Uncovering Latent Subgroups: Spectral Clustering for Fairness Analysis in Contrastive Embeddings

Hridoy Rahman[†,*], Blessing Ogbuokiri[‡]
[†] Responsible and Applied Machine Learning Laboratory (RAML Lab),
Department of Computer Science, Brock University

Abstract

Contrastive learning enables scalable representation learning in computer vision and healthcare, yet embedding spaces may encode unequal geometric structure across latent subpopulations, leading to downstream performance disparities. Conventional fairness audits relying on demographic labels often fail to detect such structural bias. This work examines whether contrastive embeddings contain fairness relevant latent subgroups that can be identified without demographic supervision. We introduce a label free spectral fairness audit that constructs similarity graphs over CLIP embeddings and applies eigengap-based spectral clustering. Experiments on CheXpert reveal stable latent subgroups with noticeable geometric distortions and performance gaps, exposing hidden fairness risks missed by demographic based evaluations. This work enables label-free discovery of hidden fairness and reliability risks in contrastive embeddings, supporting safer, more transparent deployment of foundation models in healthcare and other high-stakes domain.

Keywords: Contrastive Learning, Embeddings, Latent Subgroups, AI Fairness, Spectral Clustering, Bias.

1. Introduction

As contrastive learning drives large-scale vision models, concerns arise that representation geometry encodes latent biases, which correlate with downstream performance disparities, suggesting contrastive objectives may propagate structural bias [1]. Methods that explicitly reshape embedding geometry during training can reduce demographic information without sacrificing accuracy [2]. However, subsequent work demonstrates that subgroup performance gaps can persist even in debiased embedding spaces [3]. These findings highlight the need to directly analyze latent geometric structure rather than relying solely on demographic labels or output level fairness metrics. Prior fairness studies in chest X-ray classification primarily assess prediction level disparities using explicit demographic attributes via metrics such as equalized odds, demographic parity, and subgroup AUROC gaps as exemplified by CheXclusion [4] and FairBias [5], both of them assume reliable demographic labels and do not analyze contrastive embedding geometry or uncover latent subgroups without demographics.

Motivated by these gaps, we investigate the fairness properties of contrastive embeddings through their latent geometric structure and address the following research questions:

- RQ1: Do contrastive embeddings contain latent geometric subgroups that capture fairness-relevant variation without relying on demographic labels?
- RQ2: Can spectral graph methods reliably identify these subgroups and characterize geometric distortions within the embedding space?
- RQ3: Do the discovered latent subgroups exhibit systematic downstream performance disparities that are not detectable through standard aggregate evaluation metrics?

[*] hr21it@brocku.ca

We propose a label-free spectral fairness audit of CLIP embeddings to uncover latent subgroup disparities beyond demographics. Code is available at: Label-free Spectral Fairness Code.

2. Related Work

Fairness research is shifting beyond demographics, using reweighting and distributionally robust methods to mitigate disparities without labels, supported by survey work emphasizing label-independent approaches [6–8].

More recently, fairness research has adopted a geometric perspective, framing representation bias in terms of embedding distances, manifolds, and structural distortions [9]. This view emphasizes representation geometry as a primary mechanism through which disparities emerge, even absent demographic labels. Complementary self-supervised approaches show that directly shaping embedding structure via conceptual constraints can yield fairer representations without demographic supervision [10]. Prior work on large-scale contrastive models shows that social biases embedded during pretraining manifest as downstream performance disparities, as demonstrated for CLIP and other vision–language encoders [1, 11]. While some methods attempt to mitigate bias by modifying representation geometry [2], empirical evidence suggests that removing demographic separability alone does not eliminate subgroup performance gaps [3]. In medical imaging, fairness studies primarily analyze prediction-level disparities using explicit demographic labels [4, 5], without examining latent embedding geometry. This leaves representational biases and unlabeled subgroup structure largely unaddressed. Motivated by the limits of demographic-based fairness evaluation and the role of representation geometry, we examine whether contrastive embeddings contain latent subgroups, if spectral methods can uncover them without demographics, and whether they exhibit hidden performance disparities.

