FairSimCLR: A Fairness-Aware Contrastive Learning Framework for Demographic Bias Mitigation in Dermatology Imaging

Anonymous Submission

Self-supervised learning (SSL) has emerged as a powerful approach for learning medical image representations without labeled data. However, most SSL methods do not explicitly account for fairness, leading to biased performance across demographic subgroups (Seyyed-Kalantari et al., 2021). In this work, we introduce FairSimCLR, a fairness-aware contrastive learning framework designed to reduce demographic disparities in medical imaging, specifically within dermatology datasets. We adopt the normalized temperature-scaled cross-entropy (NT-Xent) loss from SimCLR (Chen et al., 2020). To enforce demographic invariance, FairSimCLR introduces an auxiliary fairness loss:

$$\mathscr{L}_{\text{FairSimCLR}} = \mathscr{L}_{\text{SimCLR}} + \lambda \cdot \mathscr{L}_{\text{fair}}$$

Here, $\mathscr{L}_{\text{fair}}$ penalizes representation disparities across demographic groups, and λ balances fairness and contrastive objectives.

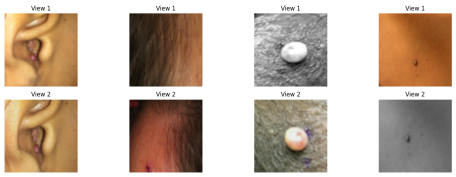

Figure 1: Augmentation pipeline used during SimCLR and FairSimCLR training, including random flips, color jitter, Gaussian blur, and cropping. These augmentations help enforce semantic invariance in learned representations.

FairSimCLR extends SimCLR by incorporating group-aware sampling to ensure balanced representation of demographic subgroups during contrastive pair selection, and a fairness-regularized contrastive loss that penalizes representational bias across groups. We pretrained both SimCLR and FairSimCLR on three dermatology datasets—PAD-UFES-20, Italian Dermatology, and Diverse Dermatology Images (DDI)—and extracted frozen embeddings. To evaluate the quality and fairness of these representations, we trained linear classifiers (logistic regression, random forest, and multi-layer perceptron [MLP]) on the embeddings to predict six diagnostic categories.

Our fairness evaluation focused on three metrics: Demographic parity difference (DPD), equal opportunity difference (EOD) and predictive equality difference (PQD) (Zhang et al., 2018) calculated between sex, skin tone and age-based subgroups. Across all datasets and classifiers, FairSimCLR consistently reduced DPD, EOD, and PQD compared to the baseline SimCLR. Notably, the MLP classifier trained on FairSimCLR embeddings achieved the highest macro-averaged F1-score and ROC AUC, demonstrating that fairness improvements did not come at the cost of classification performance.

This study shows that fairness-aware modifications to contrastive learning can improve both representational equity and predictive performance in medical imaging. Going forward, we aim to extend FairSimCLR to multi-modal pretraining with clinical metadata, generalize to other SSL frameworks, and explore adversarial and group-adaptive fairness interventions.