SEWISE: SIGNED ERROR-WEIGHTED INFORMED SUBGROUP ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many machine learning classifiers can achieve good average accuracy but sometimes fail systematically on subgroups of their inputs. Subgroup discovery methods (SDMs) surface such underperforming groups by using knowledge about which samples were misclassified. As common SDMs use a binary indicator to represent an error, when a model leaves one subgroup under-detected and its complement over-detected, the two error types cancel. For example, on CheXpert, where a pleural effusion classifier's errors track patient sex, the two groups are far apart in signed error rate but nearly indistinguishable once direction is dropped. Additionally, the discovered errors are only part of a subgroup and therefore SDMs might miss the correctly predicted inputs. We present SEWISE, which represents each mistake as a *signed error score*, which is positive for false negatives and negative for false positives and is scaled by the model's confidence, and diffuses it over a nearest-neighbour graph constructed from the classifier's features to recover additional subgroup examples from their failing neighbours. Across ISIC-2017, CheXpert, CelebA and Waterbirds, SEWISE outperforms SPOTLIGHT, DOMINO and GEORGE on the majority of tasks. We further show that error-guided discovery may not be suitable when the classifier already has a good performance, as there is little error signal to utilise.

1 INTRODUCTION

Machine learning models may appear to be accurate on average but often fail systematically on a subgroup of their inputs. One such phenomenon is *shortcut learning*, where the model exploits features that are spuriously associated with the target label in training but are not semantically meaningful (Geirhos et al., 2020). For example, in chest radiography, classifiers trained to detect a collapsed lung have been shown to focus on the chest drain used to treat it, so they fail on precisely the patients who have not yet been treated (Oakden-Rayner et al., 2020). Furthermore, classifiers trained on sex-imbalanced radiograph datasets behave differently across patient sex (Larrazabal et al., 2020), and widely used chest radiograph models selectively underdiagnose under-served populations, reporting sick patients as healthy (Seyyed-Kalantari et al., 2021). In each case, the aggregate performance metric can appear to be good while a clinically important group is failing. When the underperforming subgroup is defined by a demographic label which is available (such as sex) the performance disparities between the demographic subgroups can be tested. But often these labels are unavailable, or the subgroup is defined by an unknown attribute. If such subgroups could be identified, a practitioner could seek to address the performance disparity by collecting more data for the underrepresented subgroup or by retraining with a group-robust objective (Sagawa et al., 2019).

Oakden-Rayner et al. (2020) call this problem *hidden stratification* and show that the affected subgroups tend to be the hardest to anticipate, being low in prevalence, weakly labelled, or separated only by a subtle visual cue. In the collapsed lung example, "patient without a chest drain" is meaningful to a radiologist but is nowhere recorded. In *subgroup discovery*, given a trained classifier and an evaluation set, we seek to surface the subgroup of samples on which the model systematically fails, without the knowledge of samples belonging to that subgroup.

Recent work has sought to address this task through the proposal of *subgroup discovery methods* (SDMs). GEORGE (Sohoni et al., 2020) clusters the classifier's features within each class. DOMINO (Eyuboglu et al., 2022) fits a mixture model over embeddings jointly with labels and predictions.

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.

Others search for high-loss regions of the representation (d'Eon et al., 2022), for directions separating correct from incorrect predictions (Jain et al., 2022), for attributes extracted from a robust model (Singla et al., 2021), or for concepts in a bottleneck layer(Konforti et al., 2026). However, these methods use a binary indicator per sample to determine whether the prediction was wrong. We argue that this answer is inadequate, for the following two reasons:

Issue 1: A single indicator of wrongness discards the direction of the error, and sometimes the direction is all there is. A classifier can fail on two subgroups in opposite ways, missing positives in one while reporting false positives in the other. This behaviour has also been reported for demographic subgroups in chest radiography (Seyyed-Kalantari et al., 2021; Larrazabal et al., 2020). Both groups have errors at comparable rates, and an indicator that records only *that* a sample was misclassified assigns them the same value; the one quantity that separates the groups is removed before any detector sees it.

Issue 2: Errors mark only part of the subgroup. A detector that ranks samples only by whether the model misclassified them recovers the misclassified fraction, not the subgroup.

To address these issues, we present SEWISE (Signed Error-Weighted Informed Subgroup Estimation), which addresses each of the above mentioned issues. **Against Issue 1**, SEWISE represents each mistake as a *signed error score*: a single confidence-weighted score that is positive on a false negative and negative on a false positive, so two subgroups failing in opposite directions are not collapsed into one value. **Against Issue 2**, SEWISE *diffuses* that signed score across a nearest-neighbour graph constructed from the classifier's features, so a correctly classified sample surrounded by (similar-looking) failing ones inherits their value and the subgroup is recovered beyond just the samples the model actually got wrong.

The work is organised as follows: We state the problem in Sec. 3.1, we present SEWISE in Sec. 3.2. In Sec. 5 we compare SEWISE with the DOMINO and GEORGE on a synthetic localised shortcut in dermatology images, a real demographic shortcut in chest radiographs, a natural attribute shortcut in face images, and a composited background shortcut in bird photographs. Across nine different experiments across four different tasks, SEWISE outperforms the considered baselines on eight.

Contributions.

In summary, in this paper we make two contributions to the field of SDMs:

1. **We analyse signed and pooled error representations.** Using known subgroup membership, labels and predictions, we examine when retaining error direction increases the difference between group means (Sec. 5.3).
2. We propose **SEWISE, a detector built on the signed representation.** We assign each model error a signed score, weighted by the classifier's confidence, and that score is diffused over a nearest-neighbour graph constructed from the classifier's features, so that correctly predicted subgroup members can be recovered from their failing neighbours. SEWISE outperforms SPOTLIGHT, DOMINO and GEORGE, by up to 15.4 AUC points, across different experiments.