3. Methodology

We propose a label-free spectral auditing method (Fig. 1) to identify latent subgroups in embeddings and assess fairness via geometry, without demographics. Experimental hardware setup details shown in Table 9, Appendix A.

Embedding Extraction and Preprocessing: CLIP's contrastive embeddings capture semantic structure and intrinsic biases, making them suitable for label-free geometric fairness analysis. Therefore, visual representations are extracted using the CLIP ViT-B/32 encoder [12] applied to the CheXpert dataset. All embeddings are ℓ_2-normalized to ensure numerical stability. To reduce noise and improve spectral conditioning, we optionally apply Principal Component Analysis (PCA), retaining $d = 64$ dimensions to stabilize kNN graph construction and spectral decomposition. The resulting embedding matrix $X \in \mathbb{R}^{n \times d}$ is used for graph construction.

Similarity Graph Construction: Local geometric structure is captured using a k-nearest neighbor (kNN) graph constructed via the HNSW approximate nearest neighbor algorithm. Edges are defined using cosine similarity with binary weighting, and an optional mutual kNN constraint is applied to improve robustness. This yields a sparse, symmetric affinity matrix W emphasizing reliable local neighborhoods.

Spectral Graph Construction and Cluster Selection: From W, we compute the symmetric normalized graph Laplacian

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2},$$

where D is the degree matrix. The number of latent subgroups is selected using the eigengap heuristic, computed from a single partial eigendecomposition of the Laplacian. Specifically,

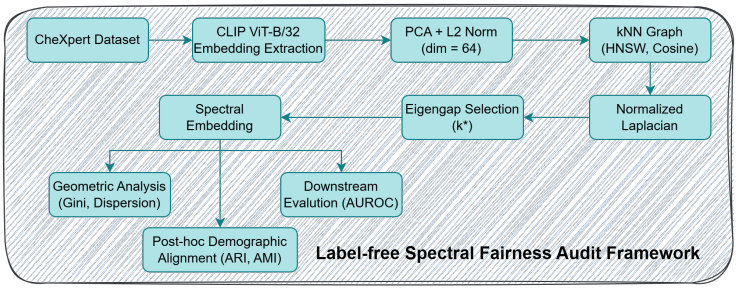


Figure 1. Label-free Spectral Fairness Audit Framework

we compute the smallest $K_{\max}+1$ eigenvalues and select

$$k^* = \arg\max_{k\in[2,K_{\max}]} (\lambda_k - \lambda_{k-1}).$$

Spectral Embedding and Clustering: Using the selected k^*, we form a spectral embedding from the corresponding eigenvectors and apply row normalization. Latent subgroups assignments are obtained via k-means clustering in this space, using multiple random initializations to ensure stability.

Geometric and Fairness Analysis: We quantify latent subgroups geometry using cluster size distributions, Gini coefficient, intra-cluster radii, inter-centroid distances, and a dispersion index. Fairness is evaluated without demographic supervision by training logistic regression classifiers for Pleural Effusion and Cardiomegaly and reporting AUROC (Area Under the Receiver Operating Characteristic curve) per cluster, with uncertainty estimated via 500 sample bootstrap resampling.

Baselines, Controls, and Robustness: We compare against k-means and random clustering, including shuffle-based negative control, and assess robustness via k-sensitivity ($k \in 10, 15, 20$).

Post-hoc Demographic Alignment and Stability: Demographic labels (sex) are used only post-hoc via ARI/AMI (Adjusted Rand Index / Adjusted Mutual Information), with stability measured across repeated initializations.

4. **Results**

We evaluate the proposed spectral fairness audit (Figure 1) on CLIP [12] embeddings from CheXpert across Pleural Effusion, Edema and Cardiomegaly which are selected as clinically distinct that share the same representation in the embedding space, reporting geometric, stability, demographic alignment, and downstream performance results (Tables 1–7).

