Generative Counterfactual Augmentation for Bias Mitigation

Jason Uwaeze, Pranav Kulkarni, Vladimir Braverman, Michael A. Jacobs, Vishwa Parekh

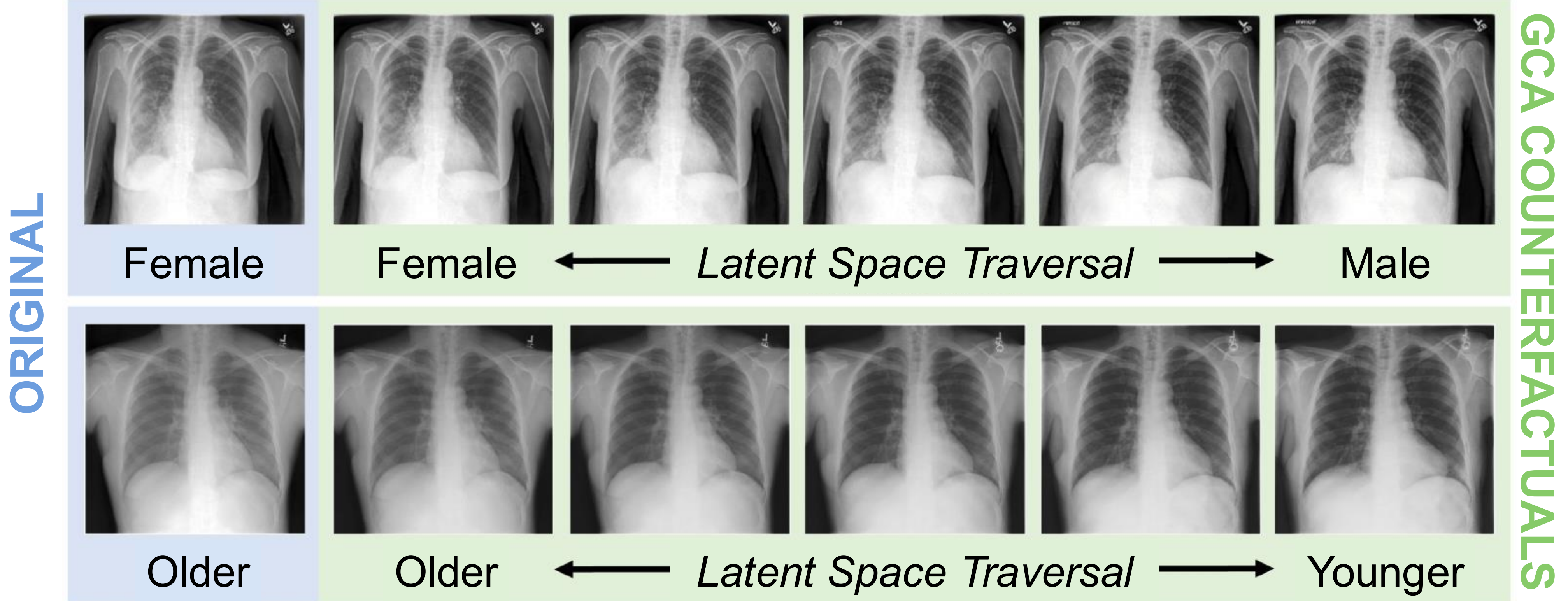
RICE · JOHNS HOPKINS UNIVERSITY · UTHealth Houston · UNIVERSITY OF MARYLAND 1856

ICCV HONOLULU HAWAII OCT 19-23, 2025

A new approach to adversarial bias mitigation

What? Constructed demographic-complete training data augmentations to mitigate adversarial bias in pneumonia chest x-ray (CXR) **DenseNet121** classifiers

