The Impact of Protected Variable Integration in Multimodal Pretraining: A Case Study on ECG Waveforms and ECG Notes Pretraining

Zhale Nowroozilarki[1], Sicong Huang[1], Sadeer Al-Kindi[2], Bobak J. Mortazavi[1,3]
Computer Science & Engineering, Texas A&M Univeristy[1]
Center for Cardiovascular Computational Health & Precision Medicine (C3-PH), Houston Methodist Hospital[2]
Center for Remote Health Technologies and Systems, Texas A&M Univeristy[3]
{zhale, siconghuang, bobakm}@tamu.edu, sal-kindi@houstonmethodist.org

***Abstract*—Electrocardiogram (ECG) interpretation using deep learning has shown promising results in detecting cardiac rhythm abnormalities. However, growing evidence suggests that model performance can vary significantly across demographic subgroups, raising concerns about algorithmic fairness in clinical deployment. In this study, we explore whether incorporating protected variables—specifically age and sex—into multimodal contrastive pretraining can reduce downstream performance disparities. We use a CLIP-style architecture to align ECG signals with machine-generated rhythm descriptions, training two variants: one with text alone and one with demographic augmentation. After pretraining, we evaluate frozen ECG embeddings using linear probing on a binary classification task distinguishing normal from abnormal rhythms. Our results show that including demographic information during pretraining can reduce performance gaps across age groups and maintains comparable or improved accuracy across sex. These findings highlight the potential of fairness-aware representation learning to improve subgroup equity in clinical machine learning applications.**

***Index Terms*—Electrocardiogram, Contrastive Pretraining, Multimodal Representation Learning, Fairness**

I. Introduction

Deep learning models have demonstrated remarkable accuracy in interpreting electrocardiograms (ECGs) for screening and diagnosing of heart conditions [1]–[3]. With the growth of these AI-ECG models comes a need to evaluate their fairness and generalizability. Model performance can vary with patient demographics: differences in patient age or sex can influence ECG waveforms and thus potentially affect algorithm predictions [4]. These findings underscore the risk that a one-size-fits-all ECG model may inadvertently favor certain demographics. Evaluating clinical AI tools in such settings using additional information about the patients to ensure more equitable outcomes across subgroups is crucial to prevent exacerbating health disparities and insufficient training datasets. Multimodal pretraining integrates complementary patient data alongside ECG signals to help models learn more robust representations of the ECG signals of the patients.

Multimodal representation learning has been shown to improve model robustness and can reduce the need for large labeled datasets [5], [6]. Specifically, contrastive multimodal approaches have shown that pretrained models create more robust patient representations across modalities, including with ECG data [7], [8]. Pretraining on paired ECGs and their corresponding clinical descriptions, for example, learns more rich feature representations that capture both electrophysiological patterns and diagnostic context [9].

However, current AI-ECG representation learning frameworks typically ignore patient demographics during training; they treat the data as if one distribution fits all. Performance can vary substantially across patient subgroups defined by age and sex [10], [11]. These disparities raise concerns regarding algorithmic bias and equitable healthcare deployment. This raises an open research question: *how might the inclusion of protected attributes (like age and sex) in the pretraining stage influence the learned representations?* On one hand, explicitly incorporating these attributes could allow the model to account for physiological differences between subgroups. On the other hand, it could also risk encoding spurious correlations with demographic factors, possibly exacerbating disparities if the model over-relies on them. There are currently conflicting perspectives: some advocate for demographically "neutral" representations by actively discouraging the encoding of sensitive information (e.g. using contrastive objectives that push apart representations sharing a protected attribute) [12], while others note that adding demographic features to models does not guarantee equitable performance [4].

We investigate the impact of incorporating protected demographic variables into multimodal pretraining on paired ECGs and text descriptions by augmenting protected demographic variables into the texts. We select a binary classification task of distinguishing sinus rhythm from abnormal rhythms to evaluate the performance gain and fairness of original pretrained vs. the augmented one.

