$\mathcal{S}$tride-Net: Fairness-Aware Disentangled Representation Learning for Chest X-Ray Diagnosis

Darakshan Rashid[1], Raza Imam[2], Dwarikanath Mahapatra[3], Brejesh Lall[1]
[1]Indian Institute of Technology Delhi, [2]MBZUAI, [3]Khalifa University
bsz228540@iitd.ac.in, raza.imam@mbzuai.ac.ae, dwarikanath.mahapatra@ku.ac.ae, brejesh.lall@ee.iitd.ac.in

***Abstract*—Deep neural networks for chest X-ray classification achieve strong average performance, yet often underperform for specific demographic subgroups, raising critical concerns about clinical safety and equity. Existing debiasing methods frequently yield inconsistent improvements across datasets or attain fairness by degrading overall diagnostic utility, treating fairness as a post hoc constraint rather than a property of the learned representation. In this work, we propose $\mathcal{S}$tride-Net ($\mathcal{S}$ensitive a$\mathcal{T}$tribute $\mathcal{R}$esilient learning v$\mathcal{I}$a $\mathcal{D}$isentanglement and learnable masking with $\mathcal{E}$mbedding alignment), a fairness-aware framework that learns disease-discriminative yet demographically invariant representations for chest X-ray analysis. $\mathcal{S}$tride-Net operates at the patch level, using a learnable stride-based mask to select label-aligned image regions while suppressing sensitive attribute information through adversarial confusion loss. To anchor representations in clinical semantics and discourage shortcut learning, we further enforce semantic alignment between image features and BioBERT-based disease label embeddings via Group-Optimal Transport. We evaluate $\mathcal{S}$tride-Net on the MIMIC-CXR and CheXpert benchmarks across race and intersectional race–gender subgroups. Across architectures including ResNet and Vision Transformers, $\mathcal{S}$tride-Net consistently improves fairness metrics while matching or exceeding baseline accuracy, achieving a more favorable accuracy–fairness trade-off than prior debiasing approaches. Our code is available Here.**

***Index Terms*—fairness, chest X-ray, disentanglement, optimal transport, underdiagnosis**

I. Introduction

In chest X-ray (CXR) analysis, deep models now approach or exceed human-level performance on several diagnostic tasks [1], [2] and are increasingly considered for clinical deployment in screening, triage, and decision support [3], [4]. Despite these advances, growing evidence shows that models with strong average performance can exhibit systematic disparities across demographic subgroups, such as race and gender [5], [6]. These disparities are not merely statistical artifacts; they translate into real clinical risks, particularly for underrepresented and intersectional populations.

A striking example arises in the prediction of the clinically critical *"No Finding"* label. This label functions as a gatekeeper in clinical workflows, signaling the absence of detectable pathology and often influencing downstream diagnostic decisions, follow-up imaging, and allocation of care. Errors in "No Finding" prediction are therefore asymmetric in their consequences: false negatives may delay diagnosis, while false positives can trigger unnecessary interventions. When such errors disproportionately affect specific demographic groups, they can exacerbate existing healthcare inequities.

Fairness challenges in chest X-ray modeling Recent studies have shown that fairness interventions in CXR models often yield inconsistent or brittle improvements across datasets, sensitive attributes, and deployment settings [7]. A core difficulty lies in the fact that demographic information can be subtly encoded in medical images through acquisition protocols, anatomical variation, or dataset-specific biases. Standard empirical risk minimization (ERM) objectives readily exploit such correlations if they improve predictive accuracy on the training distribution, even when those correlations are clinically irrelevant or harmful.

As a result, CXR models may implicitly rely on sensitive attributes as shortcuts for disease prediction. This phenomenon is particularly problematic for decision-critical labels such as "No Finding", where small representational biases can have outsized downstream effects. Moreover, these issues are often amplified under distribution shift, when models are deployed in populations that differ from the training cohort [7], [8]. Achieving equitable performance therefore requires representations that are highly predictive of pathology while being invariant to sensitive attributes; a goal that is fundamentally challenging under standard optimization.

