Analyzing Fairness in Medical Image Representations

Ziyan An*
Department of Computer Science
Vanderbilt University
Nashville, TN 372235
ziyan.an@vanderbilt.edu

Zirong Chen*
Department of Computer Science
Vanderbilt University
Nashville, TN 372235
zirong.chen@vanderbilt.edu

Xia Wang*
Department of Computer Science
Vanderbilt University
Nashville, TN 372235
xia.wang@vanderbilt.edu

Xinmeng Zhang*
Department of Computer Science
Vanderbilt University
Nashville, TN 372235
xinmeng.zhang@vanderbilt.edu *

Abstract

Diagnostic classifiers are deep learning-enabled components integrated into the clinical diagnostic process to aid decisions. Previously, successful diagnostic classifiers have achieved better performances than their human counterparts. However, it is also well-known that diagnostic classifiers are prone to various fairness issues and exhibit biased prediction results. Although plenty of previous work has revealed the unfairness in medical diagnosis [28] or even has provided approaches to mitigate unfairness [3, 16]. Considering the low applicability to image objects and limited computational resources, in this work, we propose a structure to analyze unfairness in medical image representation quantitatively with a light-weighted solution.

1 Introduction

Machine learning methods have been employed in various clinical settings to facilitate clinical decisions [21]. For example, Duke has integrated a deep learning sepsis prediction and management platform, Sepsis Watch, into routine clinical care to quickly and accurately detect patients who have a high risk of developing sepsis [22]. However, as what happens often in data-driven approaches, machine learning algorithms show a high risk of carrying bias and **unfairness** along the way [17]. With the quick integration of machine learning methods into clinical care, it is essential to audit potential problems that these methods induce and their impacts on patients [26]. However, those risks of potential unfairness are inherited with the application of machine learning algorithms in medical image processing [15, 13]. Fairness in medical images is defined by [15] as "*the existence of a complex causal graph with partially observed and potentially confounded observations, sensitive protected attributes can leak undesired information into a classification task*". If those medical image processing algorithms are further applied to medical diagnosis, it will raise more serious consequences, as those sensitive protected attributes can be gender, age, or race. This is considered unacceptable to diagnose with potential discrimination on those attributes. For example, previous studies have found that multiple deep learning methods to predict 14 diagnostic labels in 3 public chest X-ray datasets have differences in true positive rates (TPR) (i.e., TPR disparity) among patients grouped by sex, age, race, and insurance type [23].

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

One common solution to mitigate unfairness in machine learning is to obtain a fair **representation** of each input with the help of pretrained models [12, 9]. Further, in this image-related task, we accept this setting – no image is passed in directly since [28] indicates that directly inputting medical images can raise a high risk of unfairness. Instead, we obtain the image representations using pre-trained models. [16] has proposed methods to help analyze unfairness in image presentations. However, it trains huge VAE-variants based on image inputs. This approach is considered time-consuming especially when the input images are medical images which often are larger than normal images. In this work, we create a new solution to measure unfairness in medical diagnosis by task-wise analysis based on obtained representations. In the task-wise analysis, we design two separate downstream tasks and make two assumptions for the fairness analysis. From the evaluation results, the unfairness in medical diagnosis is successfully quantified.

We summarize the major contributions of this paper as follows:

- We propose a new solution to analyze unfairness in medical diagnosis based on image presentations. This approach is also considered light-weighted compared to existing ones. This proposed approach is proved to be effective in two commonly-used medical datasets: COVID-CT-MD and PAPILA.
- We design task-wise fairness analysis in both COVID-CT-MD and PAPILA datasets. The fairness analysis methods are based on the performance differences of real tasks.

Paper organization: in the rest of the paper, we provide an overview of our proposed method in Section 2, and present the technical details in Section 3. We then present the evaluation results in Section 4, discuss the related work in Section 5 and draw conclusions in Section 6.

2 System Overview

Our proposed system contains three key components, as shown in Figure. 1: (1) data pre-processing; (2) representation obtaining; (3) downstream task & fairness analysis. After data pre-processing, the inputs to our system will be well-formatted into $[I, A_{sen}, A_{insen}]$. I represents initial images in the dataset, A_{sen} and A_{insen} represent sensitive and insensitive attributes respectively. The representation obtaining module takes images and sensitive attributes as input, with an option to encode sensitive attributes or not, and outputs image embeddings using pretrained models and their initial weights. We design two downstream tasks and further analyze task-wise fairness. Those two tasks are diagnosis prediction and sensitive attribute prediction. In diagnosis prediction, the inputs are $[I_{emb}, A_{sen}]$, where I_{emb} represents image embeddings and A_{sen} can be turned on or off. The output is the diagnosis result, in this case, it will be a binary result representing whether the patient has a positive diagnosis. In sensitive attribute prediction, the input is $[I_{emb}]$ alone, and the output is set to A_{sen}, which means sensitive attributes.