We train a contrastive model to align each ECG waveform with its textual description; critically, we perform a comprehensive analysis of the downstream rhythm classification performance with and without protected attributes (age and sex) during pretraining. By comparing models pretrained with and without protected variables, we provide empirical insights into whether demographic-informed representation learning mitigates or exacerbates inequities

in model performance. Our code is publicly available at https://github.com/stmilab/ECGClip-Fair-Eval.

II. Related Work

A. Fairness-Aware Representation Learning

Fair representation learning aims to encode data into latent features that preserve task-relevant information while removing or obscuring protected attribute effects [13]. Some approaches formulated this as an optimization to maximize the predictive utility of the representation while obfuscating sensitive attributes [14]. On the other hand, Lin et al. incorporated demographic attributes (e.g., age, sex) into model pretraining to promote fairness [15].

To overcome these limitations, researchers have proposed contrastive learning frameworks that explicitly leverage protected attributes during representation learning. For example, *FairEHR-CLP* generates synthetic patient counterparts with varied demographics and uses contrastive objectives to align patient representations across sensitive attributes, thereby learning a demographically invariant embedding [16]. Likewise, Agarwal et al. [17] introduced *Debias-CLR*, which trains contrastive encoders for structured and unstructured modalities while minimizing demographic leakage in the learned embeddings. Despite these advances, the comparison between demographic-aware vs demographic-unaware contrastive pretraining remains underexplored in the context of physiological signals like ECG and has yet to be evaluated in multimodal signal–text representation settings.

B. Fairness Evaluation in ECG Models

Most prior studies assess model fairness *post hoc* by measuring performance disparities across demographic subgroups, rather than enforcing fairness during training. In the medical AI literature, evaluations stratified by protected attributes have repeatedly uncovered uneven model behavior across patient groups [10]. For example, Kaur et al. [11] found that an ECG deep learning model for heart failure risk stratification had declining accuracy with increasing patient age and performed significantly worse on some intersectional subgroups and observed that even providing the model with demographic inputs or training separate models for each subgroup did not abolish the disparity. Perez-Alday et al. [18] showed that adding a fairness-driven regularization term to penalize subgroup performance gaps could mitigate bias in arrhythmia detection, albeit at the cost of overall accuracy. There is still a growing need to shift from reactive fairness audits toward proactive fairness-aware model design and evaluation.

III. Methods

A. Pretraining with Contrastive Multimodal Learning

We adopt a CLIP-style architecture to learn joint representations of ECG waveforms and their corresponding textual descriptions. Let x^{ecg} be a one-dimensional ECG signal and x^{text} be a machine-generated note describing the cardiac rhythm. The goal of contrastive pretraining is to bring matching ECG–text pairs closer in the embedding space while pushing apart non-matching pairs.

Two text variants are used:

- **Original:** Unmodified note, e.g., "*sinus tachycardia*".
- **Augmented:** Note appended with age and sex, e.g., "*This ECG belongs to a 75-year-old male with sinus tachycardia*".

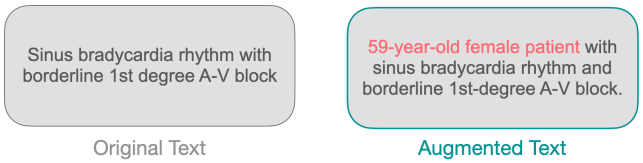

Fig. 1. ECG notes: Original vs Augmented

The ECG encoder is a 1D convolutional transformer with positional encoding, while the text encoder is a pretrained BERT model. Both outputs are projected into a shared embedding space and normalized. We apply a symmetric InfoNCE loss:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp(\text{sim}(z_i^{\text{ecg}}, z_i^{\text{text}})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^{\text{ecg}}, z_j^{\text{text}})/\tau)}\right]$$

where sim denotes cosine similarity, z are projected embeddings, and τ is a temperature parameter.

B. Incorporating Protected Variables

In the augmented setup, age and sex are explicitly embedded in the input text (Fig 1). This allows the model to condition representations on demographic information during pretraining, enabling it to learn age- and sex-aware alignments between ECG signals and text. This approach is motivated by the known physiological variation in ECG morphology across demographic groups [19].