Limitations of existing fairness approaches Most fairness-aware learning methods can be broadly categorized as pre-processing, in-processing, or post-processing techniques [8]. Group fairness objectives such as demographic parity or equalized odds often impose explicit constraints on predictions, but may induce undesirable trade-offs with accuracy or lead to "leveling down", where performance is uniformly degraded across groups to satisfy fairness criteria [9]. In clinical settings, such trade-offs are particularly problematic, as they can reduce overall diagnostic utility.

In-processing approaches, including GroupDRO [10], MMD Match [11], and adversarial debiasing [12], aim to mitigate bias during training, yet empirical evaluations show that they do not consistently outperform strong ERM baselines in medical imaging tasks [7]. Parameter-efficient fine-tuning (PEFT) methods have recently been proposed as a lightweight fairness mechanism [13], but their effectiveness remains sensitive to hyperparameter choices and computational budgets.

Disentanglement-based methods offer a principled alternative by explicitly factorizing latent representations into independent

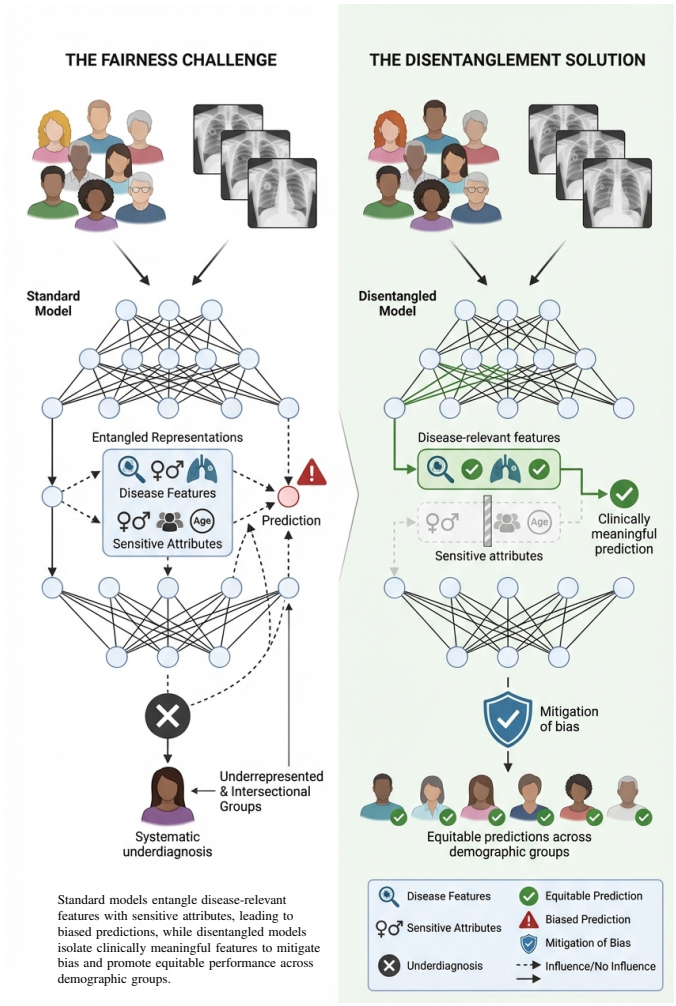

Fig. 1: Conceptual illustration of fairness and disentanglement in chest X-ray classification.

components [14], [15]. However, despite their conceptual appeal, disentanglement techniques have rarely been integrated with fairness objectives in medical imaging, and even more rarely at the level of localized, patch-based representations where clinically meaningful signals reside.

Toward fair and semantically grounded representations A promising direction to address these limitations is to intervene directly at the level of representation learning. Rather than constraining outputs post hoc, one can aim to structure the latent space so that disease-relevant features are explicitly isolated from sensitive attribute information. In the context of CXR, this suggests a model that (i) operates at the patch level to localize clinically meaningful regions, (ii) aligns visual features with disease semantics to discourage shortcut learning, and (iii) actively suppresses demographic information through adversarial disentanglement.

