How Hearts Beat: Classifying Failing Hearts into HFrEF vs HFpEF using Image Embeddings from Chest X-Rays

Darcy Kim
Wellesley College
6.8301 Final Report
darcykim@mit.edu

June 3, 2023

Abstract

Heart failure is a major public health challenge with growing costs. Ejection fraction (EF) is a key metric for the diagnosis and management of heart failure. However, estimating EF using echocardiography is expensive and requires extensive expertise. An existing deep learning model is able to identify EF from chest x-rays, which are quick, inexpensive, and require less expertise. However, this model requires enormous computational resource to train and has inequitable performance across race. This paper introduces a model that distinguishes EF in chest-x rays using image embeddings from the MIMIC dataset. Compared to the existing model trained for the same task on raw images, the proposed model performs similarly while significantly reducing the computational resources needed for training and updating. Upon further model analysis based on race, gender, and insurance, the proposed embedding model has a bias performance correlated with the representation in the dataset. Finally, further discussion on model bias in clinical settings puts the model into context of the American healthcare system.

1 Introduction/Related Work

Heart failure (HF) is a progressive chronic condition where the heart is unable to pump enough blood. Generally, there are two type of heart failure: (1) reduced ejection fraction (HFrEF), when the heart cannot contract properly; and (2) preserved ejection fraction (HFpEF), when the heart cannot relax properly. Clinicians need to distinguish the type of heart failure to provide adequate treatment and management.

Currently, cardiologists rely on trans-thoracic echocardiography (TTE) to make this distinction. Unfortunately, in many parts of the world, access to TTE is limited. It requires not only expensive ultrasound equipment, but also specially trained ultrasound technicians and cardiologists for its interpretation. This limits clinicians in lower-resourced settings trying to treat HF.

In contrast with TTE, chest X-ray (CXR) is significantly less expensive and does not require the same level of specialized training to interpret. While the human eye cannot distinguish HFrEF vs HFpEF via CXR, Williams et. al. introduce a deep learning model that can [2]. They studied a dataset of 3488 CXRs from the MIMIC CXR-jpg dataset and established binary classification benchmarks using convolution neural network architectures. Their subsequent analysis shows increasing model sizes from 8M to 23M parameters improved classification performance without overfitting the dataset.

As fantastic as their model is, there are some problems with it. For one, training models on images requires enormous computational resource since a single image contains many features, considering each pixel has multiple channels to describe its color. Models used in clinical settings need to be updated often, so efficiently using computational resources will ensure a sustainable model in practice. Further, the size of their patient cohort, and hence the dataset, is limited. Models tend to perform better when trained on more data.

Another problem with Williams et. al.'s models are their bias. Their model performs significantly better on white populations compared to black populations. One reason for this could be due to the skewed representation in their dataset. There are significantly more white patients than patients of any other race. Another reason for the bias performance could be due to short cut features. Degrave et. al. demonstrated that recent deep learning systems to detect COVID-19 from chest x-rays rely on confounding factors, called short cut features, rather than medical pathology [3]. One such short cut feature could be race, where the model learns the race of a patient and uses this information to provide a diagnosis, instead of the medical signals found in the chest x-ray.

In this work, I seek to address these problems of data size limitations and short cut features. To do so, I will leverage the work of Sellergren et. al., who introduce image embeddings for chest x-rays [1]. They used supervised contrastive [SupCon] learning to generate chest x-ray networks that distinguish normal x-rays from abnormal x-rays. They then showed that freezing the weights of this large base model then finetuning it on specific on various prediction tasks (e.g. airspace opacity, fracture, tuberculosis, and COVID-19 outcomes) yielded competitive results when trained on small datasets, like the one used in this paper. Essentially, the last hidden layer of their SupCon network is a smaller, dense representation of an image, called an image embedding. Using these smaller image embeddings to train networks would significantly reduce computational resource when training and updating models. Further, it might be impossible for models to learn short cut features from these embeddings since each embedding holds significantly less information.

2 Methodology

The methodology investigates if image embeddings can address data size limitations and short cut features in the existing heart failure diagnosis model.