2 RELATED WORK

Hidden subgroups and shortcut learning. A classifier that is accurate on average can still fail on a coherent subgroup of its inputs. Geirhos et al. (2020) frame this as *shortcut learning*: models prefer easily-learnt decision rules that succeed on the training distribution but rely on spuriously associated features, rather than causal ones. Oakden-Rayner et al. (2020) document the same phenomenon in medical imaging as *hidden stratification*, and show the affected subgroups are precisely those hardest to anticipate, so aggregate metrics conceal clinically meaningful failure. Therefore, in our work, we consider the scenario where the subgroup labels are unknown at evaluation time.

Subgroup discovery from model behaviour. SDMs automate the search for underperforming subgroups directly from a trained model, with no subgroup annotation. d'Eon et al. (2022) find contiguous regions of the final hidden layer on which loss is high, without clustering. Singla et al. (2021) extract interpretable features from a separate robust model and characterise failures by the attributes accompanying them. Jain et al. (2022) fit a linear decision boundary between correct and incorrect

predictions and read failure modes off as directions in latent space. DOMINO (Eyuboglu et al., 2022), our principal baseline, fits a Gaussian mixture model over cross-modal embeddings jointly with the model's predictions and true labels, and returns the components most associated with error. Importantly, because it models the label and the prediction as separate per-subgroup variables, it retains information about which *kind* of error a subgroup makes. More recently, Konforti et al. (2026) investigate subgroup discovery for concept bottleneck models, so discovered subgroups are explained by the concepts responsible for them. What separates SEWISE from these methods is therefore not that they discard the direction of an error but where that information sits. In each case, direction is a property of a *candidate subgroup*: a region or mixture component is proposed by a search over the feature space, and the errors falling inside it then describe it. Direction can only be attached to a subgroup the geometry has already isolated, and two subgroups failing in opposite ways become unrelated candidates with nothing holding them apart. SEWISE reverses the order. The signed error is a *field defined on the samples* and the grouping is produced by propagating it, which is what lets it recover subgroup members which the model did not misclassify, and what places oppositely-failing groups at opposite ends of a single score.

Subgroup discovery without error information. GEORGE (Sohoni et al., 2020) occupies the opposite end of the design space and is our second baseline. Within each class, GEORGE reduces activations with UMAP (McInnes et al., 2018), fits a Gaussian mixture model and then retrains the model with group distributionally robust optimisation (Sagawa et al., 2019). With regard to our work, only the discovery stage of GEORGE is comparable here, and it is a useful contrast precisely because it consults no error signal: a subgroup GEORGE recovers is one that is visually coherent, whereas a subgroup SEWISE recovers is one that is also defined by how the model fails.

Because GEORGE clusters strictly *within* each class (target label), it can only find a subgroup that is associated with one class, never one that cuts across classes, which is what the target subgroup is in all four of our experimental tasks. SEWISE scores every sample on a single scale, so it is not limited this way. Note that discovery is the first half of a longer pipeline: once a group is identified it can be used to reweight or retrain, as in Group DRO (Sagawa et al., 2019) or in methods that infer the group from a deliberately biased model's mistakes (Nam et al., 2020).

Directional error in medical imaging. Seyyed-Kalantari et al. (2021) show that chest radiograph classifiers selectively *underdiagnose* under-served populations i.e. the affected subgroup receives false negatives, not simply more errors. Larrazabal et al. (2020) show that sex imbalance in the training set alone is enough to produce systematically different behaviour across patient sex. Both describe error structure that a direction-blind approach cannot discover, which is the gap we study in this work.

3 METHODS

We first discuss the preliminaries for subgroup discovery methods and then describe our proposed approach, SEWISE in detail.

3.1 SUBGROUP DISCOVERY PRELIMINARIES

Problem Setting. Consider a binary classification setting with input $X \in \mathcal{X}$ and label $Y \in \{0, 1\}$. An auditor holds a trained classifier f and an evaluation set of n samples. For each sample i they observe the label y_i, the prediction $\hat{y}_i = f(x_i)$, the predicted probability $p_i = P_f(Y{=}1 \mid x_i)$, and an embedding $z_i \in \mathbb{R}^d$ produced either by the classifier's penultimate layer or by a frozen encoder of the same input.

A subgroup is a binary indicator $s_i \in \{0, 1\}$. The classifier *underperforms* on it with respect to a performance metric ℓ (higher is better) if $\mathbb{E}[\ell(f(X), Y) \mid s{=}1] < \mathbb{E}[\ell(f(X), Y) \mid s{=}0] - \epsilon$. The indicator is available to the evaluator but never to the method; it is what makes the problem well posed and what the method must recover.

The task. A subgroup discovery method is a function

$$\mathcal{D} : \big(\{z_i\}, \{y_i\}, \{\hat{y}_i\}, \{p_i\}\big)_{i=1}^{n} \longmapsto \{\sigma_i\}_{i=1}^{n}, \qquad \sigma_i \in \mathbb{R},$$

assigning each sample a score that ranks subgroup members above non-members. The three methods we compare share this signature and differ only in how the score, σ is computed.

The signed error. Comparing y_i with $\hat{y}_i$ places each sample in one of four confusion cells: true negative, false positive, false negative and true positive. Two summaries of that outcome will concern us. The *pooled* indicator e_i takes the value 1 when $\hat{y}_i \neq y_i$ and 0 otherwise, recording *that* a sample was misclassified. The *signed error*

$$\varsigma_i \;=\; \begin{cases} +1 & \text{if } y_i = 1 \text{ and } \hat{y}_i = 0 \quad \text{(a false negative)}, \\ -1 & \text{if } y_i = 0 \text{ and } \hat{y}_i = 1 \quad \text{(a false positive)}, \\ \;\;0 & \text{if the model was correct}, \end{cases}$$

additionally records *how* it was misclassified. The two are related by $e_i = |\varsigma_i|$, so pooling is exactly the operation that discards the sign.