Motivated by these principles, we introduce $\mathcal{S}$tride-Net, a fairness-aware representation learning framework for chest X-ray classification. As illustrated in Fig. 2, $\mathcal{S}$tride-Net employs a learnable stride mask to select label-aligned image patches and route them into a shared latent space. Semantic alignment between visual patches and disease labels is enforced via Group-Optimal Transport (GOT) using BioBERT-based label embeddings, anchoring representations in clinically meaningful semantics. In parallel, adversarial confusion losses discourage the encoding of sensitive attributes, promoting demographic invariance without sacrificing predictive power.

Contributions Our contributions are summarized as follows:

- We propose $\mathcal{S}$tride-Net, a fairness-aware disentanglement framework for chest X-ray classification that explicitly separates disease-relevant features from sensitive demographic information using a learnable stride-based mask.
- We integrate semantic supervision via Group-Optimal Transport alignment between image patches and BioBERT-based disease label embeddings, grounding visual representations in clinical semantics.
- We demonstrate consistent improvements in fairness across race and intersectional race–gender subgroups, with a particular focus on underdiagnosis in the clinically important "No Finding" task.
- We validate our approach on MIMIC-CXR and CheXpert, showing that $\mathcal{S}$tride-Net achieves a favorable accuracy–fairness trade-off compared to ERM baselines and established debiasing methods.

II. Proposed Methodology

The central goal of $\mathcal{S}$tride-Net is to learn disease-discriminative visual representations from chest X-rays that are explicitly aligned with clinical label semantics while being invariant to sensitive demographic attributes such as race or gender. As illustrated in Fig. 2, the proposed architecture integrates patch-level visual encoding, label-aware semantic alignment, and adversarial fairness regularization within a unified, end-to-end trainable framework. By operating at the level of localized image patches and explicitly enforcing alignment with disease label embeddings, $\mathcal{S}$tride-Net mitigates shortcut learning driven by demographic confounders and promotes fair, interpretable decision-making.

Problem Formulation Let $X \subset \mathbb{R}^{H \times W}$ denote the space of chest X-ray images and Y the space of disease labels. We consider a supervised learning setting with a dataset of N samples

$$\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N},$$

where $x_i \in X$ is a chest X-ray image, $y_i \in Y$ is the corresponding disease label, and $s_i \in S$ denotes a sensitive attribute such as race or gender. The sensitive attribute is assumed to be available during training but not at inference time.

The objective of $\mathcal{S}$tride-Net is threefold: (i) to accurately predict disease labels $\hat{y}$ from chest X-rays, (ii) to minimize the information about sensitive attributes s encoded in the learned latent representation, and (iii) to ensure that disease predictions are grounded in clinically relevant image regions rather than spurious demographic correlations.