C. Downstream Task: ECG Rhythm Classification

To evaluate the learned representations, we perform a binary classification task: **normal** rhythm (sinus) vs. **abnormal** (all other rhythms). We freeze the pretrained encoders and extract embeddings from the ECG encoder for each sample. These embeddings are then used as input to a downstream multilayer perceptron classifier:

D. Fairness Evaluation

We assess model fairness by stratifying results across two protected attributes:

- **Sex:** Male vs. Female
- **Age:** Young (<60 years) vs. Old (≥60 years)

We compute the F1-score for each subgroup and calculate the absolute performance gaps between groups to quantify disparities:

$$\Delta_{\text{gap}} = |\text{Metric}_{\text{Group A}} - \text{Metric}_{\text{Group B}}|$$

We report Δ_{gap} for both metrics across both protected attributes.

IV. Experiments

A. Dataset and Preprocessing

We use a subset of the MIMIC-IV ECG dataset, consisting of 795,517 diagnostic ECG signals paired with rhythm interpretations [20]. We select lead-II and preprocessed it via bandpass and notch filtering, then resampled to 1,000 time points. We tokenize notes using BERT's tokenizer, and append age/sex metadata for the augmented setup.

We derive labels by mapping each rhythm to a binary label: "*sinus rhythm*" is considered normal; all other types (e.g., atrial fibrillation, bradycardia, tachycardia) are treated as abnormal.

B. Training and Evaluation

Contrastive pretraining is performed using the AdamW optimizer (learning rate $1e$-4, batch size 256) for 100 epochs. For probing, we train all classifiers using 5-fold cross-validation on extracted ECG embeddings. We report macro F1-score on the held-out test set. For subgroup analysis, performance is measured separately within each age and sex group. All experiments are repeated across both pretraining variants (original and augmented) to assess the impact of including protected variables on model fairness and representation quality.

V. Result

A. Overall Classification Performance

We evaluate the downstream classification performance of ECG embeddings using F1-score for the *abnormal* rhythm class (Class 1). Figure 2 compares the results for models pretrained with and without demographic augmentation using the MLP classifier. Overall, the inclusion of age and sex information in pretraining yielded comparable or slightly improved F1-scores across most subgroups.

B. Effect of Augmentation Across Demographics

When comparing the augmented vs. non-augmented models:

- **Overall performance** is marginally higher for the augmented model, suggesting a slight benefit in global classification accuracy.
- **Sex breakdown:** For female patients, the non-augmented model performs slightly better. However, for male patients, the augmented variant leads to higher F1-score, indicating a potential benefit in male subgroup alignment.
- **Age breakdown:** Notably, the *augmented* model improves performance for the 30–60 age group compared to the non-augmented baseline (0.64 vs. 0.52), while performance for the $\geq$60 group is competitive across both setups.

C. Fairness Implications

The addition of age and sex information during contrastive pretraining appears to reduce performance gaps in certain subgroups, particularly across age groups. While not uniformly beneficial across all subgroups, the results suggest that demographic-informed pretraining can help mitigate underperformance in younger patients, who are often underrepresented in cardiac risk models but that the selection of which must be an informed decision.

D. ECG Latent Embedding

Examining t-SNE plots of ECG embeddings from models pretrained on original text versus those trained on augmented text reveals notable differences in cluster density and separation. Models exposed to augmented text often form more distinct clusters, reflecting a richer representation of underlying signal patterns. In contrast, embeddings derived from models trained solely on the original text can appear more dispersed or show overlapping clusters.

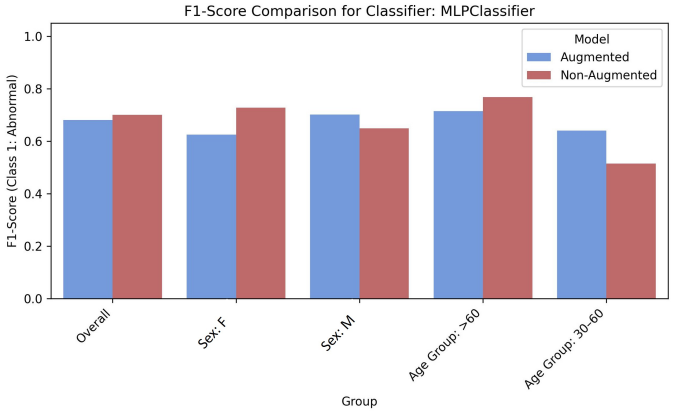

Fig. 2. F1-score comparison for MLP classifier across demographic groups. The "Augmented" model includes age and sex in pretraining text; "Non-Augmented" uses rhythm description only.

