ICML 2025 Workshop on Actionable Interpretability

Persistent Demographic Information in X-ray Foundation Embeddings: a Risk for a Safe and Fair Deployment in Healthcare

Filipe Santos[1*], Aldo Marzullo[3], Alessandro Quarta[2,5*], João M. C. Sousa[1], Susana M. Vieira[1], Leo Anthony Celi[4], Francesco Calimeri[2**], Laleh Sayeed-Kalantari[6**]

[1] IDMEC, University of Lisbon, Lisbon, Portugal • [2]University of Calabria, Rende, Italy • [3]IRCCS Humanitas Research Hospital, Rozzano, Italy • [4]Massachusetts Institute of Technology, Cambridge, MA, USA • [5] Sapienza University of Rome, Rome, Italy • [6] York University, Toronto, ON, Canada

* Corresponding authors **These authors contributes equally as last

alessandro.quarta@unical.it - filipempsantos@tecnico.ulisboa.pt

Background

Medical imaging foundation models generate vector embeddings (vembs) that improve efficiency but may inadvertently encode sensitive demographic information, raising concerns about:

- **Fairness**: Biased predictions for underrepresented groups
- **Privacy**: Unintended leakage of sensitive attributes
- **Safety**: Perpetuation of healthcare disparities

Research questions

- How extensively is demographic information encoded in chest X-ray embeddings from foundation models?
- How well can we recover Demographic Information from Embedding space?

Methodology & Analytical Framework

Foundation Models

- CXR Foundation: 1376-dimensional embeddings
- BiomedCLIP: 512-dimensional embeddings

Predictive Modeling Assessment:

- Multi-Layer Perceptron (MLP) classifiers
- Stratified 10-fold cross-validation
- Patient-level isolation to prevent data leakage
- Evaluation: F1-score and ROC-AUC

Feature Removal Analysis:

- Input gradient computation
- Feature ranking by gradient magnitude
- Iterative ablation (10%, 20%, 50% of most important features)
- Performance re-evaluation after retraining

Clinical Implications

- **Risk:**
  a. Models may implicitly use demographic proxies, potentially reinforcing healthcare disparities and leading to biased diagnostic decisions

Results

Feature Removal Impact

- Removing up to 50% of most informative features had minimal impact on demographic prediction performance:
  a. Sex prediction ROC-AUC dropped by less than 0.01
  b. Age, ethnicity, and insurance predictions showed negligible changes
  c. Disease prediction performance remained stable
- Implication:
  a. Demographic information is redundantly distributed across the entire embedding space, not confined to specific dimensions.

Table 2. % feature perturbation (demographic prediction VS. disease prediction - mean performance across folds ± standard deviation).