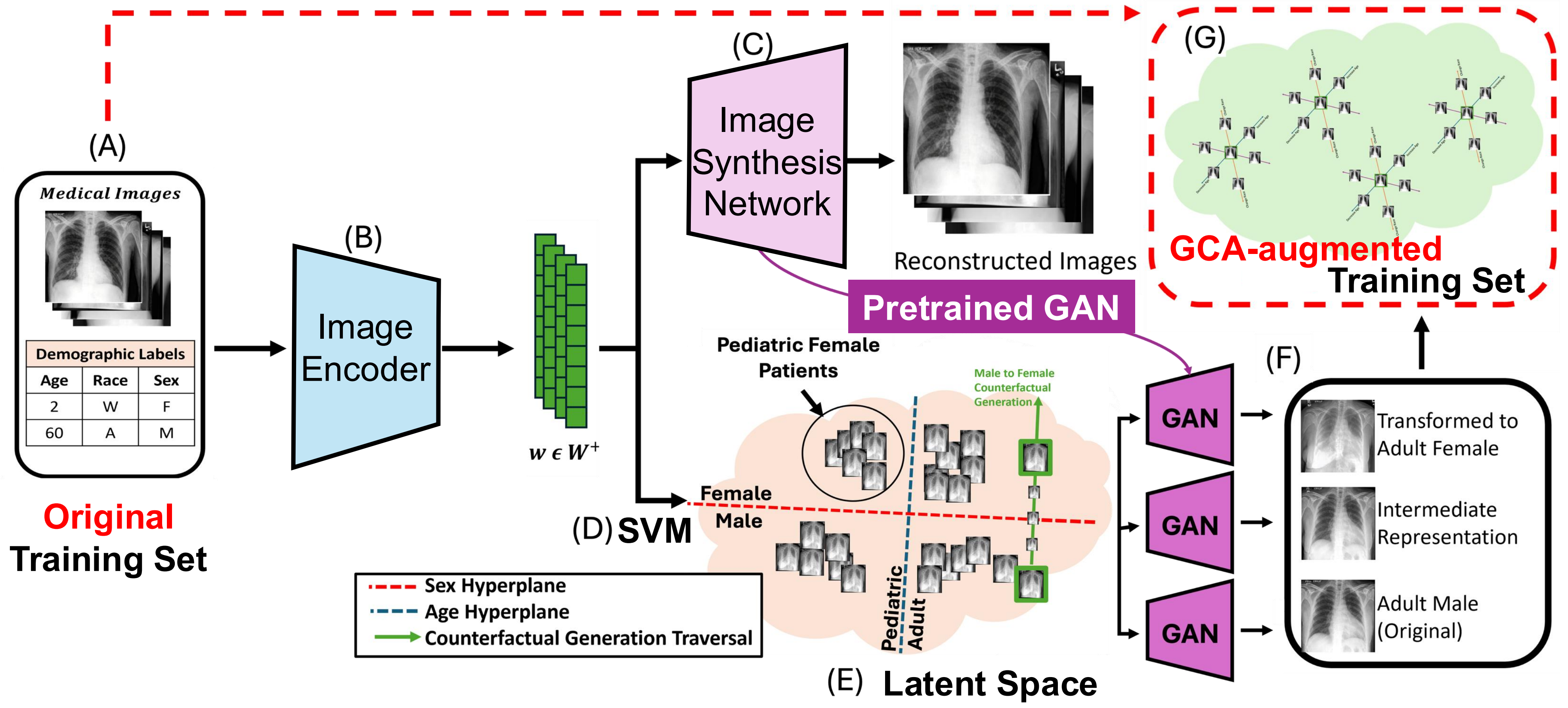
How? Generated counterfactuals using StyleGAN3 synthesis network and SVM-guided latent space traversals for **sex** and **age** attributes