Evaluation. Following Eyuboglu et al. (2022), the subgroup indicator is used only after a method has produced its scores. We report the AUC of σ against s on a held-out split, oriented so AUC $\geq$ 0.5.

3.2 SEWISE

Figure 1 summarises the complete SEWISE procedure.

Algorithm 1: SEWISE

Require: embeddings $z_{1:n}$, labels $y_{1:n}$, predictions $\hat{y}_{1:n}$, probabilities $p_{1:n}$; depth T. *The subgroup indicator s is not an input.*
Ensure: scores $\sigma_{1:n}$

1: $Z \leftarrow \textsc{Pca}_{128}(\textsc{Standardise}(z_{1:n}))$ *// shared feature space*
2: $\phi_i \leftarrow c_i\,\varsigma_i$ for all i *// signed score, Eq. 1*
3: $P \leftarrow \textsc{RowNormalise}(\textsc{Sym}(\textsc{kNNAffinity}(Z)))$ *// geometry only*
4: **repeat** T times: $\phi \leftarrow \lambda\phi + (1-\lambda)P\phi$ *// diffusion, Eq. 3*
5: **return** $\sigma \leftarrow \textsc{Rescale}_{[0,1]}(\phi)$

Figure 1: SEWISE in full. Line 2 is the fix for Issue 1, line 4 the fix for Issue 2. The graph in line 3 is built from feature distances alone and never consults the confusion outcome.

In this section, we describe the components of SEWISE as follows:

3.3 THE SIGNED ERROR SCORE

Recall from Sec. 3.1 that ς_i is $+1$ on a false negative, -1 on a false positive and 0 when the model was correct. SEWISE weights this by how confident the model was, since a confident mistake is stronger evidence of a systematic failure as compared to a borderline confidence. Writing $c_i = \max(p_i, 1-p_i) \in [\frac{1}{2}, 1]$ for that confidence, the *signed error score* of sample i is

$$\phi_i \;=\; c_i\,\varsigma_i. \tag{1}$$

The pooled error indicator, $e_i = |\varsigma_i|$, records whether a prediction is wrong but discards its direction. Groups with similar pooled error rates may therefore differ in their signed error rates. Section 5.3 examines these differences using known subgroup membership, while Sec. 5.1 measures the effect of retaining error direction on retrieval.

Which sign corresponds to the subgroup of interest is not assumed; it is resolved on validation data (Sec. 3.4.1). Eq. 1 reads only y_i, $\hat{y}_i$ and p_i, never the subgroup indicator.

3.4 DIFFUSION OVER FEATURE NEIGHBOURS

The score is non-zero only on samples the model got wrong, and a failing subgroup is rarely misclassified in its entirety. Ranking by Eq. 1 directly would therefore return the subgroup's misclassified

fraction and stop there, which is Issue 2. The members it misses are classified correctly but resemble those that were not, and that resemblance is the only available route to them.

SEWISE exploits this by letting the score spread between neighbours. Note that SEWISE operates not on the raw embeddings z_i introduced in Sec. 3.1 but on a standardised, reduced version of them, written Z: each feature dimension is rescaled to zero mean and unit variance, then projected onto its first 128 principal components. This rescaling process is important because everything downstream is based on distances between samples, and one large-magnitude coordinate would otherwise dominate these distances. Each sample i is joined to its k nearest neighbours in Z to form a graph, and each edge is given a weight that is largest for the closest neighbours and decays with distance,

$$W_{ij} = \begin{cases} \exp\!\left(-\|z_i - z_j\|^2 / h_i^2\right), & j \in \mathrm{kNN}(i), \\ 0, & \text{otherwise}, \end{cases} \qquad h_i \;=\; \frac{1}{k} \sum_{j \in \mathrm{kNN}(i)} \|z_i - z_j\|, \tag{2}$$

where the decay rate h_i is set *per sample* as the mean distance to that sample's own neighbours, so samples in sparse regions are not swamped by the denser population of correctly classified samples. Nearest-neighbour relations are not symmetric, so we symmetrise the weights, $A_{ij} = \frac{1}{2}(W_{ij} + W_{ji})$, giving each sample the neighbourhood $\mathcal{N}(i) = \mathrm{kNN}(i) \cup \{\, j : i \in \mathrm{kNN}(j) \,\}$, and then row-normalise, $P_{ij} = A_{ij} / \sum_l A_{il}$, so that $\sum_j P_{ij} = 1$. One diffusion step replaces each sample's value with a blend of its own and its neighbours',

$$\phi_i \;\longleftarrow\; \lambda\, \phi_i \;+\; (1 - \lambda) \sum_{j \in \mathcal{N}(i)} P_{ij}\, \phi_j, \tag{3}$$

where $\lambda \in [0, 1]$ is how much of its own value a sample keeps. Applying Eq. 3 T times and rescaling linearly to $[0, 1]$ yields the scores $\sigma_{1:n}$ that SEWISE returns. We fix $k = 15$ and $\lambda = 0.5$ throughout and never tune them; T is the only quantity selected per experiment.

Note that the graph is built from feature distances *only*: the confusion outcome enters through ϕ and never through the edges. Errors of the same kind are not declared adjacent, and two distant failures are not merged merely because they failed similarly; each propagates within its own neighbourhood, so a shortcut-induced failure and an unrelated hard case remain separable. Furthermore, there is no random initialisation and no iterative fitting, hence none of the run-to-run variance that mixture-model read-outs incurs, and we assume nothing about the shape of the failing region, unlike a mixture model, which presumes errors form a few blob-shaped clusters.