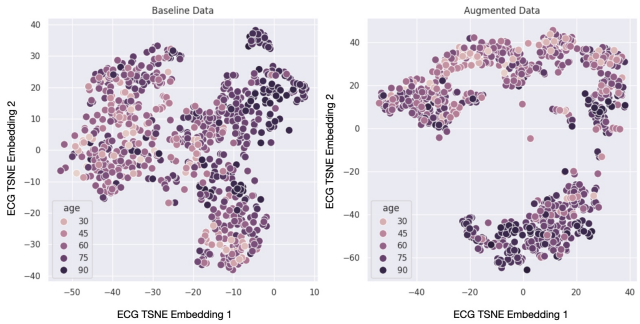

Fig. 3. Visualization of the ECG embedding space with original and augmented ECG text, with each data point color-coded by patient age.

VI. Limitations and Future Work

While our study provides promising insights into fairness-aware multimodal pretraining for ECG analysis, several limitations remain. First, we focus only on two demographic attributes—age and sex. Although clinically important, these variables do not capture the full spectrum of potential bias factors. Future work should consider additional protected attributes such as race, ethnicity, and comorbidities, as well as explore intersectional subgroup analysis. Second, our evaluation relies on linear probing to assess the quality of pretrained

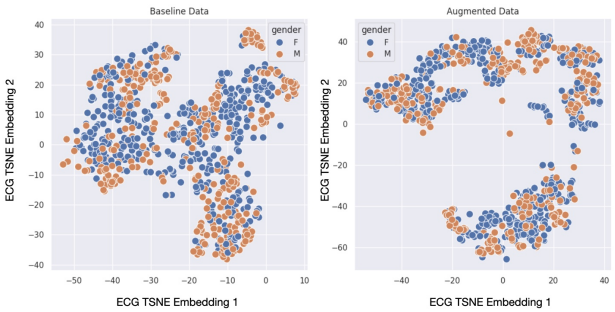

Fig. 4. Visualization of the ECG embedding space with original and augmented ECG text, with each data point color-coded by patient sex.

representations. While this isolates the effects of pretraining, clinical models are often fine-tuned end-to-end. It remains an open question whether fairness improvements persist when transferred to other clinical tasks. We would aim to extend our evaluation to additional datasets in future studies. Finally, while we stratify results by subgroup, our contrastive training objective does not explicitly enforce fairness constraints. Future extensions could incorporate fairness-aware loss terms, adversarial debiasing, or counterfactual contrastive examples to further mitigate subgroup disparities during representation learning. These provide opportunities for future research into equitable, multimodal, and physiologically grounded AI systems for healthcare.

VII. Conclusion

In this work, we investigated whether incorporating protected demographic attributes—age and sex—during contrastive pretraining can improve fairness in downstream ECG classification tasks. Using a CLIP-style multimodal framework that aligns ECG signals with rhythm descriptions, we trained models with and without demographic augmentation. We then evaluated their performance on abnormal rhythm detection using frozen embeddings and linear probing classifiers.

Our results show that demographic-aware pretraining can help reduce performance disparities across subgroups, particularly in age-based cohorts. The augmented model demonstrated improved or comparable F1-scores across most groups, and notably outperformed the baseline in younger patients. These findings suggest that introducing demographic context during representation learning can enhance fairness without compromising overall accuracy.

This work highlights the importance of incorporating fairness considerations early in the model development pipeline. Rather than relying solely on post hoc evaluation or mitigation, we show that pretraining strategies can be explicitly designed to support equitable performance across diverse patient populations.

References

[1] V. Sangha, A. Khunte, G. Holste, B. J. Mortazavi, Z. Wang, E. K. Oikonomou, and R. Khera, "Biometric contrastive learning for data-efficient deep learning from electrocardiographic images," *Journal of the American Medical Informatics Association*, vol. 31, no. 4, pp. 855–865, 2024.