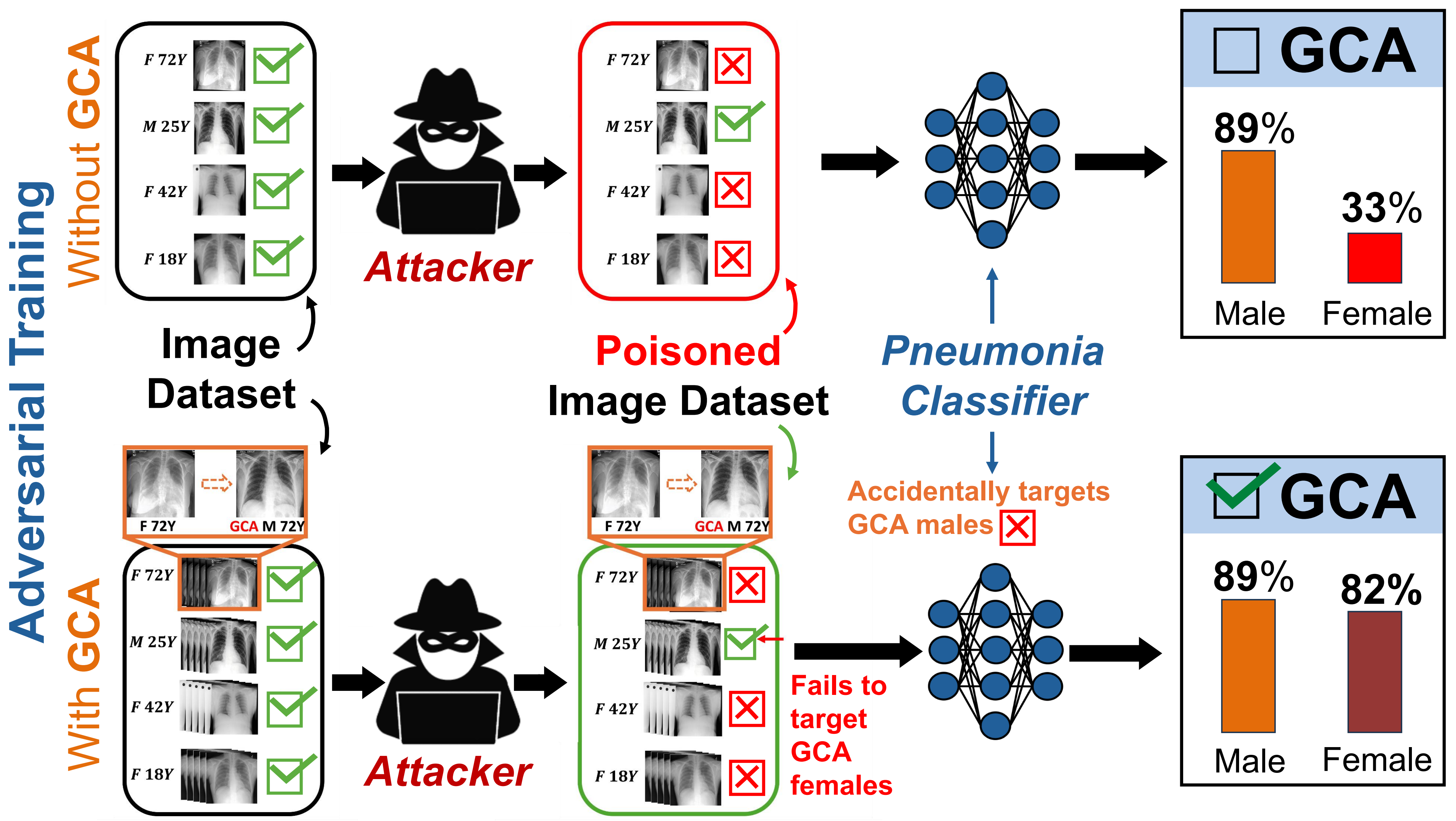
Why?

- Deep learning methods learn protected characteristics (e.g., sex and age) to achieve excellent classification performance
- Potentially amplifying existing systemic disparities in healthcare and worsening patient outcomes
- Thus, we introduce **generative counterfactual augmentations (GCA)**