The depth T controls the reach of the diffusion process. A small T stays close to the observed errors and may under-recover the subgroup; a large one washes the score across the whole graph and dilutes it. Sec. 5.2 isolates how much of SEWISE's performance is attributable to this step.

3.4.1 SELECTING THE DIFFUSION DEPTH

T is chosen on a held-out validation portion, where the subgroup indicator s_i is used only to score candidates and never enters the computation of any candidate's error score ς_i. We sweep $T \in \{1, 2, 3, 5, 8, 12, 20, 30, 50\}$, take the mean validation AUC over that configuration's folds, and apply the argmax to the test split. The sign convention of Eq. 1 is fixed by the same validation data, so it too is never resolved against test labels. The selected diffusion depths for the CLIP experiments are listed in Table 2.

4 EXPERIMENTAL SETUP

Datasets and shortcuts. We evaluate subgroup retrieval on four datasets.

For **ISIC-2017** (Codella et al., 2018), we follow Boland et al. (2025) and insert a coloured square into 15% of positive images. We vary its placement between a fixed and a random location, and its side length between 30, 60 and 90 pixels. We train a separate ResNet-18 (He et al., 2016) for each of the six combinations using five-fold cross-validation. The target subgroup consists of images containing the square.

On **CheXpert** (Irvin et al., 2019), we train a ResNet-18 to predict pleural effusion on a sex-imbalanced training samples, which induces a *real, diffuse* demographic shortcut.

Table 1: **SEWISE effectively identifies subgroups.** Subgroup-retrieval AUC (%) in the **CLIP embedding** space, oriented so AUC ≥ 0.5, as mean $\pm$ standard deviation across five fold cross-validation. Bold marks experiments where SEWISE exceeds all three baselines.

Expt.	SEWISE	SPOTLIGHT	DOMINO	GEORGE
ISIC-2017				
const_p30	**83.4 ± 14.1**	68.0 ± 8.4	66.2 ± 2.0	55.7 ± 0.0
const_p60	**93.3 ± 8.1**	84.2 ± 15.4	77.2 ± 3.0	59.4 ± 0.0
const_p90	**97.3 ± 1.4**	96.0 ± 3.0	82.2 ± 1.4	56.3 ± 0.0
rand_p30	**85.2 ± 4.5**	73.9 ± 14.1	70.2 ± 6.4	54.6 ± 0.0
rand_p60	**96.6 ± 4.1**	93.9 ± 8.2	76.1 ± 2.2	54.6 ± 0.0
rand_p90	**99.5 ± 0.2**	98.8 ± 1.2	86.3 ± 3.4	65.1 ± 0.0
CheXpert				
pleural eff.	**69.6 ± 1.9**	54.8 ± 3.8	54.1 ± 0.6	53.5 ± 1.4
CelebA				
Male	61.8 ± 8.3	72.1 ± 5.0	83.9 ± 1.0	71.6 ± 4.4
Waterbirds				
wrong bg.	**78.7 ± 0.6**	67.4 ± 0.8	67.2 ± 1.9	76.4 ± 1.4

On **CelebA** (Liu et al., 2015), we follow Konforti et al. (2026), where a sex classifier uses hair-colour and makeup cues and fails on samples lacking the cue for their sex. This setting represents a *natural, visually coherent* shortcut.

For **Waterbirds** (Sagawa et al., 2019), we train a ResNet-18 to distinguish waterbirds from landbirds. The background agrees with the bird class in 95% of training images and 50% of evaluation images. The target subgroup includes waterbirds on land backgrounds and landbirds on water backgrounds.

Further dataset and split details are provided in Appendix A.2.

Embedding spaces. We evaluate all subgroup discovery methods using embeddings from a frozen CLIP encoder (Radford et al., 2021) and features from the penultimate layer of the trained ResNet-18 classifier. We report the CLIP results in Table 1 and the ResNet-18 results in Table 3.

Baselines. We compare SEWISE with SPOTLIGHT, DOMINO and GEORGE. For SPOTLIGHT (d'Eon et al., 2022), we fit a single high-loss region covering 5% of the validation split and apply it to the test split. For DOMINO (Eyuboglu et al., 2022), we fit a mixture model using embeddings, labels and model predictions, with five subgroups, 25 mixture components, 128 PCA components, diagonal covariance and likelihood weight $\gamma = 10$. For GEORGE (Sohoni et al., 2020), we cluster representations within each target class. Further implementation details are given in Appendix A.3.

Hyperparameter selection for DOMINO. We also evaluate DOMINO with $\gamma \in \{1, 2, 5, 10, 20, 40, 50, 100\}$. For each experiment, we select the value with the highest mean subgroup-retrieval AUC on the validation data and use it for test evaluation. Tables 1 and 3 report results with $\gamma = 10$, while Appendix A.5 reports results with the validation-selected values.

Metric. We measure subgroup retrieval using the area under the ROC curve (AUC) on the held-out test split, treating subgroup members as positive examples and all other samples as negative examples. We report AUC as a percentage, oriented so that AUC $\geq 50\%$, with the mean and standard deviation across folds.