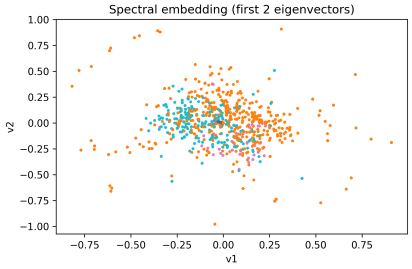
Using the eigengap criterion, the proposed method identifies $k^* = 9$–11 latent subgroups for Pleural Effusion (PE) across kNN graph sizes $k \in \{10, 15, 20\}$ (Table 1). Under the default configuration ($k = 15$), Cardiomegaly and Edema similarly yield $k^* = 10$ and $k^* = 11$ latent subgroups, respectively, demonstrating that the discovered subgroup structure generalizes beyond the primary task. Spectral analysis reveals clear latent geometric subgroup structure in CLIP contrastive embeddings without demographic labels. As shown in Fig. 2 and Table 2, the spectral embedding reveals a dense central manifold with peripheral samples, while the discovered clusters exhibit substantial size imbalance ($Gini = 0.54 - 0.57$ for PE, 0.49 for Edema, and 0.84 for Cardiomegaly), indicating meaningful structure rather than random partitioning. Inter-centroid distances ($0.63 - 0.76$) further suggest subgroup separation in the embedding space. Importantly, alignment with available demographic attributes is negligible (ARI/AMI ≈ 0), suggesting that the identified subgroups reflect latent distributional structure rather than trivial recovery of known labels which supports the existence of fairness-relevant geometric variation in contrastive embeddings.

Component	Value
Dataset / Encoder	CheXpert (n=223,414) / CLIP ViT-B/32
Embedding	512 → 64 (PCA)
Graph	HNSW (cosine), k=10, 15, 20; mutual kNN: Yes; edge weight: Binary
Laplacian	Symmetric normalized
Clusters (k^*)	9–11 (eigengap)
Metrics (Silhouette / DBI)	0.54–0.65 / 0.53–0.79
Stability (ARI)	> 0.95

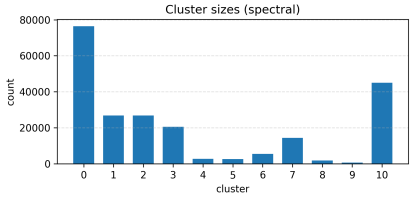
Table 1. Experimental Setup and Spectral Clustering Summary Across k-Sensitivity

Metric	Range
Cluster size Gini	0.54 – 0.57 (PE), 0.49 (Edema), 0.84 (Cardiomegaly)
Mean cluster radius	0.82 – 0.92
Inter-centroid distance	0.63 – 0.76
Dispersion index	0.73 – 0.92 (PE), 0.79 (Edema), 0.47 (Cardiomegaly)
Demographic alignment (ARI / AMI)	≈ 0 / ≈ 0

Table 2. Geometric Properties of Discovered Latent Subgroups for Pleural Effusion(PE), All $k = 10, 15, 20$, k=15 for Edema and Cardiomegaly



(a) Spectral embedding (first two eigenvectors) for k = 10.



(b) Cluster size distribution for k = 10.

Figure 2. Latent geometric subgroups revealed by spectral clustering for Pleural Effusion

Spectral clustering consistently outperforms k-means and random partitioning in uncovering geometrically meaningful and stable subgroups (Tables 3 and 5, Appendix A). For Pleural Effusion, spectral clustering exhibits higher cluster size imbalance ($Gini \approx 0.55$–0.57) and lower dispersion indices (0.73-0.93) than k-means ($Gini \approx 0.11$, $dispersion \approx 0.98$), while random clustering exhibits negligible structure (Tables 2 and 3, Appendix A). As shown in Fig. 2, spectral embeddings exhibit dense central clusters and sparse peripheral subgroups, indicating uneven latent geometry. Sensitivity analysis across neighborhood sizes (Table 5, Appendix A) shows that clustering quality remains stable, with silhouette scores between 0.60-0.65, Davies-Bouldin indices between 0.53-0.79, indicating changing neighborhood size does not collapse the cluster structure, and ARI consistently above 0.95, indicating latent subgroups structure is consistent. In contrast, the shuffle control produces inconsistent and task-dependent clustering quality, indicating that stable latent subgroups structure does not emerge when embedding geometry is disrupted (Table 8, Appendix A). These results suggests, spectral graph methods reliably identify latent subgroups and characterize geometric distortions that remain stable across neighborhood choices, with shuffle control confirming