[2] D. Ouyang, J. Theurer, N. R. Stein, J. W. Hughes, P. Elias, B. He, N. Yuan, G. Duffy, R. K. Sandhu, J. Ebinger *et al.*, "Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study," *The Lancet Digital Health*, vol. 6, no. 1, pp. e70–e78, 2024.

[3] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ecg diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107187, 2021.

[4] D. Kaur, J. W. Hughes, A. J. Rogers, G. Kang, S. M. Narayan, E. A. Ashley, and M. V. Perez, "Race, sex, and age disparities in the performance of ecg deep learning models predicting heart failure," *Circulation: Heart Failure*, vol. 17, no. 1, p. e010879, 2024.

[5] P. P. Liang, C. K. Ling, Y. Cheng, A. Obolenskiy, Y. Liu, R. Pandey, A. Wilf, L.-P. Morency, and R. Salakhutdinov, "Multimodal learning without labeled multimodal data: Guarantees and applications," *arXiv preprint arXiv:2306.04539*, 2023.

[6] R. King, T. Yang, and B. J. Mortazavi, "Multimodal pretraining of medical time series and notes," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 244–255.

[7] R. King, S. Kodali, C. Krueger, T. Yang, and B. J. Mortazavi, "An efficient contrastive unimodal pretraining method for ehr time series data," in *Proceedings of the 2024 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2024.

[8] Z. Nowroozilarki, S. Huang, R. Khera, and B. J. Mortazavi, "Non-invasive electrolyte estimation using multi-lead ecg data via semi-supervised contrastive learning with an adaptive loss," in *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2024, pp. 1–8.

[9] M. Pham, A. Saeed, and D. Ma, "C-melt: Contrastive enhanced masked auto-encoders for ecg-language pre-training," *arXiv preprint arXiv:2410.02131*, 2024.

[10] P. Rajpurkar, A. Hannun, A. Y. Ng *et al.*, "Ai-aided ecg interpretation: performance gaps and fairness considerations," *Nature Medicine*, vol. 28, no. 8, pp. 1406–1412, 2022.

[11] H. Kaur, A. Y. Ng, and P. Rajpurkar, "Fairness in cardiovascular risk prediction: examining subgroup disparities and model choices in deep learning using ecg," *NPJ Digital Medicine*, vol. 5, no. 1, pp. 1–10, 2022.

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

[13] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Fairness and disentangled representations in deep learning," *arXiv preprint arXiv:1905.13662*, 2019.

[14] S. Xie, Y. Wu, J. Li, M. Ding, and K. B. Letaief, "Privacy for fairness: Information obfuscation for fair representation learning with local differential privacy," *arXiv preprint arXiv:2402.10473*, 2024.

[15] M. Lin, T. Li, Z. Sun, G. Holste, Y. Ding, F. Wang, G. Shih, and Y. Peng, "Improving fairness of automated chest radiograph diagnosis by contrastive learning," *Radiology: Artificial Intelligence*, vol. 6, no. 5, p. e230342, 2024.

[16] L. Seyyed-Kalantari, H. Zhang, and M. Ghassemi, "Fairehr-clp: Contrastive learning with counterfactual augmentations for equitable ehr representations," *arXiv preprint arXiv:2302.00674*, 2023.

[17] C. Agarwal, P. Robinson, and S.-Y. Lee, "Debias-clr: Debiased contrastive learning of visual representations," in *NeurIPS Workshop on Algorithmic Fairness through the Lens of Time*, 2021.

[18] E. Perez-Alday *et al.*, "The physionet/computing in cardiology challenge 2021: Classification of 12-lead ecgs," in *Computing in Cardiology*, vol. 48, 2021, pp. 1–4.

[19] J. Zheng, C. Ani, I. Abudayyeh, Y. Zheng, C. Rakovski, E. Yaghmaei, and O. Ogunyemi, "A review of racial differences and disparities in ecg," *International Journal of Environmental Research and Public Health*, vol. 22, no. 3, p. 337, 2025.

[20] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.