5 RESULTS

We report subgroup-retrieval results using CLIP embeddings in Table 1. It can be seen that SEWISE achieves a higher mean AUC than all three baselines in the six ISIC-2017 experiments, as well as on CheXpert and Waterbirds. The largest increase over the strongest baseline is 15.4 percentage points on ISIC-2017 const_p30. On CheXpert, SEWISE achieves a mean AUC of 69.6%, compared with 54.8% for SPOTLIGHT, 54.1% for DOMINO and 53.5% for GEORGE. On Waterbirds, SEWISE achieves 78.7%, compared with 76.4% for the strongest baseline, GEORGE. Examples of the highest-ranked waterbirds are shown in Appendix A.1. The exception is CelebA, where SEWISE achieves 61.8%, below all three baselines. Appendix A.6 shows a visualisation of subgroup labels and method scores on t-SNE projections of CLIP features. Results based on features from the ResNet-18 classifier are presented in Appendix A.4.

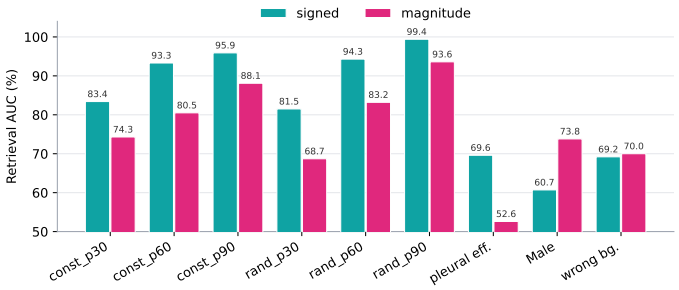


Figure 2: **The signed confidence is generally more effective than absolute confidence.** We use a fixed depth of 3 for the diffusion (kNN). Values are mean Retrieval AUC (%), mean across five fold cross-validation.

Table 2: Permutation and no-diffusion controls, in CLIP features. *signed* is SEWISE as reported in Table 1. *permuted* randomly reassigns the error scores across samples, keeping the embeddings, graph, diffusion settings and score distribution unchanged. Results are averaged over ten permutations per fold. *no diffusion* ranks by the raw signed score with no propagation. Retrieval AUC (%), mean over five folds (ISIC, CheXpert, Waterbirds) or three repeats (CelebA).

Expt.	T	signed	permuted	no diffusion
const_p30	3	83.4	56.6	65.3
const_p60	3	93.3	56.9	72.6
const_p90	8	97.3	65.7	74.6
rand_p30	5	85.2	60.1	63.7
rand_p60	8	96.6	63.9	72.7
rand_p90	30	99.5	80.4	76.7
pleural eff.	3	69.6	53.1	62.0
Male	12	61.8	58.7	52.1
wrong bg.	50	78.7	74.8	51.8
mean		85.0	63.3	65.7

5.1 EFFECT OF PRESERVING THE ERROR DIRECTION

We examine the effect of retaining error direction by comparing signed and unsigned error scores within SEWISE. We keep the features Z, neighbourhood graph P, confidence weights c_i and self-retention λ unchanged, and fix the diffusion depth at $T = 3$. We then use the following two scores:

$$\phi_i^{\text{signed}} = c_i\big(\mathbb{I}[\text{FN}_i] - \mathbb{I}[\text{FP}_i]\big), \qquad \phi_i^{\text{unsigned}} = c_i\big(\mathbb{I}[\text{FN}_i] + \mathbb{I}[\text{FP}_i]\big).$$

For both scores, we keep the confidence weighting, but the unsigned score treats false positives and false negatives alike.

It can be seen from Figure 2 that retaining error direction increases mean subgroup-retrieval AUC in seven of the nine experiments. The largest increase is 17.0 percentage points on CheXpert, while the increases in the six ISIC-2017 experiments range from 5.8 to 12.8 points. On CelebA and Waterbirds, the unsigned score achieves a higher mean AUC by 13.1 and 0.8 points, respectively. These results show that retaining error direction can improve retrieval, but is not beneficial for every target subgroup.

5.2 EFFECT OF SCORE ASSIGNMENT AND DIFFUSION

We analyse whether retrieval depends on assigning error scores to the corresponding samples and on propagating these scores through the neighbourhood graph. For the first control, we randomly permute the scores across samples while keeping the embeddings, graph and validation-selected diffusion depth unchanged. We average the results over ten permutations per fold. For the second control, we remove diffusion and rank samples directly by their signed, confidence-weighted error scores.

Table 2 shows that, on average across the experiments, mean subgroup-retrieval AUC is 85.0% for SEWISE, 63.3% with permuted scores and 65.7% without diffusion. The corresponding decreases are 21.7 and 19.3 percentage points. The decrease after permutation is less than five points on CelebA and Waterbirds, indicating that the contribution of the original score assignment varies across experiments. Overall, both controls reduce average retrieval AUC, supporting the use of sample-specific error scores together with diffusion.

5.3 ERROR DIRECTION WITHIN AND OUTSIDE THE SUBGROUP

We investigate how retaining error direction changes the difference in mean error scores between the target subgroup and the remaining samples. This analysis uses known subgroup membership to interpret the retrieval results.

Let $g_i \in \{0, 1\}$ indicate subgroup membership. Within each group g, let m_g and f_g denote the fractions of all samples that are false negatives and false positives, respectively. For this analysis, we use error indicators without confidence weighting or diffusion. The mean signed error indicator is $m_g - f_g$, while the mean unsigned error indicator is $m_g + f_g$. The absolute differences between the group means are therefore

$$\Delta_{\text{signed}} = |(m_1 - f_1) - (m_0 - f_0)|, \qquad \Delta_{\text{pooled}} = |(m_1 + f_1) - (m_0 + f_0)|. \tag{4}$$

Writing $a = m_1 - m_0$ and $b = f_1 - f_0$, these differences become $|a - b|$ and $|a + b|$. When a and b have opposite signs, retaining error direction gives a larger difference between the group means. For example, if $a = -b \neq 0$, the unsigned means are equal, while the signed means differ by $2|a|$. When a and b have the same sign, the unsigned mean difference is larger. If either is zero, the two differences are equal.