that the structure arises from intrinsic embedding geometry rather than hyperparameter artifacts.

Although aggregate AUROC remains largely stable across clustering configurations (Table 5, Appendix A), latent subgroups identified by spectral clustering exhibit substantially larger downstream performance disparities than those revealed by k-means or random clustering baselines. For Pleural Effusion, spectral clustering produces AUROC ranges of 0.12–0.13 across subgroups, compared to 0.06–0.10 for k-means and ≤ 0.03 for random clustering (Table 4, Appendix A). Similar patterns emerge across conditions: Cardiomegaly shows extreme per-cluster AUROC variation under spectral clustering (range ≈ 0.41), while Edema exhibits non-trivial subgroup heterogeneity under both spectral and k-means methods but not under random assignment. These findings indicate that standard aggregate evaluation masks systematic subgroup-level performance variation that becomes visible only when auditing embeddings through latent geometric structure. While k-means occasionally identifies subgroup performance variation, these effects are less consistent and accompanied by near-uniform cluster geometry, whereas spectral clustering uniquely surfaces both geometric imbalance and amplified worst-case performance disparities absent under random clustering.

Overall, our results show that CLIP embeddings encode stable latent geometric subgroups independent of demographic attributes, exhibiting imbalance and task-dependent performance differences. By auditing representation geometry rather than demographics or aggregate accuracy, the proposed spectral approach reveals latent reliability and fairness risks not consistently detected by k-means or random baselines.

5. Discussion

RQ1 - Latent subgroup existence without demographics: Prior demographic-free fairness work has largely focused on optimization-level interventions, without examining whether representation spaces themselves encode latent subgroups structure [6–8]. Our results show that CLIP contrastive embeddings contain stable latent geometric subgroups independent of demographic labels, characterized by substantial cluster imbalance and coherent geometry, supporting geometric perspectives on fairness [9]. *RQ2 - Reliability of spectral methods:* Spectral clustering consistently outperforms k-means and random baselines in capturing uneven manifold structure and stable subgroup partitions, highlighting the importance of manifold-aware methods for auditing representation spaces beyond enforcing demographic invariance [2, 3]. *RQ3 - Downstream disparities beyond aggregate evaluation:* Despite weak demographic alignment, spectrally identified subgroups exhibit task-dependent performance heterogeneity that is largely invisible under aggregate AUROC; while k-means occasionally surfaces variation, these effects are less consistent and lack corresponding geometric structure. Rather than constituting fairness violations per se, such disparities signal reliability and fairness-relevant risks that complement prediction-level audits in domains with limited demographic supervision, including medical imaging [4, 5].

6. Conclusion and Future work

This work introduces a label-free method for auditing fairness in contrastive embeddings. Through spectral analysis of CLIP representations, we show that latent geometric subgroups emerge independently of demographic supervision, exhibit systematic geometric distortions, and correspond to substantial downstream performance disparities. These findings demonstrate that conventional demographic audits can overlook hidden risks, while geometric aware spectral analysis provides a necessary complementary lens for fairness evaluation in large-scale representation learning. Future work will focus on improving scalability through approximate spectral methods, extending the method to multimodal and temporal embeddings, and integrating causal and clinical validation to support geometric aware fairness

mitigation. Spectral decomposition is computationally demanding at scale, latent clusters may lack semantic interpretability, and downstream evaluation is limited to a linear classifier under incomplete demographic labels.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2026-05435 and DGECR-2026-00353]. The authors also acknowledge the support of the Responsible and Applied Machine Learning Lab (RAML Lab) at Brock University's Department of Computer Science for providing a collaborative research environment and resources that contributed to this work.