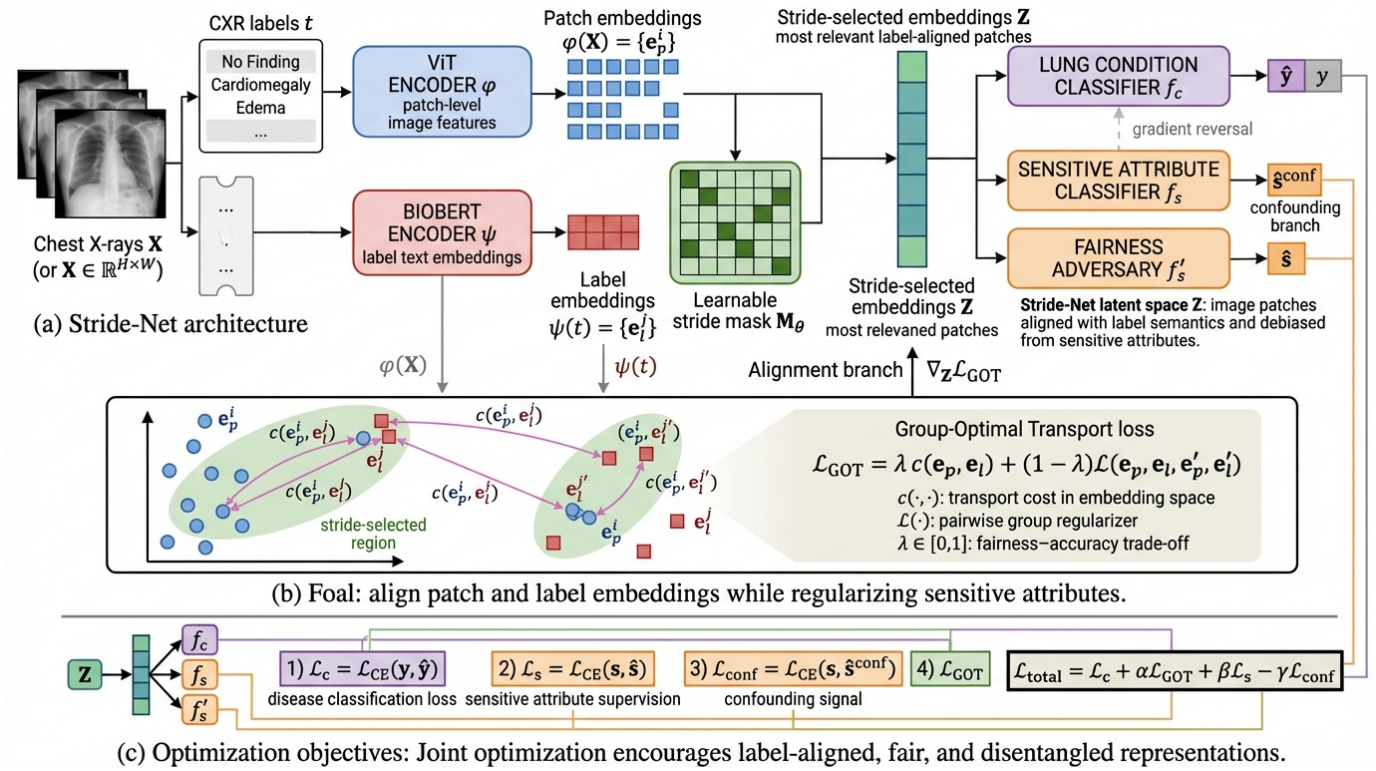

Fig. 2: $\mathcal{S}$tride-Net architecture for fair chest X-ray classification. A ViT encoder ϕ extracts patch embeddings from an input X-ray, while disease labels are encoded by BioBERT ψ. A learnable stride mask selects label-aligned patches to form the latent representation Z, which is used for disease prediction. Group-Optimal Transport enforces semantic alignment and fairness, while adversarial sensitive-attribute classifiers promote demographic invariance.

Unlike post-hoc debiasing approaches, $\mathcal{S}$tride-Net enforces fairness directly during representation learning by structuring the latent space to be label-aligned and demographically invariant.

Patch-Level Visual Encoding Given an input chest X-ray image $x \in X$, we first extract localized visual features using a Vision Transformer (ViT) encoder ϕ. The image is divided into a grid of non-overlapping patches, and each patch is mapped to a patch embedding through the encoder:

$$\phi(x) = \{e_p^i\}_{i=1}^{P},$$

where $e_p^i \in \mathbb{R}^d$ denotes the embedding of the i-th image patch, P is the total number of patches, and d is the embedding dimension.

Patch-level representations preserve spatial granularity, enabling the model to selectively attend to localized anatomical regions such as lung fields or cardiac silhouettes. This design is crucial for mitigating bias, as demographic signals often manifest globally, whereas disease evidence is typically localized.

Label Semantic Encoding To inject clinical semantics into the learning process, we encode disease labels using a pre-trained BioBERT encoder ψ. Each disease label $t \in Y$ is mapped to a semantic embedding:

$$\psi(t) = \{e_l^j\}_{j=1}^{L},$$

where $e_l^j \in \mathbb{R}^d$ represents the embedding of the j-th disease label token, and L is the number of labels.

By embedding labels in the same latent space as image patches, $\mathcal{S}$tride-Net enables direct semantic alignment between visual evidence and disease concepts. This alignment acts as a form of weak supervision that encourages the model to rely on medically meaningful image regions.