Generative counterfactual augmentation framework

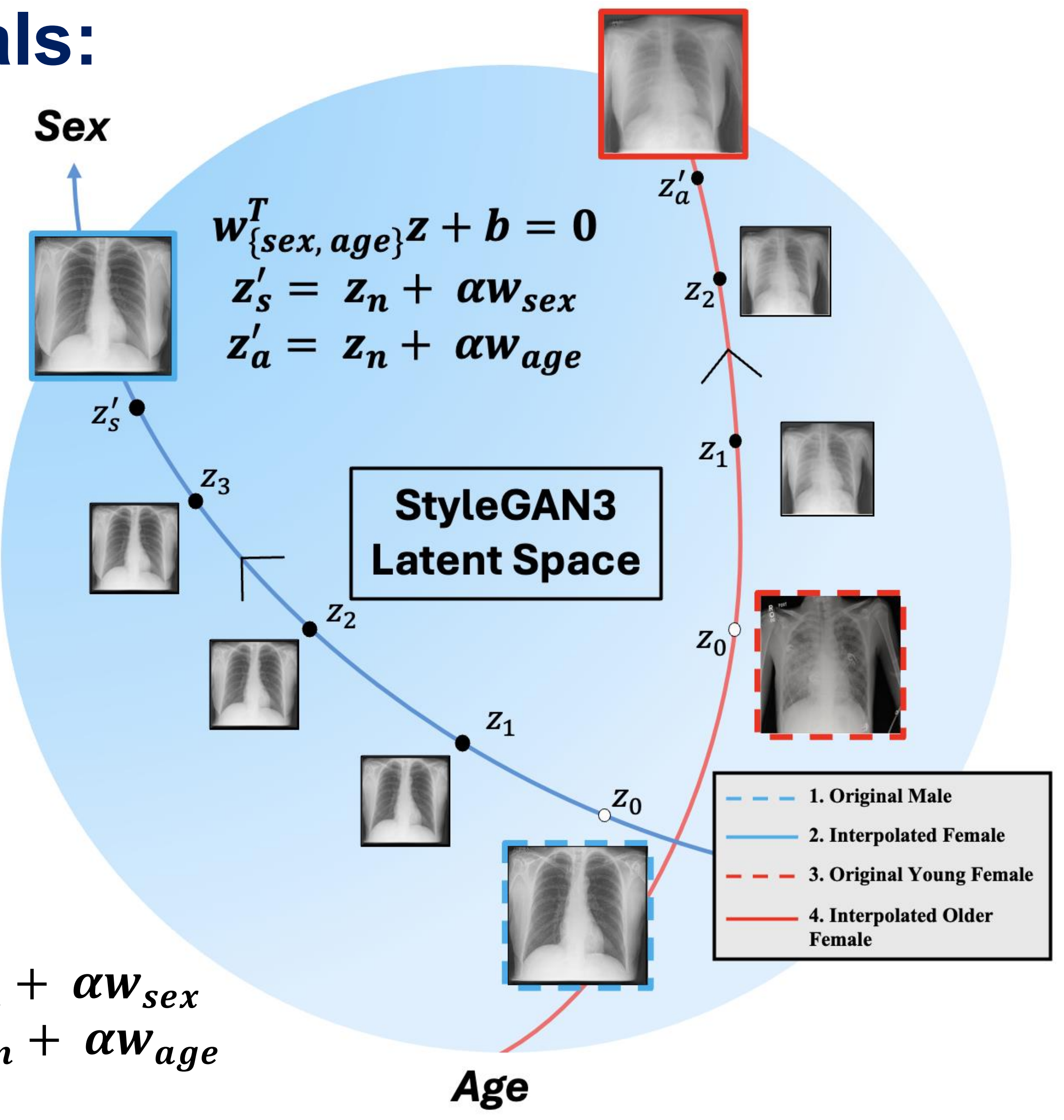


Medical image classifiers are susceptible to adversarial bias attacks:

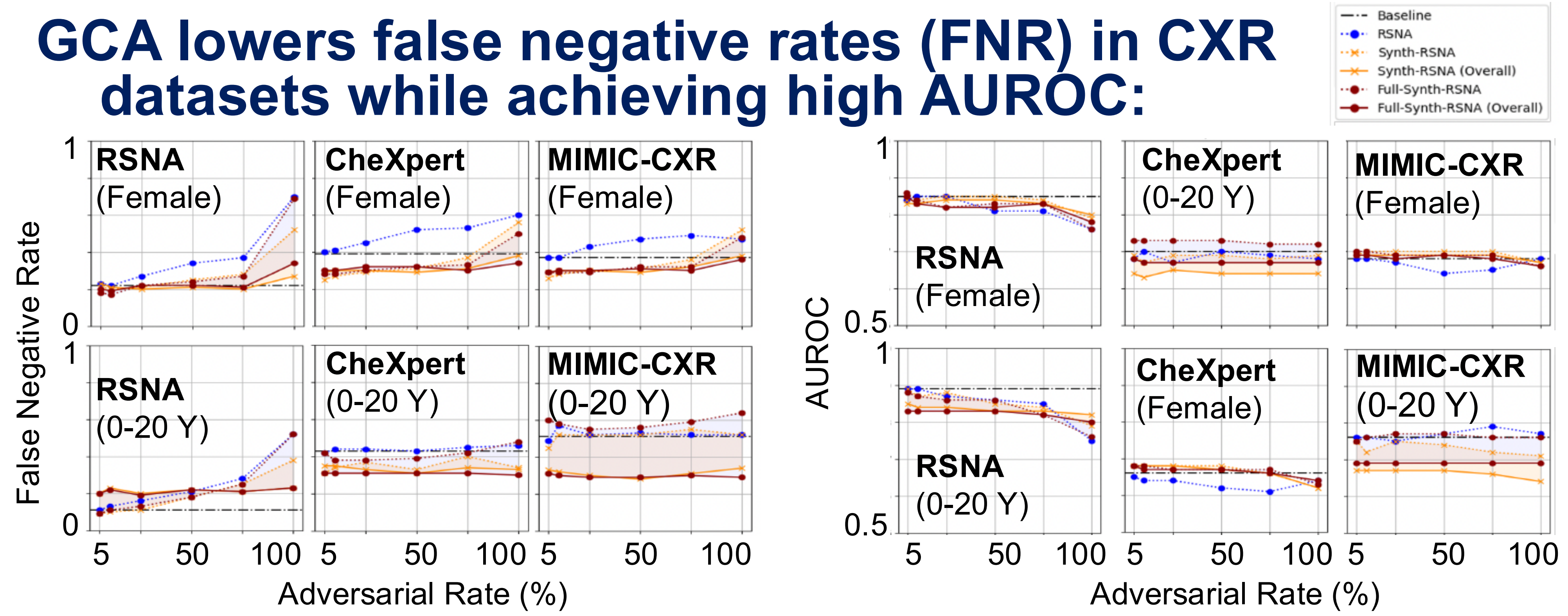


Latent space interpolations for generating CXR counterfactuals:

- Train SVM classifiers in the StyleGAN3 latent space
- Use learned SVM hyperplanes to guide latent interpolations across sex and age attributes
- Let w_{sex} and w_{age} denote normal vectors separating demographic groups
- Interpolate by traversing perpendicular to each hyperplane:
  - ***Sex interpolation***: $z_s' = z_n + \alpha w_{sex}$
  - ***Age interpolation:*** $z_a' = z_n + \alpha w_{age}$