Declaration of AI Use

The authors used ChatGPT to improve language and readability. All content was subsequently reviewed and edited, and the authors take full responsibility. No AI tools were used to generate scientific content, data, analysis, or conclusions.

References

[1] K. Ghate, I. Slaughter, K. Wilson, M. Diab, and A. Caliskan. "Intrinsic bias is predicted by pretraining data and correlates with downstream performance in vision-language encoders". In: *arXiv preprint arXiv:2502.07957* (2025).

[2] A. Shen, X. Han, T. Cohn, T. Baldwin, and L. Frermann. "Contrastive learning for fair representations". In: *arXiv preprint arXiv:2109.10645* (2021).

[3] A. Shen, X. Han, T. Cohn, T. Baldwin, and L. Frermann. "Does representational fairness imply empirical fairness?" In: *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. 2022, pp. 81–95.

[4] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. World Scientific. 2020, pp. 232–243.

[5] S. Iqbal, X. Zhong, M. A. Khan, Z. Wu, N. A. Almujally, W. Liu, and A. Hussain. "FairBias: Mitigating Bias in Medical Image Diagnosis with Mixed Noise and Class Imbalance". In: *Neurocomputing* (2025), p. 130910.

[6] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi. "Fairness without demographics through adversarially reweighted learning". In: *Advances in neural information processing systems* 33 (2020), pp. 728–740.

[7] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. "Fairness without demographics in repeated loss minimization". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1929–1938.

[8] C. Ashurst and A. Weller. "Fairness without demographic data: A survey of approaches". In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2023, pp. 1–12.

[9] A. Maggio, L. Giuliani, R. Calegari, M. Lombardi, M. Milano, et al. "A geometric framework for fairness". In: *CEUR WORKSHOP PROCEEDINGS*. Vol. 3523. CEUR-WS. 2023, pp. 1–17.

[10] J. Chai and X. Wang. "Self-supervised fair representation learning without demographics". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27100–27113.

[11] S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, and M. Brundage. "Evaluating clip: towards characterization of broader capabilities and downstream implications". In: *arXiv preprint arXiv:2108.02818* (2021).

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. Preprint / OpenAI CLIP ViT-B/32 model. 2021. URL: https://huggingface.co/openai/clip-vit-base-patch32.

Appendix A. Baseline K-means vs Spectral Clustering vs Random Clustering

A.1. Geometry and stability comparison for Pleural Effusion

Run	Method	#Clusters	Cluster Size Gini ↑	Dispersion Index ↓
K=10	k-means (Baseline) ($k = 11$)	11	0.109	0.975
	Spectral ($k^* = 11$)	11	**0.557**	**0.747**
	Random ($k = 11$)	11	0.004	0.010
K=15	k-means (Baseline) ($k = 9$)	9	0.138	0.979
	Spectral ($k^* = 9$)	9	**0.569**	**0.730**
	Random ($k = 9$)	9	0.003	0.009
K=20	k-means (Baseline) ($k = 11$)	11	0.109	0.975
	Spectral ($k^* = 11$)	11	**0.547**	**0.925**
	Random ($k = 11$)	11	0.004	0.010

Table 3. Subgroup geometry and stability comparison across spectral clustering, k-means, and random partitioning. Higher cluster size Gini indicates greater size imbalance; lower dispersion index indicates tighter cluster cohesion.

