambient inactive dimensions, the injected samples follow the corresponding background component with shrinkage factor 0.7.

A.5 Localized domain shift: overlap-based recovery and matched baseline comparison

The pruned-to-injected ratio of Fig. 2B compares the cardinality of the pruned set with that of the injected population and does not measure their overlap. Since injected and background samples occupy the same local region, the procedure aims at removing the localized excess mass rather than at reconstructing the identity of each injected point. Table 2 therefore reports two overlap-based measures: the injected-set recall, $|\text{Pruned} \cap \text{Injected}|/|\text{Injected}|$, computed both in the full feature space and in the final subspace selected by the workflow, and the fraction of pruned samples lying within the 95.45% probability-mass ellipsoid of the Gaussian used to generate the injected population in the true active subspace.

Table 1 in the main text compares the proposed workflow with pointwise scoring baselines on the same sample-localization task. For each injected cardinality n_{inj}, every method ranks the target samples by its discrepancy score, the n_{inj} highest-scoring samples are selected, and the fraction of injected samples among them is reported. Besides the MLP discriminator of Sec. A.3, the baselines are: the EagleEye score Υ_i computed in the initial full feature space; a plain KNN cohort-enrichment score, defined as the fraction of $\mathcal{Y}$ samples among the $K = 400$ nearest neighbours of each target sample in the pooled set; and the direct density-ratio estimator uLSIF (Kanamori et al., 2009). The proposed workflow uses the pruned set obtained after subspace refinement.

Table 2: Overlap-based recovery of the injected localized shift as a function of the injected cardinality n_{inj}. Cells report the median and the 25th–75th percentile range across 11 random seeds.

n_{inj}	Recall (full space)	Recall (final subspace)	Pruned within 95.45% mass
100	0.02 [0.01, 0.03]	0.00 [0.00, 0.01]	0.05 [0.01, 0.09]
150	0.07 [0.03, 0.11]	0.03 [0.01, 0.62]	0.22 [0.15, 0.46]
200	0.21 [0.16, 0.28]	0.57 [0.46, 1.00]	0.36 [0.30, 0.82]
250	0.39 [0.35, 0.45]	1.00 [0.99, 1.00]	0.77 [0.69, 0.89]
300	0.55 [0.52, 0.62]	1.00 [1.00, 1.00]	0.93 [0.89, 0.95]
350	0.68 [0.66, 0.74]	1.00 [1.00, 1.00]	0.91 [0.87, 0.96]
400	0.78 [0.74, 0.80]	1.00 [1.00, 1.00]	0.93 [0.89, 0.95]
450	0.85 [0.84, 0.88]	1.00 [1.00, 1.00]	0.92 [0.90, 0.95]
500	0.90 [0.89, 0.90]	1.00 [1.00, 1.00]	0.94 [0.92, 0.95]

A.6 Sensitivity analysis

We assess the sensitivity of the two stages introduced in this work to their hyperparameters, in the localized-shift setting of Sec. 3.1, with 11 independent seeds per configuration.

Equalization parameters. With an injected population of 300 points, we run a single equalization pass without subspace refinement, so that the measured effect is attributable to α and q_{tail} alone, and vary the two parameters jointly on a 3×3 grid around their nominal values. Table 3 reports the size of the pruned set, the fraction of pruned samples lying within the 95.45% probability-mass region R_{95} of the injected Gaussian, and the same intersection normalized by n_{inj}. The procedure is stable across the grid, with the expected precision–coverage trade-off as q_{tail} increases.

Table 3: Sensitivity of a single equalization pass to α and q_{tail} ($n_{\mathrm{inj}} = 300$). Cells report the median and the 25th–75th percentile range across 11 seeds.

α	q_{tail}	|Pruned|	$\lvert\text{Pruned} \cap R_{95}\rvert / \lvert\text{Pruned}\rvert$	$\lvert\text{Pruned} \cap R_{95}\rvert / n_{\mathrm{inj}}$
0.03	0.95	194 [172, 220]	0.73 [0.67, 0.78]	0.46 [0.44, 0.48]
0.03	0.97	178 [162, 200]	0.75 [0.72, 0.80]	0.44 [0.42, 0.46]
0.03	0.99	148 [135, 170]	0.81 [0.78, 0.83]	0.41 [0.38, 0.42]
0.05	0.95	200 [184, 220]	0.71 [0.67, 0.75]	0.46 [0.44, 0.48]
0.05	0.97	187 [163, 210]	0.74 [0.70, 0.78]	0.44 [0.42, 0.47]
0.05	0.99	148 [135, 170]	0.80 [0.78, 0.83]	0.41 [0.38, 0.42]
0.07	0.95	209 [184, 234]	0.71 [0.65, 0.75]	0.46 [0.44, 0.48]
0.07	0.97	187 [163, 220]	0.74 [0.70, 0.78]	0.44 [0.42, 0.47]
0.07	0.99	148 [135, 170]	0.80 [0.78, 0.83]	0.41 [0.38, 0.42]

NESL parameters. With an injected population of 400 points, we start for each seed from the pruned set produced by the first equalization pass at nominal parameters and keep it fixed, so that the measured effect is attributable to the NESL parameters alone. Mini-batch size, number of neighbours in the enrichment objective, and ℓ_1 penalty are varied one at a time. Since this is the first pass of an iterative procedure that subsequently refines the subspace, we assess the ordering of the coordinates by learned weight: with $G = \{2, 4, 6, 8, 9\}$ the true active support, Table 4 reports the fraction of G falling within the 5 and the 8 highest weights, and the rank of the lowest-weighted true coordinate. In every configuration, a median of four to five of the five true features lies within the top eight.