Stride-Based Patch Selection Not all image patches contribute equally to disease prediction. To identify the most label-relevant visual regions, we introduce a learnable stride mask $M_\theta \in \mathbb{R}^{P \times L}$, parameterized by θ. This mask computes relevance scores between patch embeddings $\{e_p^i\}$ and label embeddings $\{e_l^j\}$.

The stride mask selectively activates a subset of patches that are most semantically aligned with the disease labels. Formally, the stride-selected latent representation is given by:

$$Z = \text{StrideSelect}(\phi(x), \psi(t), M_\theta),$$

where $Z \subset \{e_p^i\}$ consists of the most relevant, label-aligned patch embeddings.

This stride-based mechanism enforces sparsity and interpretability, ensuring that downstream predictions are driven

by clinically relevant regions rather than background artifacts or demographic proxies.

Group-Optimal Transport Alignment To further enforce semantic consistency and fairness, $\mathcal{S}$tride-Net employs a Group-Optimal Transport (GOT) loss to align patch embeddings with label embeddings while regularizing sensitive attribute disparities. The GOT loss is defined as:

$$\mathcal{L}_{\mathrm{GOT}} = \lambda\, c(e_p, e_l) + (1-\lambda)\, \mathcal{L}(e_p, e_l, e'_p, e'_l),$$

where $c(\cdot,\cdot)$ denotes the transport cost in the embedding space, $\mathcal{L}(\cdot)$ is a pairwise group regularizer, and $\lambda \in [0,1]$ controls the trade-off between accuracy and fairness.
Intuitively, the first term encourages patch embeddings to align closely with their corresponding disease label embeddings, while the second term enforces group-wise consistency across different sensitive attribute values. This alignment reduces representational discrepancies that could otherwise lead to biased predictions.

Disease Classification Branch The stride-selected latent representation Z is fed into a disease classifier f_c to predict disease labels:

$$\hat{y} = f_c(Z).$$

The disease classification loss is defined using standard cross-entropy:

$$\mathcal{L}_c = \mathcal{L}_{\mathrm{CE}}(y, \hat{y}).$$

Because Z contains only label-aligned, stride-selected patches, the classifier is forced to rely on clinically meaningful evidence rather than spurious correlations.

Sensitive Attribute Prediction and Adversarial Fairness To explicitly remove sensitive information from the latent space, we introduce two sensitive attribute classifiers. The first classifier f_s predicts the sensitive attribute from Z:

$$\hat{s} = f_s(Z),$$

with supervision loss

$$\mathcal{L}_s = \mathcal{L}_{\mathrm{CE}}(s, \hat{s}).$$

In parallel, a fairness adversary f'_s is trained using gradient reversal to maximize confusion with respect to sensitive attribute prediction:

$$\hat{s}^{\mathrm{conf}} = f'_s(Z),$$

with corresponding loss

$$\mathcal{L}_{\mathrm{conf}} = \mathcal{L}_{\mathrm{CE}}(s, \hat{s}^{\mathrm{conf}}).$$

The gradient reversal mechanism ensures that the feature extractor learns representations that are predictive of disease while being uninformative of sensitive attributes.

Joint Optimization Objective All components of $\mathcal{S}$tride-Net are trained jointly in an end-to-end manner. The total loss function is given by:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_c + \alpha\mathcal{L}_{\mathrm{GOT}} + \beta\mathcal{L}_s - \gamma\mathcal{L}_{\mathrm{conf}},$$

where α, β, and γ are scalar hyperparameters controlling the relative contributions of semantic alignment, sensitive attribute supervision, and adversarial debiasing.
This joint objective encourages the model to learn latent representations that are simultaneously discriminative, label-aligned, and fair.