2.1 Data

I used the "Generalized Image Embeddings for the MIMIC Chest X-Ray" dataset [1] released by the authors who invented the embeddings introduced above. These embeddings are produced using the same dataset that the heart failure model introduced above was trained on, and I follow their methodology for generating labels using ICD codes as follows [2]. The images are collected from Electronic Health Records from the Beth Israel Deaconess Medical Center (BIDMC) emergency room from 2011-2016 [4]. The dataset is filtered to only include those CXRs diagnosed with HFpEF or HRrEF, which is determined by their International Classification of Diseases (ICD) codes recorded in MIMIC IV [4]. The patient cohort consists of 2,967 CXR embeddings with 1,246 CXRs diagnosed with HFrEF and the remaining 1,718 CXRs diagnosed with HFpEF. Further patient demographics are discussed in the results section below.

2.2 Model

The model uses image embeddings as input for a binary classification neural network with architecture described in Table 1. There are 839,681 total parameters, 838,145 trainable parameters, and 1,536 non trainable parameters. The model essentially freezes the weights of a SupCon general model then finetunes it for this specific task on this dataset. I used an 80-20 train-test split on the data ensuring their is no data leakage by using the subject ID attached to every image.

Layer Type	Output Size	Params
Input	1376	0
Dense	512	705 204
Batch Norm	512	2048
Dropout	512	0
Dense	256	131 328
Batch Norm	256	1024
Dropout	256	0
Dense	1	257
Sigmoid (Label)	1	0

Table 1: Model Architecture

2.3 Metrics

The model seeks to identify if heart failure is HFrEF or HFpEF. Thus, the metrics of performance, recall, and f1 score for both classes most readily reflects how the model will perform in a clinical setting. These metrics also serve as a basis of comparing the model trained on embeddings to the model trained on images [2] across class, race, and sex.

3 Results / Discussion

Overall, the model I trained on embeddings has comparable performance to the existing model that was trained on images [2]. Thus, this methodology reduces computational resources while preserving performance. Further model analysis broken down by race, sex, and insurance indicates a bias correlated with representation.

3.1 Reducing Data Size

The overall embedding model performance compared to the previous image model performance is summarized in Figures 1 and 2. Performance is calculated for both HFrEF and HRpEF.

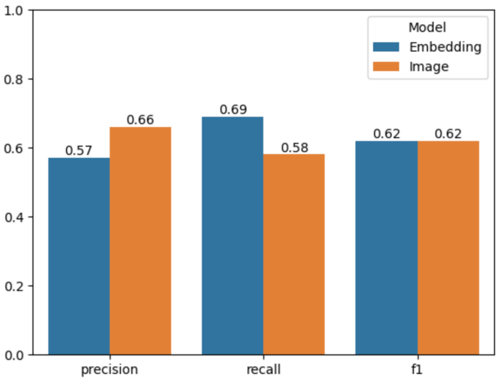

Figure 1: Performance on Reduced EF.

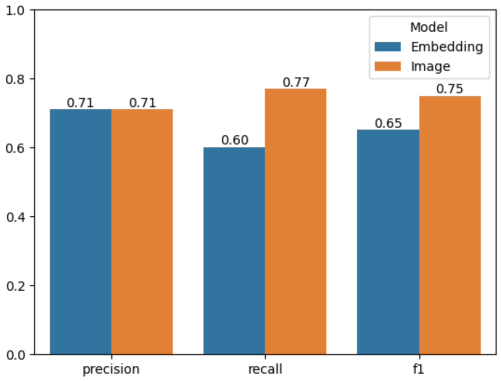

Figure 2: Performance on Preserved EF.

The models have the same f1 score when identifying HFrEF, but the image model has a slightly better f1 score when identifying HFpEF. The proposed embeddings model significantly reduces the amount of computational resources needed to train and update the model while preserving substantial performance. This would allow the model to be more readily implemented into clinical settings where efficient use of computational resources is necessary for the upkeep of the model.

3.2 Short Cut Feature Mitigation

Below is further model analysis investigating bias across race, sex, and insurance. The model bias is then put into the context of representation in the dataset.

3.2.1 Race

Figures 3 and 4 illustrate this embedding model's performance compared to the existing image model's performance across race.

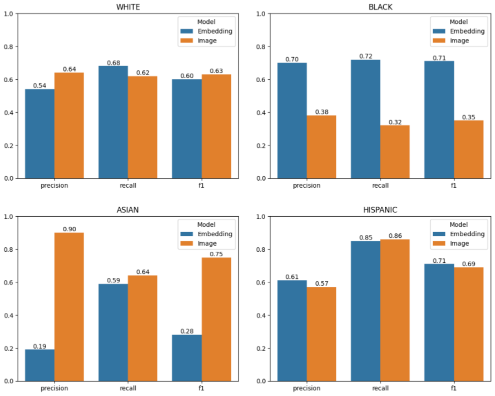

Figure 3: Performance on Reduced EF by race.