For the target subgroup on CheXpert, we obtain $\Delta_{\text{signed}} = 0.264$ and $\Delta_{\text{pooled}} = 0.040$. This larger signed mean difference is consistent with the improvement from retaining error direction in Figure 2. For the Waterbirds subgroup containing both waterbirds on land and landbirds on water, we obtain $\Delta_{\text{signed}} = 0.037$ and $\Delta_{\text{pooled}} = 0.221$. Here, combining false positives and false negatives gives a larger difference between the group means, and the unsigned variant achieves a slightly higher mean retrieval AUC at $T = 3$. Appendix A.7 shows the signed and pooled group-level differences for every experiment.

6 DISCUSSION AND LIMITATIONS

The CelebA results suggest that high classifier accuracy may limit the information available for error-guided subgroup discovery. However, the number of errors alone does not determine whether a subgroup can be recovered; their distribution within and outside the subgroup also matters. The present experiments do not isolate whether SEWISE's lower AUC on CelebA is due to the small number of errors, their distribution, or the relationship between the errors and the embedding space. Establishing when error-guided discovery becomes unreliable would require varying these factors separately.

We evaluate the methods over a small number of folds, which limits the precision of the performance estimates, particularly where variation across folds is large. We also use a single synthetic shortcut protocol, varying the size and placement of coloured squares, so generalisation to other synthetic cues is currently untested. Further experiments could vary the number of errors, their direction within and outside the subgroup, and the separation of the subgroup in the embedding space. Studying these factors separately would help establish when signed error scores improve retrieval and when the available error information is insufficient.

We introduced SEWISE to retrieve subgroups by propagating signed, confidence-weighted error scores through an embedding space. Our experiments show that retaining error direction can improve retrieval, but its benefit depends on how false positives and false negatives are distributed within and outside the target subgroup. The ablations also show that the correspondence between errors and samples, together with diffusion through neighbouring samples, contributes to retrieval. These findings support considering both error direction and sample similarity when designing subgroup discovery methods.

6.1 Reproducibility statement

The code will be made publicly available upon acceptance of the paper.

AI use statement

An LLM was used solely for refining text clarity and grammar. No AI tools were involved in research ideation, experimental design, or data analysis. The authors verify the accuracy of the final manuscript and assume full responsibility for its content. We did not use generative AI to design the experiments, run training or evaluation, or produce the reported results. The authors take responsibility for the final content of this work.

References

Christopher Boland, Sotirios Tsaftaris, and Sonia Dahdouh. Preventing shortcut learning in medical image analysis through intermediate layer knowledge distillation from specialist teachers. *arXiv preprint arXiv:2511.17421*, 2025.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1962–1981, 2022.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

Yael Konforti, Mateo Espinosa Zarlenga, Elaf Almahmoud, and Mateja Jamnik. Cb-slice: Concept-based interpretable error slice discovery. *arXiv preprint arXiv:2605.29836*, 2026.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12848–12857. IEEE, 2021.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Han Yu, Hao Zou, Jiashuo Liu, Renzhe Xu, Yue He, Xingxuan Zhang, and Peng Cui. Error slice discovery via manifold compactness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27925–27933, 2026.

A Appendix

A.1 Qualitative comparison of discovered subgroups

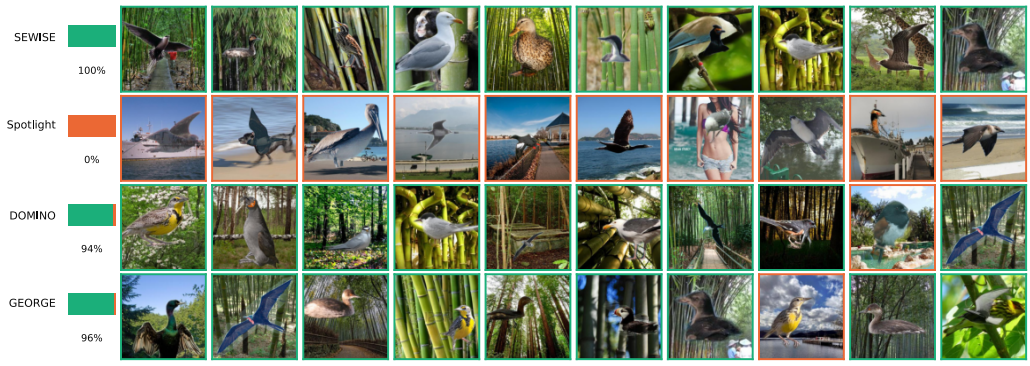


Figure 3: **Top-ranked images based CLIP features for Waterbirds.** Following Yu et al. (2026), we select each method's 50 highest-ranked waterbirds from one test split. The target is waterbirds on land, which make up 50% of the category. Bars show all 50 selections (green: land; orange: water). The ten images per method illustrate those selections and include a waterbird on water whenever the top 50 contain one.

Figure 3 shows ten examples from each method's top-ranked waterbirds, alongside the proportion on land among its top 50. All SEWISE examples are on land, as are most GEORGE and DOMINO examples; SPOTLIGHT's are on water. The corresponding top-50 proportions are 100%, 96%, 94% and 0%.

A.2 DATASETS, SHORTCUTS AND TRAINING PROTOCOLS

ISIC-2017 (Codella et al., 2018). A red-coloured square is injected into 15% of positive images following Boland et al. (2025), at two placements (constant location, random location) and three sizes (30, 60 and 90 pixels on a side, in a 224×224 image), giving six experimental setups. We write these as const_p30, const_p60, const_p90 for constant locations and rand_p30, rand_p60, rand_p90 for random locations, where the prefix is the placement and the number is the square's side length. A ResNet-18 (He et al., 2016) classifier is trained per setup over five-fold cross-validation. Each evaluation split consists of 251 samples.