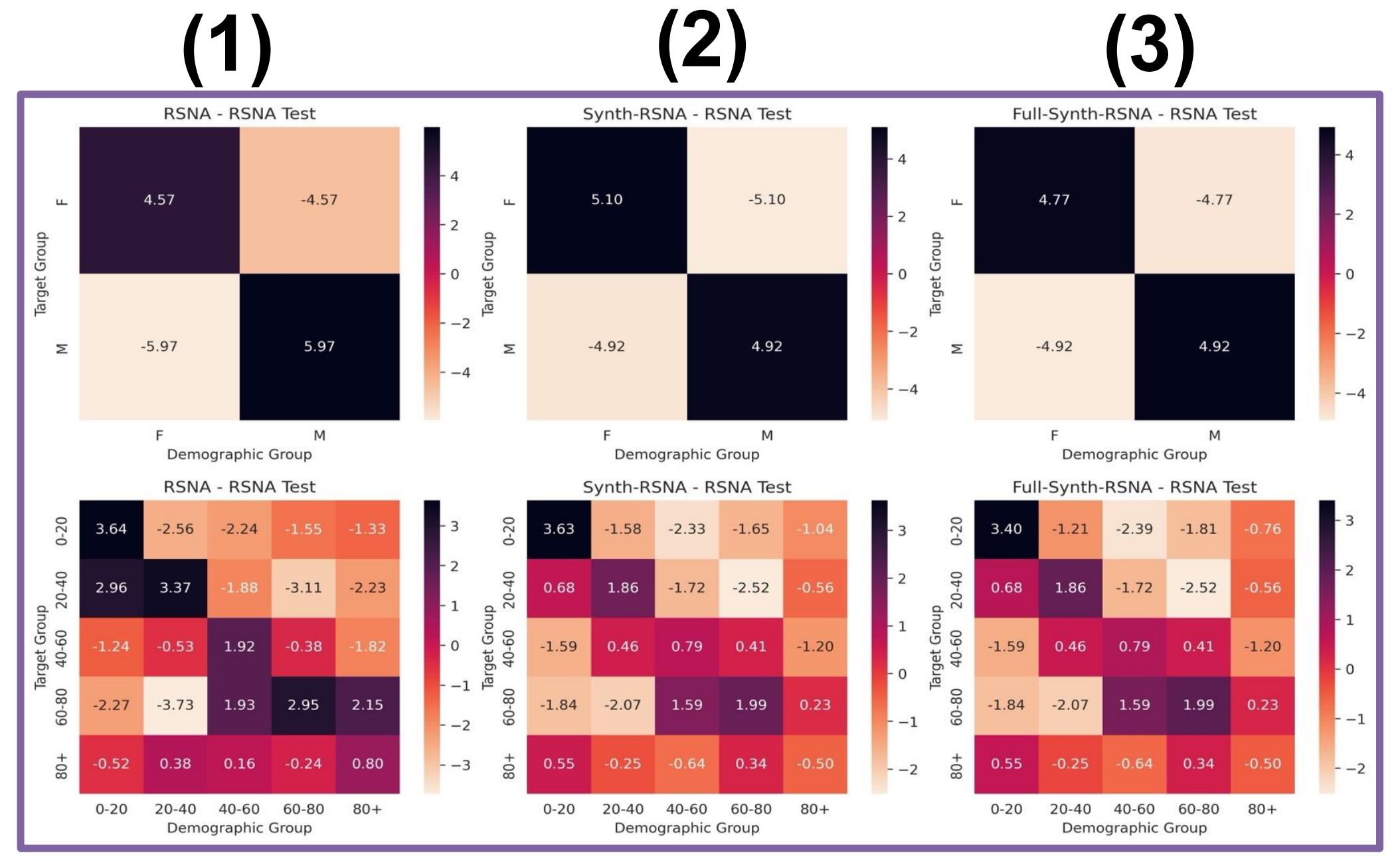


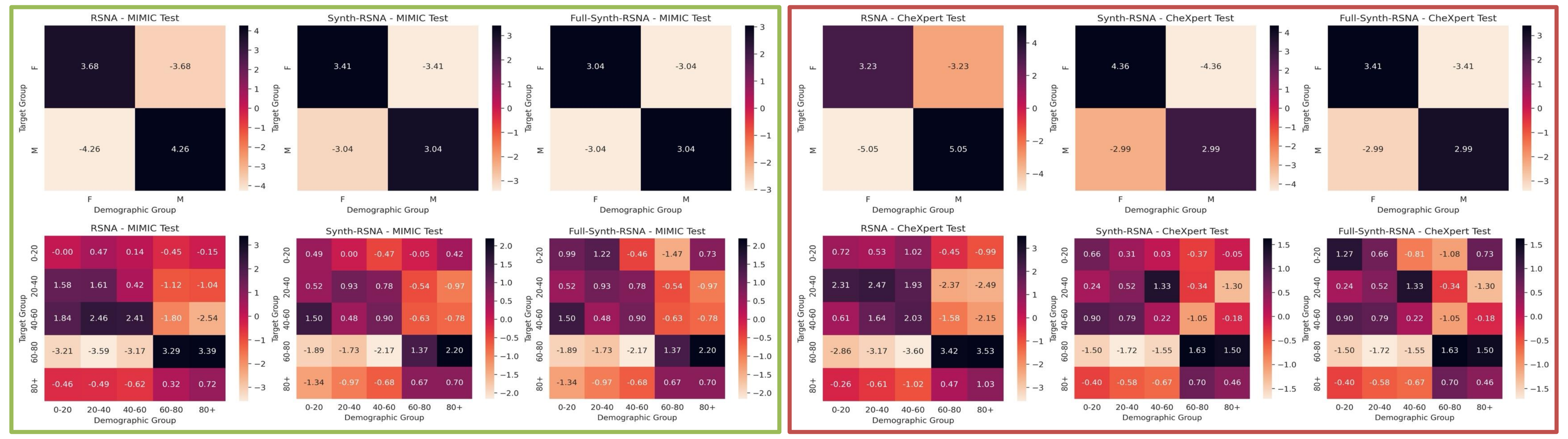
GCA lowers false negative rates (FNR) in CXR datasets while achieving high AUROC:



GCA reduces vulnerability of the targeted and non-targeted subgroups:

- Models trained on **(1)** original, **(2)** demographic-specific synthetic, and **(3)** demographic-complete synthetic RSNA datasets
- Models tested on **RSNA**, **MIMIC-CXR**, and **CheXpert** datasets
- Impact of GCA remained consistent across external MIMIC-CXR and CheXpert datasets





Summary/Conclusion:

- We propose GCA, a counterfactual demographic-complete data augmentation method to mitigate adversarial bias in deep learning.
- Using controlled injections of underdiagnosis bias across age and sex groups, we show that GCA reduces FNR disparities, preserves high AUROC, and generalizes across external datasets.
- **Future work**: to adapt GCA for other imaging modalities and tasks, ensuring trustworthiness in real-world clinical settings.

Paper & Code

Scan Me