A.2. Downstream AUROC Disparities for Pleural Effusion

Run	Method	$\text{AUROC}_{\min}$	$\text{AUROC}_{\max}$	Range (max–min)
K=10	k-means (Baseline) ($k = 11$)	0.658	0.723	0.065
	Spectral ($k^* = 11$)	0.630	0.756	**0.126**
	Random ($k = 11$)	0.697	0.729	0.032
K=15	k-means (Baseline) ($k = 9$)	0.646	0.741	0.095
	Spectral ($k^* = 9$)	0.604	0.731	**0.127**
	Random ($k = 9$)	0.704	0.719	0.015
K=20	k-means (Baseline) ($k = 11$)	0.628	0.742	**0.115**
	Spectral ($k^* = 11$)	0.662	0.728	0.066
	Random ($k = 11$)	0.702	0.730	0.027

Table 4. Downstream AUROC heterogeneity across discovered subgroups. Disparity is measured as the range (max–min) of cluster-level AUROC means.

A.3. Sensitivity Analysis of Spectral Clustering across Neighborhood Sizes

Sensitivity Run	Spectral k^*	Silhouette ↑	Davies–Bouldin ↓	Stability (ARI) ↑	AUROC ↑
$k = 10$	11	0.6032	0.7882	0.9750	**0.7170**
$k = 15$	9	**0.6486**	**0.5332**	**0.9811**	0.7157
$k = 20$	11	0.5386	0.5521	0.9587	0.7116

Table 5. Sensitivity analysis of spectral clustering across neighborhood sizes. The $k = 15$ configuration achieves the best balance of geometric separation (highest silhouette, lowest Davies–Bouldin) and clustering stability, while downstream AUROC remains relatively stable across settings, indicating robustness of predictive performance to clustering hyperparameters.

A.4. **Cardiomegaly — Baseline k-means vs Spectral vs Random**

Method	#clust	Silh ↑	DB ↓	Gini ↓	ARI ↑	AUROC ↑	Per Cluster AUROC min–max	Range
Spectral	10	0.9868	1.0946	0.8487	0.9999	0.6566	0.2697–0.6777*	0.4080*
k-means	10	–	–	0.1046	–	0.6538	0.6296–0.6752	0.0457
Random	10	–	–	0.0027	–	0.6509	0.6313–0.6670	0.0357

Table 6. Cardiomegaly: comparison of spectral clustering against k-means and a random assignment baseline. spectral clustering discovers highly skewed subgroup structure (Gini=0.8487) and wider per-cluster AUROC variation, but this spread is likely inflated by very small clusters; k-means produces more balanced partitions with tighter AUROC dispersion. Note: Spectral per-cluster AUROC is only reported for a subset of clusters in the provided output and includes very small clusters (e.g., $n \approx 31$), which can yield unstable AUROC estimates.

A.5. **Edema — Baseline k-means vs Spectral vs Random**

Method	#clust	Silh ↑	DB ↓	Gini ↓	ARI ↑	AUROC ↑	Per Cluster AUROC min–max	Range
Spectral	11	0.5386	0.5866	0.4943	0.8670	0.7519	0.6547–0.7504	0.0957
k-means	11	–	–	0.1090	–	0.7512	0.6184–0.7539	0.1355
Random	11	–	–	0.0040	–	0.7521	0.7380–0.7630	0.0250

Table 7. Edema: comparison of spectral clustering against k-means and a random assignment baseline. Unlike the random control (low Gini, narrow AUROC spread), both spectral and k-means identify subgroups with non-trivial per-cluster AUROC variation, with k-means showing the largest spread in this run.

A.6. **Shuffle Negative Control across All Pathologies**

Task (kNN k)	k^*	**Silhouette**	**DBI**
Pleural Effusion ($k = 10$)	12	0.696	0.993
Pleural Effusion ($k = 15$)	9	0.663	0.431
Pleural Effusion ($k = 20$)	11	0.540	0.552
Cardiomegaly ($k = 15$)	9	0.998	0.987
Edema ($k = 15$)	11	0.544	0.579

Table 8. Shuffle-based negative control results across pathologies and kNN graph sizes.

A.7. **Experimental Hardware Setup**

Component	**Specification**
CPU	Intel Xeon @ 2.00GHz (4C/8T)
GPU	NVIDIA Tesla T4 (15 GB VRAM)
CUDA	13.0
Environment	Google Colab (KVM)

Table 9. System Specifications