CheXpert (Irvin et al., 2019). A ResNet-18 classifier is trained to predict pleural effusion on a sex-imbalanced training set, which induces a real, diffuse demographic shortcut between the label and patient sex. Evaluated over five stratified re-splits of a fixed 1000-sample set, 500 per split.

CelebA (Liu et al., 2015). Following the protocol in (Konforti et al., 2026), we use a sex classifier that relies on hair-colour and makeup cues and fails on samples lacking the cue for their sex. Three repeats, each with a validation and test split of 2000 samples. The classifier results in an overall accuracy of 0.983 on the test set with roughly 35 errors.

Waterbirds (Sagawa et al., 2019). A ResNet-18 bird classifier is trained on a set where the composited background agrees with the label on 95% of the 4795 images and is evaluated on a set balanced by background and labels. Following (Sagawa et al., 2019), we use a validationn set consisting of 1199 images which is also used for early stopping. For the cross-validation, we create five folds that are stratified halves of the test set consisting of 5794 images, resulting in 2897 images per fold. The trained classifier results in an overall accuracy of 87.0% but misclassifies 48.8% of waterbirds placed on land against 8.6% elsewhere.

A.3 BASELINE IMPLEMENTATIONS

We provide brief implementation and source details for the baseline subgroup discovery methods.

SPOTLIGHT (d'Eon et al., 2022) and **GEORGE** (Sohoni et al., 2020). We use the source code[1] with its defaults.

DOMINO (Eyuboglu et al., 2022). We use the official source code[2] and fit the approach with number of subgroups as 5 (n_subgroups=5), the number of mixture components as 25 (n_mixture_components=25), 128 components from the PCA projection of the representation n_pca_components=128 and the diagnonal covariance type (covariance_type='diag').

A.4 SUBGROUP DISCOVERY ON RESNET

We also evaluate subgroup discovery using features from the penultimate layer of the ResNet-18 classifier. Table 3 shows that SEWISE achieves the highest mean subgroup-retrieval AUC in six of the nine experiments, including four ISIC-2017 experiments, CheXpert and Waterbirds. On CheXpert, its mean AUC exceeds the strongest baseline, SPOTLIGHT, by 8.3 percentage points. On ISIC-2017 rand_p60 and rand_p90, SPOTLIGHT achieves a higher mean AUC by 0.7 and 0.3 percentage points, respectively. As with CLIP embeddings, DOMINO and GEORGE result in a higher mean AUC than SEWISE on CelebA. These results show that SEWISE generally performs well with features from the classifier itself, although its advantage over the baselines depends on the experiment.

Table 3: **SEWISE effectively identifies subgroups.** Subgroup-retrieval AUC (%) in the **ResNet-18** based embedding space, oriented so AUC ≥ 0.5, as mean $\pm$ standard deviation across five-fold cross validation. Bold marks experiments where SEWISE exceeds all three baselines.

Expt.	SEWISE	SPOTLIGHT	DOMINO	GEORGE
ISIC-2017				
const_p30	**77.0 ± 15.0**	73.6 ± 17.2	72.7 ± 11.1	56.5 ±6.5
const_p60	**96.9 ± 3.1**	96.2 ± 6.3	93.0 ± 9.6	61.2 ± 5.9
const_p90	**99.5 ± 1.0**	96.3 ± 6.9	97.5 ± 3.5	72.0 ± 7.6
rand_p30	**78.1 ± 12.1**	72.0 ± 14.0	75.6 ± 13.4	57.4 ± 6.0
rand_p60	98.7 ± 1.3	99.4 ± 1.4	95.6 ± 4.4	67.6 ± 6.3
rand_p90	99.6 ± 0.2	99.9 ± 0.1	99.5 ± 0.6	68.2 ± 5.3
CheXpert				
pleural eff.	**77.0 ± 1.5**	68.7 ± 5.0	59.4 ± 2.3	60.6 ± 3.4
CelebA				
Male	81.3 ± 4.6	57.8 ± 10.1	89.1 ± 1.3	84.7 ± 0.7
Waterbirds				
wrong bg.	**75.8 ± 0.5**	69.9 ± 0.6	72.4 ± 1.6	73.9 ± 0.6

Table 4: DOMINO subgroup-retrieval AUC (%, mean $\pm$ sample standard deviation) with fixed $\gamma = 10$ and with γ selected using the validation split.

		Fixed	Validation-selected	
Expt.	Emb.	AUC ($\gamma = 10$)	γ	AUC
ISIC-2017				
const_p30	CLIP	66.2 ± 2.0	2	68.6 ± 0.4
const_p30	ResNet	72.7 ± 11.1	5	71.9 ± 13.8
const_p60	CLIP	77.2 ± 3.0	10	77.2 ± 3.0
const_p60	ResNet	93.0 ± 9.6	2	96.7 ± 5.8
const_p90	CLIP	82.2 ± 1.4	10	82.2 ± 1.4
const_p90	ResNet	97.5 ± 3.5	2	98.7 ± 3.0
rand_p30	CLIP	70.2 ± 6.4	1	73.0 ± 5.1
rand_p30	ResNet	75.6 ± 13.4	5	77.6 ± 13.0
rand_p60	CLIP	76.1 ± 2.2	2	75.5 ± 3.6
rand_p60	ResNet	95.6 ± 4.4	2	97.1 ± 2.9
rand_p90	CLIP	86.3 ± 3.4	2	88.9 ± 1.5
rand_p90	ResNet	99.5 ± 0.6	1	99.4 ± 0.9
CheXpert				
pleural eff.	CLIP	54.1 ± 0.6	20	59.0 ± 3.3
pleural eff.	ResNet	59.4 ± 2.3	5	67.4 ± 1.4
CelebA				
Male	CLIP	83.9 ± 1.0	10	83.9 ± 0.9
Male	ResNet	89.1 ± 1.3	10	89.1 ± 1.3
Waterbirds				
wrong bg.	CLIP	67.2 ± 1.9	50	72.8 ± 1.0
wrong bg.	ResNet	72.4 ± 1.6	50	75.6 ± 0.9