III. Experimental Setup

Datasets and Preprocessing We evaluate $\mathcal{S}$tride-Net on two widely used large-scale chest X-ray benchmarks: MIMIC-CXR [16] and CheXpert [1]. Following prior fairness-focused work in medical imaging [17], we consider the binary *No Finding* classification task and analyze performance disparities across sensitive demographic attributes, specifically race and gender.
Sensitive attribute annotations are defined using the same protocol as [17]. To ensure reliable subgroup analysis, samples with missing sensitive attribute information are excluded. This filtering step is critical for fairness evaluation, as inaccurate subgroup membership can obscure or distort disparity measurements.
All chest X-ray images are resized to $224\times224\times3$ to match the input resolution of the Vision Transformer backbone. Labels annotated as "Not mentioned" are treated as "uncertain" and assigned a value of -1, consistent with standard preprocessing practices. We split each dataset into training and testing sets using an 80/20 stratified split that preserves the distribution of sensitive subgroups. During training, standard data augmentations; including random horizontal flips and small-angle rotations; are applied to improve robustness and generalization [18].

Training Configuration All models are trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. Training is performed for 20 epochs on a single NVIDIA RTX A6000 GPU. Unless otherwise specified, hyperparameters are held constant across datasets and model variants to ensure fair comparison.
Semantic supervision for the Group-Optimal Transport (GOT) module is provided by fixed label embeddings extracted using BioBERT [19]. These embeddings encode clinical semantics of disease labels and remain frozen throughout training, allowing the model to align visual features with stable semantic anchors. For Vision Transformer-based variants, we fine-tune a ViT-B model pre-trained on ImageNet .

Fairness Evaluation Metrics We evaluate fairness using two complementary metrics that capture different aspects of subgroup performance equity.
1) Predictive Quality Disparity (PQD). Let R denote the set of sensitive subgroups, and let acc_j be the classification accuracy for subgroup $j \in R$. PQD measures the uniformity of predictive performance across subgroups:

$$PQD = \frac{\min_{j\in R} \mathrm{acc}_j}{\max_{j\in R} \mathrm{acc}_j}. \tag{1}$$

Higher PQD values indicate smaller performance gaps between the best- and worst-performing subgroups.

TABLE I: Overall accuracy and fairness comparison across sensitive attributes, evaluated across race and intersectional race-gender subgroups.

Methods ↓	MIMIC-CXR			CheXpert		
	Acc/Avg	PQD	EOM	Acc/Avg	PQD	EOM
Individual Subgroup: "Race"						
ResNet-18	0.780	0.850	0.680	0.898	0.896	0.480
UBAIA [17]	0.789	0.935	0.830	0.895	0.943	0.745
CheXclusion [6]	0.777	0.850	0.839	0.887	0.903	0.700
$\mathcal{S}$tride-Net (Ours)	0.805	0.922	0.870	0.903	0.928	0.803
Intersection Subgroup: "Race-Gender"						
ResNet-18	0.789	0.918	0.685	0.900	0.881	0.460
UBAIA [17]	0.788	0.821	0.703	0.907	0.868	0.452
$\mathcal{S}$tride-Net (Ours)	0.806	0.879	0.705	0.909	0.897	0.466

High highlight intensity denote high performance

2) Equality of Opportunity Measure (EOM). Let M denote the number of disease classes, and let $p(\hat{y} = i \mid y = i, r = j)$ be the true positive rate for class i within subgroup j. EOM evaluates whether true positive rates are consistent across subgroups:

$$EOM = \frac{1}{M} \sum_{i=1}^{M} \frac{\min_{j \in R} p(\hat{y} = i \mid y = i, r = j)}{\max_{j \in R} p(\hat{y} = i \mid y = i, r = j)}. \quad (2)$$

Both metrics lie in $[0, 1]$, with higher values indicating more equitable model behavior.

IV. Results

Overall Performance Across Sensitive Subgroups We compare $\mathcal{S}$tride-Net against three representative baselines: (1) a standard ResNet-18 trained with empirical risk minimization (ERM) [20], (2) UBAIA [17], a fairness-aware method designed to reduce underdiagnosis, and (3) CheXclusion [6], which removes spurious correlations via exclusion mechanisms.

Table I summarizes results across race and intersectional race-gender subgroups. On MIMIC-CXR, $\mathcal{S}$tride-Net improves average accuracy over UBAIA by $+1.6$ percentage points (0.805 vs. 0.789) while also increasing EOM by $+4.0$ points. On CheXpert, similar gains are observed, with accuracy improving by $+0.8$ points and EOM by $+5.8$ points.