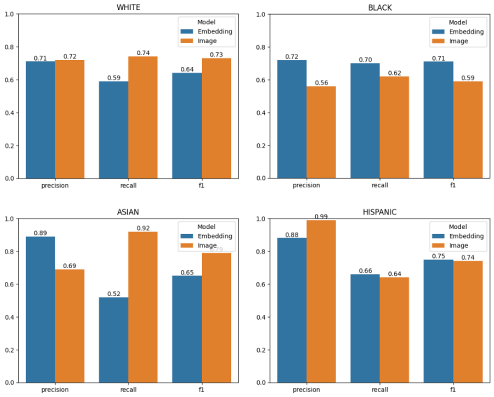

Figure 4: Performance on Preserved EF by race.

Notably, the existing model trained on images performed significantly worse on black populations. This model trained on embeddings greatly improves performance on black populations. In fact, the embedding model performs best on black and hispanic populations for both HFrEF and HFpEF. However, the embedding model performs significantly worse on Asian population for HFrEF. This is the only category where the embedding model performs severely worse.

Investigating the demographic distribution of the dataset provides some insight. The representation of each race in the dataset is illustrated in Figure X. Figure X shows that the dataset was majority white, with black populations being the second majority, then Hispanic, then Asian. The bias performance of this embeddings model more closely reflects the representation of the dataset. Perhaps the embeddings model bias lies more within the representation than the reliance on short cut features.

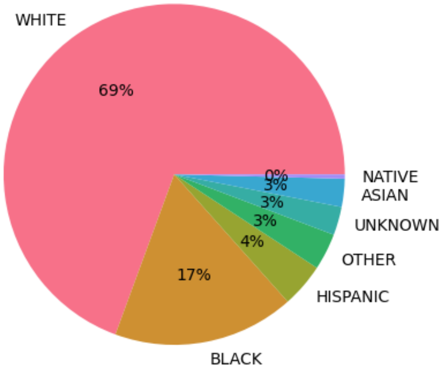

Figure 5: Race distribution in dataset.

3.2.2 Sex

Figures 6 and 7 compare the embeddings model to the image model across sex. The dataset reflects an even 50-50 split across sex. The embedding model has similar results to the image model. Both models perform better on males for HFrEF, and both models perform better on females for HFpEF.

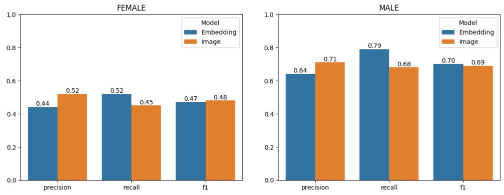

Figure 6: Performance on Reduced EF by sex.

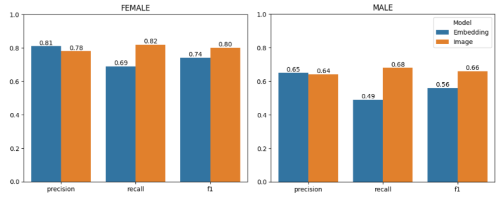

Figure 7: Performance on Preserved EF by sex.

3.2.3 Insurance

Degrave et. al. did not report their image model's performance by insurance. However, insurance is a reflection of socio-economic status and so it is relevant to the interpretation of model bias in clinical settings. In clinical settings, people of higher socio-economic status often have access to better health care, so a model might then perform better for these patients. Figures 8 and 9 illustrate model performance across patients with medicare (i.e. public health insurance for seniors and some people with disabilities), medicaid (i.e. public health insurance for low income peoples), and private insurance.

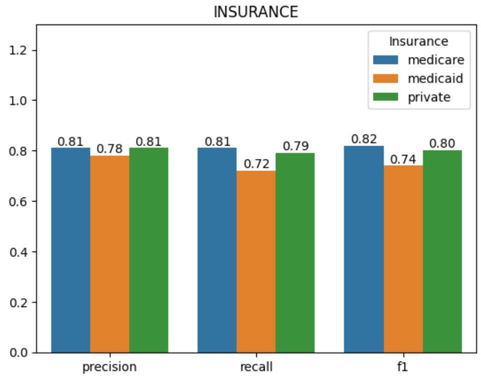

Figure 8: Embedding model performance on Reduced EF by insurance.