Dataset	Model	Task	Percentage of features modified 0%	1%	5%	10%	20%	50%
MIMIC-CXR	CXR-F	Sex	99.2 ± 0.1	99.2 ± 0.0	99.1 ± 0.0	98.9 ± 0.1	98.7 ± 0.0	97.9 ± 0.1
		Disease*	86.4 ± 0.1	86.4 ± 0.1	86.3 ± 0.1	86.3 ± 0.1	86.4 ± 0.1	86.3 ± 0.1
		Age	88.9 ± 0.2	88.7 ± 0.1	88.6 ± 0.2	88.2 ± 0.1	88.0 ± 0.1	87.4 ± 0.1
		Disease*	86.4 ± 0.1	86.4 ± 0.1	86.3 ± 0.2	86.4 ± 0.1	86.3 ± 0.1	86.4 ± 0.1
		Eth. B.	84.5 ± 0.2	84.2 ± 0.3	83.9 ± 0.2	83.3 ± 0.2	82.6 ± 0.2	80.9 ± 0.2
		Disease*	86.4 ± 0.1	86.3 ± 0.2	86.3 ± 0.2	86.3 ± 0.1	86.3 ± 0.2	86.3 ± 0.1
		Eth. M.	79.9 ± 0.3	79.4 ± 0.3	79.1 ± 0.2	78.7 ± 0.3	78.1 ± 0.2	76.4 ± 0.2
		Disease*	86.4 ± 0.1	86.4 ± 0.1	86.3 ± 0.1	86.3 ± 0.1	86.4 ± 0.1	86.3 ± 0.1
		Insurance	76.2 ± 0.2	75.8 ± 0.2	75.8 ± 0.3	75.8 ± 0.2	75.6 ± 0.2	75.4 ± 0.2
		Disease*	86.4 ± 0.1	86.3 ± 0.1	86.4 ± 0.1	86.4 ± 0.1	86.4 ± 0.1	86.3 ± 0.1
	B-CLIP	Sex	92.6 ± 0.1	92.2 ± 0.0	92.2 ± 0.1	92.1 ± 0.1	92.0 ± 0.1	91.7 ± 0.1
		Disease*	82.9 ± 0.0	82.8 ± 0.1	82.8 ± 0.1	82.8 ± 0.0	82.8 ± 0.0	82.8 ± 0.1
		Age	77.8 ± 0.1	77.0 ± 0.1	77.0 ± 0.1	76.9 ± 0.1	76.8 ± 0.2	76.7 ± 0.1
		Disease*	82.9 ± 0.0	82.8 ± 0.0	82.7 ± 0.0	82.7 ± 0.1	82.7 ± 0.1	82.8 ± 0.1
		Eth. B.	72.4 ± 0.1	72.1 ± 0.3	71.9 ± 0.3	72.0 ± 0.2	71.8 ± 0.3	71.8 ± 0.2
		Disease*	82.9 ± 0.0	82.7 ± 0.1	82.8 ± 0.0	82.8 ± 0.0	82.8 ± 0.1	82.8 ± 0.1
		Eth. M.	69.4 ± 0.4	68.2 ± 0.3	68.1 ± 0.2	68.1 ± 0.4	68.2 ± 0.2	67.8 ± 0.2
		Disease*	82.9 ± 0.0	82.7 ± 0.1	82.7 ± 0.1	82.8 ± 0.1	82.8 ± 0.1	82.7 ± 0.1
		Insurance	69.8 ± 0.1	69.5 ± 0.1	69.5 ± 0.1	69.4 ± 0.2	69.4 ± 0.1	69.1 ± 0.1
		Disease*	82.9 ± 0.0	82.8 ± 0.0	82.8 ± 0.0	82.8 ± 0.0	82.7 ± 0.1	82.8 ± 0.1
CheXpert	CXR-F	Sex	98.7 ± 0.1	98.7 ± 0.1	98.5 ± 0.0	98.3 ± 0.1	98.1 ± 0.0	97.1 ± 0.1
		Disease*	95.5 ± 0.1	95.5 ± 0.0	95.5 ± 0.1	95.5 ± 0.0	95.5 ± 0.0	95.4 ± 0.0
		Age	88.8 ± 0.2	88.5 ± 0.1	88.3 ± 0.1	87.9 ± 0.1	87.7 ± 0.1	86.8 ± 0.1
		Disease*	95.5 ± 0.1	95.5 ± 0.0	95.5 ± 0.0	95.5 ± 0.0	95.5 ± 0.1	95.5 ± 0.0
		Eth. B.	74.3 ± 0.3	74.4 ± 0.3	74.0 ± 0.3	73.3 ± 0.4	72.8 ± 0.1	71.4 ± 0.3
		Disease*	95.5 ± 0.1	95.5 ± 0.0	95.5 ± 0.0	95.5 ± 0.0	95.5 ± 0.0	95.4 ± 0.0
		Eth. M.	74.6 ± 0.7	74.3 ± 0.2	74.3 ± 0.3	73.8 ± 0.3	73.4 ± 0.4	71.8 ± 0.3
		Disease*	95.5 ± 0.1	95.5 ± 0.1	95.5 ± 0.0	95.5 ± 0.0	95.5 ± 0.0	95.4 ± 0.0
	B-CLIP	Sex	90.1 ± 0.1	89.7 ± 0.2	89.6 ± 0.1	89.5 ± 0.1	89.2 ± 0.1	88.9 ± 0.1
		Disease*	91.7 ± 0.0	91.6 ± 0.1	91.7 ± 0.0	91.6 ± 0.0	91.7 ± 0.0	91.7 ± 0.0
		Age	75.9 ± 0.0	75.1 ± 0.2	75.0 ± 0.1	74.9 ± 0.1	74.9 ± 0.2	74.8 ± 0.2
		Disease*	91.7 ± 0.0	91.6 ± 0.0	91.6 ± 0.1	91.6 ± 0.0	91.6 ± 0.0	91.6 ± 0.1
		Eth. B.	66.3 ± 0.1	65.8 ± 0.2	65.9 ± 0.2	65.8 ± 0.2	65.7 ± 0.2	65.4 ± 0.3
		Disease*	91.7 ± 0.0	91.6 ± 0.0	91.6 ± 0.0	91.6 ± 0.1	91.6 ± 0.1	91.6 ± 0.1
		Eth. M.	66.5 ± 0.3	64.6 ± 0.5	64.7 ± 0.3	64.7 ± 0.2	64.7 ± 0.3	64.8 ± 0.5
		Disease*	91.7 ± 0.0	91.6 ± 0.0	91.6 ± 0.1	91.7 ± 0.0	91.6 ± 0.0	91.5 ± 0.1

CXR-F – CXR Foundation, B-CLIP – BiomedCLIP, Eth. – Ethnicity, B. – Binary, M. – Multi-class

* Modifying key features to predict the demographic variable in the same group of tests

Conclusion

- High Recoverability: Machine learning models can reliably predict demographic attributes even when not explicitly included
- Robustness: Demographic encoding shows high resistance to simple feature removal strategies
- Redundancy: Information is distributed across the entire embedding space

Future Directions

- Advanced Debiasing: Develop advanced debiasing and disentangled representation learning methods
- Embedding Auditing: Create techniques to audit and interpret embeddings for demographic content
- Regulatory Frameworks: Establish guidelines for safe deployment of medical AI with demographic awareness