These improvements highlight a key advantage of $\mathcal{S}$tride-Net: fairness gains are not achieved by flattening model predictions or sacrificing discriminative power. Instead, stride-based patch selection and semantic alignment enable the model to retain strong disease-relevant signals while suppressing demographic shortcuts.

Intersectional Fairness Analysis Fairness challenges are often most severe for intersectional subgroups. When evaluated across race-gender intersections, $\mathcal{S}$tride-Net continues to outperform competing approaches. Accuracy increases to 0.806 on MIMIC-CXR and 0.909 on CheXpert, while both PQD and EOM improve relative to prior methods.

These gains are particularly notable because intersectional groups often suffer from compounding biases. The results suggest that explicitly disentangling disease-relevant features from sensitive attributes at the representation level; as enforced by the GOT alignment and adversarial confusion losses; helps stabilize performance even in low-support subgroups.

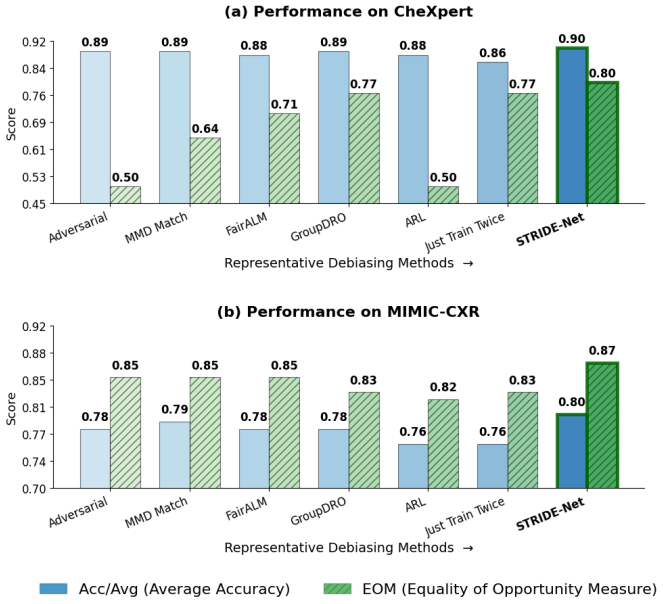

Fig. 3: **Accuracy-fairness trade-offs on CheXpert and MIMIC-CXR.** Comparison of $\mathcal{S}$tride-Net with representative debiasing methods under the Equality of Opportunity (EOM) metric. $\mathcal{S}$tride-Net consistently occupies the favorable upper-right region, indicating improved fairness without sacrificing predictive accuracy, especially for underrepresented demographic groups. (Please Zoom In).

Accuracy-Fairness Trade-off Table I reveals distinct trade-off patterns across methods. UBAIA achieves strong PQD but comparatively lower accuracy and EOM, indicating that it equalizes subgroup performance partly by reducing true positive rates for well-performing groups. CheXclusion improves certain fairness metrics but lags behind in overall accuracy.

In contrast, $\mathcal{S}$tride-Net consistently achieves the best balance between accuracy and fairness. By constraining the latent space to contain only label-aligned, stride-selected patches, the model avoids the typical trade-off where fairness improvements degrade clinical utility.

Comparison with Debiasing Baselines Figure 3 situates $\mathcal{S}$tride-Net within the broader landscape of debiasing approaches, including adversarial training [12], MMD Match [11], FairALM [21], GroupDRO [10], ARL [22], and Just Train Twice [23]. While several baselines achieve modest fairness improvements, these gains often come at the expense of accuracy or exhibit instability across datasets.

$\mathcal{S}$tride-Net consistently occupies the upper-right region of the accuracy-EOM plane, indicating improved fairness without sacrificing predictive performance. This behavior reflects the complementary nature of stride-based localization, semantic alignment, and adversarial disentanglement.

TABLE II: Component-wise ablation on ViT-based models. Impact of semantic alignment (GOT) and stride-based disentanglement on accuracy and fairness across MIMIC-CXR and CheXpert.