3 Methods

In this section, we present the major components of our proposed system 1. We first introduce the data preprocessing methods we applied to have medical images well-formatted as inputs. We then present the representation obtaining module. Last we show the details of two downstream tasks and their task-wise fairness analysis.

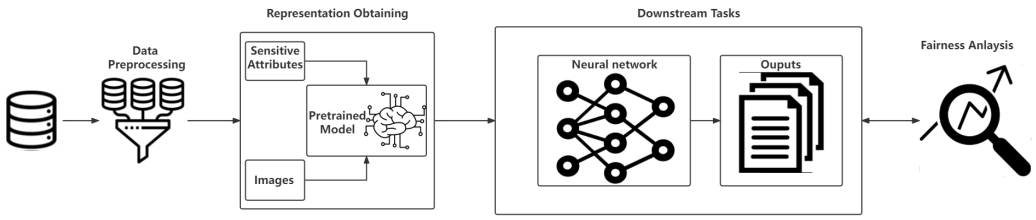

Figure 1: System Overview

3.1 Preprocessing

For data preprocessing procedures, we normalized the images and resized them to smaller sizes. We randomly split the whole dataset into training/validation/testing sets with a proportion of 80/10/10. Age is transformed into binary categories: equal or larger than 60 years old or younger than 60 years old.

3.2 Representation obtaining

Our intuition for obtaining image representation is inspired by most Natural Language Processing (NLP) tasks. Pretrained models like BERT [5] or GPT [19] are often used as representations of input tokens, phrases, or sentences [18, 25, 20]. We apply similar ideas on images, however, unlike [24], we are not retraining the existing model or training a new model from scratch. Instead, we directly use the outputs from pretrained models as image representations. ResNet-18 [10] (along with its variant) and Visual Transformers [6] (ViT) are selected as the pretrained models for representation obtaining. The application of skip connection in ResNet-18 allows information to be passed even if the network is deep. ViT treats images as separate patches and injects prior knowledge by pertaining. We leverage the characteristics of both powerful pretrained models to gain our image representations.

Before further experiments, we justify our idea about using the outputs of pretrained models as representations by feeding images with different objects and analyzing the similarity (see details Fig.2 and Fig.3). The red dots show the dog ResNet representations and the blue ones are cat ResNet representations. We can clearly tell the differences between these two clusters although they are not widely separated.

3.3 Downstream tasks

The obtained image representations are passed to a neural network for the downstream tasks. We experimented with two neural network structures for fine-tuning: a fully connected MLP with a sigmoid activation function and Convolutional Neural Network (CNN). We would fine-tune network structures and hyperparameters, which include epoch, optimizer, and weight initializer, for each specific task and dataset.

3.3.1 Diagnosis prediction

The diagnosis prediction uses image representation embedding as input and predicts the diagnosis label. The assumption of fairness on this prediction task is that under each subgroup, for instance, in each gender subgroup, the prediction performance should be very similar. Also, the fairness metrics