A.5 LIKELIHOOD WEIGHT ABLATION FOR DOMINO

Table 4 shows a comparison of subgroup discovery using DOMINO when the default likelihood weight (γ) value of 10 is used and when an optimal value of γ is inferred based on the held-out validation set. Specifically, we use the following values of γ: $\{1, 2, 5, 10, 20, 40, 50, 100\}$.

[1]https://github.com/yaelkon/CB-SLICE/tree/main/baselines

[2]https://github.com/HazyResearch/domino

It can be seen from the results that choosing γ on the held-out validation set improves the meansubgroup retrieval AUC DOMINO in 11 and lowers the mean AUC in 3 of the 18 dataset, subgroup, and embedding comparisons. Averaged across all 18 comparisons, the increase is 2.0 percentage points; the largest increase is 8.0 points on CheXpert with ResNet features. Overall, SEWISE generally achieves a higher mean AUC than DOMINO, even after γ is selected using these validation sets. The exceptions are the two CelebA comparisons, where DOMINO results in a higher mean AUC.



Figure 4: **Subgroup membership and method scores on t-SNE projections of CLIP features.** Each row shows one test split on a shared layout: orange points belong to the target subgroup, and darker blue indicates higher method scores. The results are based on one cross validation fold. However, the trends are similar across folds.

A.6 SUBGROUP SCORES IN A T-SNE PROJECTION

Figure 4 shows subgroup membership and each method's scores on the same t-SNE layout within each dataset. Subgroup members appear more concentrated in parts of the CelebA and Waterbirds maps and more dispersed in CheXpert; ISIC lies between these patterns. On the considered CheXpert fold, SEWISE achieves the highest AUC despite the absence of an obvious isolated subgroup in this projection.

A.7 ERROR DIRECTION IN SUBGROUPS

Signed errors can distinguish two groups that make different kinds of mistakes, even when their overall error rates are similar.To examine this, we use known subgroup labels to calculate the fractions of samples within each group that are false negatives and false positives. Their difference gives the mean signed error, and their sum gives the pooled error rate. Figure 5 shows these signed rates in panel (a) and the absolute difference between the groups under each measure in panel (b). On

CheXpert, the target subgroup is false-negative-dominated ($+0.126$), while its complement is false-positive-dominated (-0.138). Their signed rates differ by 0.264, but their pooled error rates differ by only 0.040.

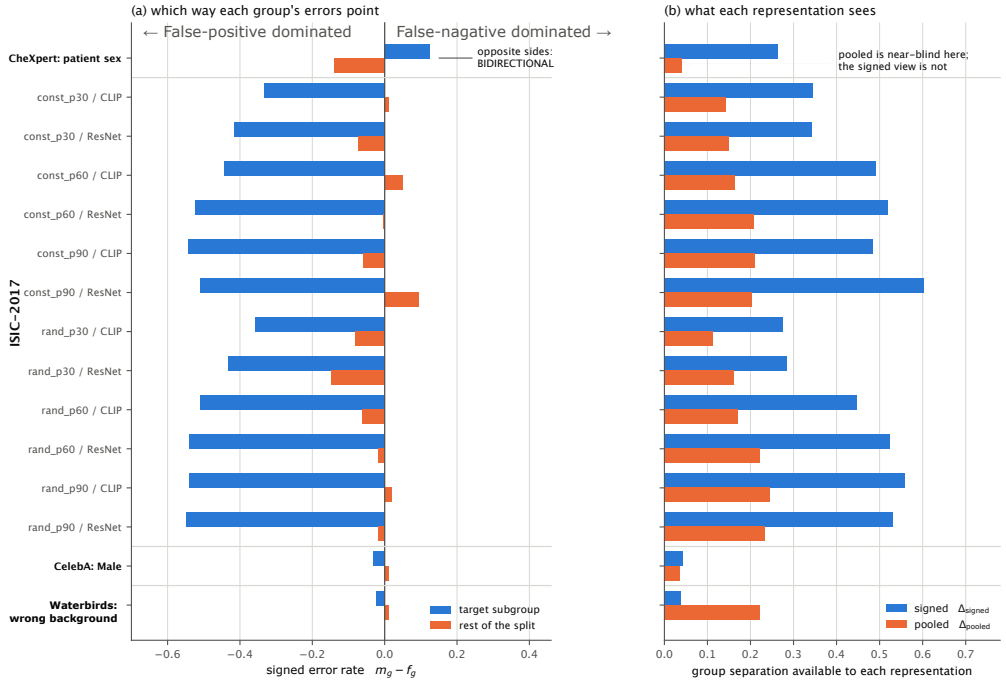


Figure 5: **Error direction and available separation on every experiment.** Panel (a) shows the signed error rate $m_g - f_g$ of the target subgroup and of the rest of the split; negative is false-positive-dominated, positive is false-negative-dominated. Panel (b) shows the separation each representation can see. The plotted rates are computed from subgroup membership, true labels and predicted labels. For CheXpert, CelebA and Waterbirds, the two embedding variants share the classifier's predictions exactly and appear once, while ISIC's variants are based on separate training runs and are shown separately.