The model performance for both HFrEF and HFpEF correlate with the representation in the dataset, which is illustrated in Figure X.

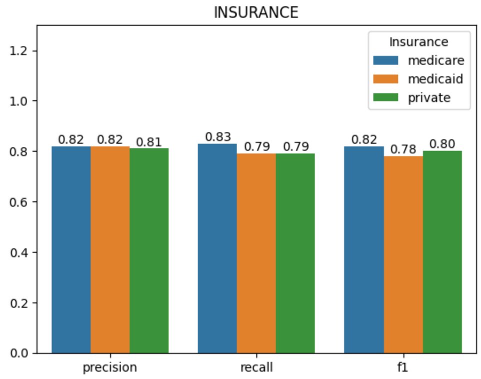

Figure 9: Embedding model performance on Preserved EF by insurance.

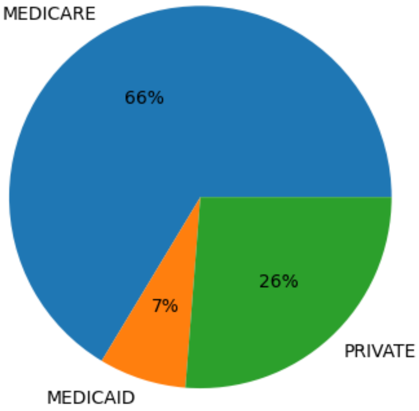

Figure 10: Insurance distribution in dataset.

3.3 Limitations

One limitation is the impossibility of explaining the reason a model performs better or worse on certain demographics. Although we can hypothesize that the embedding model relies less on short-cut features since its bias more closely reflects the representation of the dataset, its impossible to prove this as fact.

Another limitation of this study is the reliance on ICD-10 codes for identifying HFpEF and HFrEF. ICD heart failure codes have weaker accuracy compared to TTE in predicting ejection fraction. The ICD for HF should have reasonable accuracy in the MIMIC database. However, if the ICD HF subtype code is inaccurate or not up to date with the corresponding CXR, this would cause mislabeling in the model's training set.

4 Conclusion

This paper introduces a model that distinguishes HFrEF from HFpEF in chest-x rays using image embeddings. Compared to an existing model trained for the same task on raw images, the proposed model performs similarly while significantly reducing the computational resources needed for training and updating. Upon further model analysis based on race, gender, and insurance, the proposed embedding model has a bias performance correlated with the representation in the datset. This suggests the models bias lies more within an unbalanced dataset than short cut features.

Bias is a problem across any machine learning application that leverages data collected from humans. Although algorithmic innovation can mitigate bias, like image embeddings, there is no such thing as an unbiased model or unbiased dataset. Data is collected from a racist, sexist, classist, ableist society, so these oppressive patterns are inherent in the data itself. A machine learning model will then learn these patterns and perpetuate historical injustices. Providing reparations for the structural violence ingrained in America is necessary when deploying models in clinical settings. This requires intimate collaborations with physicians and the people we actually hope to serve with our technologies. Although current models perpetuate inequity, I'm hopeful that machine learning will provide tools to help doctors that are based on an abstraction of our reality instead of the beliefs of one person or committee.

References

[1] Sellergren, A. B., Chen, C., Nabulsi, Z., Li, Y., Maschinot, A., Sarna, A., Huang, J., Lau, C., Kalidindi, S. R., Etemadi, M., Garcia-Vicente, F., Melnick, D., Liu, Y., Eswaran, K., Tse, D., Beladia, N., Krishnan, D., Shetty, S. (2022). Simplified Transfer Learning for Chest Radiography Models Using Less Data. Radiology, 305(2), 454–465. https://doi.org/10.1148/radiol.212482

[2] Williams, W., Liang, K., Li, Y., Doshi, R., Shah, K., Wang, R., Chen, L., Dagan, A., Max, K., Celi, L. A. (2023). Classifying Chest X-rays with Pulmonary Edema into HFrEF vs HFpEF. Unpublished.

[3] DeGrave, A.J., Janizek, J.D. Lee, SI. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3, 610–619. https://doi.org/10.1038/s42256-021-00338-7

[4] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv. PhysioNet. https://physionet.org/content/mimiciv/1.0/