Methods ↓	MIMIC-CXR			CheXpert		
	Acc/Avg	PQD	EOM	Acc/Avg	PQD	EOM
Individual Subgroup: "Race"						
ViT-B	0.772	0.943	0.767	0.853	0.938	0.751
ViT-B+GOT	0.784	0.929	0.779	0.865	0.931	0.763
$\mathcal{S}$tride-Net (Ours)	0.805	0.922	0.870	0.875	0.927	0.803
Intersection Subgroup: "Race-Gender"						
ViT-B	0.780	0.761	0.377	0.889	0.837	0.433
ViT-B+GOT	0.795	0.733	0.389	0.892	0.837	0.435
$\mathcal{S}$tride-Net (Ours)	0.806	0.879	0.705	0.909	0.897	0.466

High highlight intensity denote high performance

Component-wise Ablation Study To better understand the contribution of each architectural component, we conduct ablation studies using a ViT-B backbone. As shown in Table II, the baseline ViT-B (ERM) achieves strong accuracy but exhibits pronounced fairness gaps, particularly for intersectional subgroups.

Adding the GOT alignment module improves both accuracy and fairness, confirming that aligning patch embeddings with label semantics stabilizes subgroup performance. The full $\mathcal{S}$tride-Net; incorporating the learnable stride mask and adversarial confusion loss; yields the best EOM and substantial PQD improvements while further boosting accuracy.

Hyperparameter Sensitivity Hyperparameter sweeps reveal that $\lambda_{\text{GOT}} = 0.8$ provides a favorable balance between semantic alignment and task performance. Larger values over-regularize the representation and degrade accuracy. Similarly, setting the confusion loss weight to $\alpha_{\text{conf}} = 2$ yields strong fairness gains with minimal impact on utility, underscoring the importance of moderation in adversarial disentanglement.

V. Conclusion

$\mathcal{S}$tride-Net offers a fairness-aware, disentangled representation learning framework for chest X-ray classification that combines adaptive masking, confusion regularization, and GOT-based semantic alignment. On MIMIC-CXR and CheXpert, $\mathcal{S}$tride-Net consistently improves fairness across race and race-gender subgroups while preserving or improving diagnostic accuracy. Although this study focuses on chest X-rays and the "No Finding" task, the principles are applicable to other modalities and label sets where sensitive information can be encoded in visual features. Future work will target computational efficiency, interpretability of learned masks and validation in larger, more heterogeneous cohorts.

References

[1] J. Irvin, P. Rajpurkar, M. Ko, and et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[2] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep convolutional neural networks in medical image analysis: A review," *Information*, vol. 16, no. 3, p. 195, 2025.

[3] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.

[4] W. Liawrungrueang, W. Cholamjiak, and et al., "Artificial intelligence for cervical spine fracture detection: a systematic review of diagnostic performance and clinical potential," *Global spine journal*, vol. 15, 2025.

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[6] 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.

[7] Y. Zong, Y. Yang, and T. Hospedales, "Medfair: benchmarking fairness for medical imaging," *arXiv preprint arXiv:2210.01725*, 2022.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[9] D. Zietlow, M. Lohaus, G. Balakrishnan, and et al., "Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 410–10 421.

[10] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[11] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of biomedical informatics*, vol. 113, p. 103621, 2021.

[12] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, 2018.

[13] R. Dutt, O. Bohdal, S. A. Tsaftaris, and T. Hospedales, "Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis," *arXiv preprint arXiv:2310.05055*, 2023.

[14] E. Tartaglione, C. A. Barbano, and M. Grangetto, "End: Entangling and disentangling deep representations for bias correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 508–13 517.

[15] U. Rahman, R. Imam, M. Yaqub, and D. Mahapatra, "Decoupling clinical and class-agnostic features for reliable few-shot adaptation under shift," in *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Springer, 2025, pp. 123–133.

[16] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[17] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature medicine*, vol. 27, no. 12, pp. 2176–2182, 2021.

[18] R. Imam, R. Marew, and M. Yaqub, "On the robustness of medical vision-language models: Are they truly generalizable?" in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2025, pp. 233–256.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.

[22] P. Lahoti, A. Beutel, J. Chen, and et al., "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[23] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6781–6792.