A.7 ECG dataset: classification

To assess whether device-associated equalization can affect downstream prediction, we performed a diagnostic classification check under the same Case C device-composition split used in the ECG equalization experiment. We considered five PTB-XL+ pathology labels, used here as illustrative NORM-vs-PATH classification tasks: first-degree atrioventricular block (1AVB), nonspecific ST changes (NST_), intraventricular conduction disturbance (IVCD), ischemic ST-T changes (ISC_), and ischemic anterolateral ST-T changes (ISCAL). The 782-dimensional Marquette 12SL feature

Table 4: Sensitivity of the first NESL pass to its main parameters ($n_{\text{inj}} = 400$, pruned set fixed). Cells report the median and the 25th–75th percentile range across 11 seeds.

Batch size	K	λ	Top-5	Top-8	Worst true rank
100	100	0.01	0.80 [0.60, 0.80]	1.00 [0.80, 1.00]	8 [7, 10]
200	100	0.01	0.60 [0.40, 0.60]	0.80 [0.60, 0.80]	12 [9, 13]
300	100	0.01	0.60 [0.40, 0.60]	0.80 [0.60, 0.80]	12 [11, 13]
300	200	0.01	0.40 [0.40, 0.60]	0.80 [0.60, 0.80]	12 [10, 13]
100	100	0.003	0.60 [0.50, 0.60]	0.80 [0.80, 0.80]	9 [9, 13]
100	100	0.03	0.60 [0.40, 0.60]	0.80 [0.60, 0.80]	9 [9, 11]



Figure 4: **Confusion matrices for NORM-vs-PATH binary classification before and after EagleEye equalization.** Rows correspond to the five illustrative pathology tasks (ISC_, 1AVB, NST_, IVCD, ISCAL); columns correspond to six conditions: three classifiers (LR-L2, LR-EN, HGBT) evaluated before equalization (left block) and after equalization (right block). Each 2×2 matrix reports prediction counts with predicted class on the horizontal axis and true class on the vertical axis, using NORM and PATH labels. Cohorts are constructed under the Case C device-composition split; see Section A.7 for classifier details.

vectors were used as input after the standardization described in Section 3.2 of the main text, with no additional task-specific preprocessing.

We evaluated three classifiers. *(i) LR-L2*: logistic regression with ℓ_2 regularization (`penalty="l2"`, $C = 0.01$, `solver="lbfgs"`, `max_iter=2000`), used as a ridge-regularized linear baseline. *(ii) LR-EN*: logistic regression with ElasticNet regularization (`penalty="elasticnet"`, $C = 0.01$, ℓ_1`_ratio=0.5`, `solver="saga"`, `max_iter=3000`), providing a sparse linear baseline through the mixed ℓ_1/ℓ_2 penalty. *(iii) HGBT*: histogram gradient-boosted trees (`max_depth=3`, `learning_rate=0.03`, `min_samples_leaf=30`, `max_iter=500`), with early stopping on

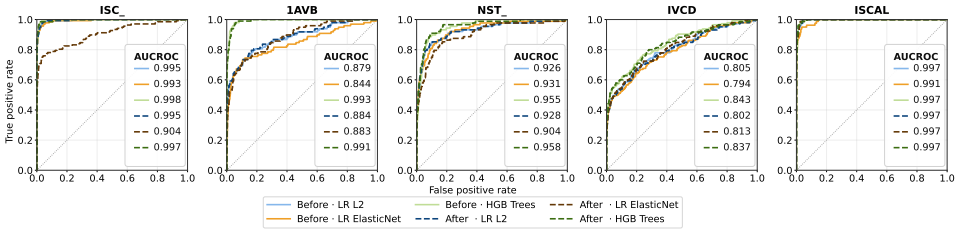

Figure 5: **ROC curves for NORM-vs-PATH binary classification before and after EagleEye equalization.** Each panel corresponds to one of the five illustrative pathology tasks. Solid curves show performance before equalization; dashed curves show performance after equalization. Colors distinguish LR-L2, LR-EN, and HGBT, and AUCROC values are reported inside each panel. The classifier shown in row E of Figure 3 of the main text is selected as the one with the largest mean absolute relative AUCROC change across the five pathology tasks; this supplementary figure reports all three classifiers for completeness. Cohorts are constructed under the Case C device-composition split; see Section A.7 for classifier details.

a 15% validation fraction (`n_iter_no_change=20`). The shallow depth and minimum leaf-size constraint were used to limit overfitting in the imbalanced binary tasks.

Class imbalance was handled using inverse-frequency class weights computed on the training set, $w_0 = 1$ and $w_1 = n_{\mathrm{NORM}}/n_{\mathrm{PATH}}$. No oversampling or resampling was applied. EagleEye equalization was first estimated on the healthy cohorts of Case C only. The downstream NORM-vs-PATH tasks were then constructed a posteriori by adding pathology-positive recordings before and after equalization according to the same device-composition design: pathology-positive training samples matched cohort $\mathcal{X}$, pathology-positive test samples matched cohort $\mathcal{Y}$, and samples from the device shared by $\mathcal{X}$ and $\mathcal{Y}$ were split 50–50. For each of the 5 cases, classifiers were trained on cohort $\mathcal{X}$ plus the corresponding pathology-positive training samples and tested on cohort $\mathcal{Y}$ plus the corresponding pathology-positive test samples, before and after equalization. Performance is summarized by ROC curves, AUCROC, and confusion matrices at the default 0.5 decision threshold. The full before/after results for all three classifiers are shown in Figure 5, with the corresponding confusion matrices reported in Figure 4.