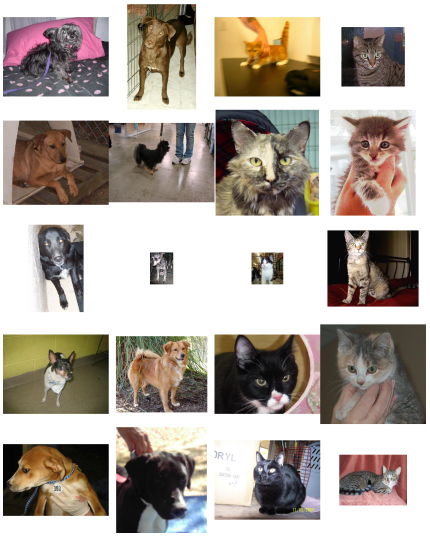

Figure 2: Images of cats and dogs

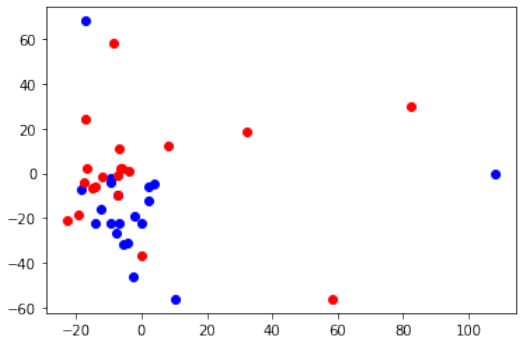

Figure 3: Representation of those images

will be defined in the fairness analysis section. In the diagnosis prediction task, CNN and MLP will be used as the classification models, which not aiming to pursue outstanding results, but to match a baseline performance and to verify whether the task implies unfairness.

3.3.2 Sensitive attribute prediction

The sensitive attributes are gender (female and male) and age (equal or older than 60 years old and younger than 60 years old). We used the obtained representation as input to the model and trained a model for each sensitive attribute. We use the Area Under the curve of the Receiver Operating Characteristic (AUROC) for model evaluation since this metric does not depend on threshold selection. AUROC score means the probability that a randomly selected patient who has an attribute will have a higher predicted risk score than a randomly selected patient who does not have this attribute. An AUROC of 0.5 corresponds to random choice, which shows that a model has no ability in the prediction task.

3.4 Fairness analysis

For the diagnosis prediction task, fairness is measured by the equality of different sensitive subgroups. Further, given the image representation, we predict its diagnosis label and expect the model will not bias towards a certain subgroup of people. Here we consider the two most salient fairness definitions for healthcare, i.e., group fairness and Max-Min fairness. Firstly, we measure the performance gap in diagnosis AUC between the advantaged and disadvantaged subgroups as an indicator of group fairness, the smaller group fairness indicator is better. Secondly, Max-Min fairness treats the model that reduces the worst-case error rates as the fairer one, the larger Max-Min fairness indicator is better.

For the sensitive attribute prediction, fairness is measured by the less mutual information between the image representation embedding and the sensitive attribute. Here we directly use the image representation embedding as input to predict the sensitive attribute. In case of the sensitive attribute prediction is random or not convergent, it means less mutual information between the image representation embedding and the sensitive attribute, which is the fairness pursued in this scenario.

4 Evaluation

We conducted the experiments with two public datasets: PAPILA [14] and COVID-CT-MD [1]. We choose these two datasets since they are publicly available and contain similar sensitive attributes for comparison. In both of the two datasets, the evaluation summary shows an imbalance distribution in diagnosis labels and sensitive attributes of gender and age. Also, for subgroups, the diagnosis rate is different under sensitive attribute subgroups. For example, in two gender subgroups, the male subgroup and female subgroup have different diagnosis rates, shown in Table 1 and Table 2.

4.1 PAPILA dataset characteristics

PAPILA dataset contains fundus images and clinical data of both eyes of patients for glaucoma assessment. The dataset contains records of 244 patients. Each record consists of fundus images

Table 1: Evaluation Summary of PAPILA

PAPILA		gender				age				data size
		Male		Female		<60		>=60		
Train	healthy	72	48	123	89	84	63	111	74	195
	unhealthy		24		34		21		37	
Test	healthy	11	4	14	13	15	12	10	5	25
	unhealthy		7		1		3		5	
Validation	healthy	10	5	14	11	10	7	14	9	24
	unhealthy		5		3		3		5	
Train (%)	healthy	36.92%	66.67%	63.08%	72.36%	43.08%	75.00%	56.92%	66.67%	-
	unhealthy		33.33%		27.64%		25.00%		33.33%	
Test (%)	healthy	44.00%	36.36%	56.00%	92.86%	60.00%	80.00%	40.00%	50.00%	-
	unhealthy		63.64%		7.14%		20.00%		50.00%	
Validation (%)	healthy	41.67%	50.00%	58.33%	78.57%	41.67%	70.00%	58.33%	64.29%	-
	unhealthy		50.00%		21.43%		30.00%		35.71%	

Table 2: Evaluation Summary of COVID_CT_MD

COVID_CT_MD		gender				age				data size
		Male		Female		<60		>=60		
Train	healthy	147	60	97	50	179	81	65	29	244
	unhealthy		87		47		98		36	
Test	healthy	18	10	13	6	25	11	6	5	31
	unhealthy		8		7		14		1	
Validation	healthy	18	5	12	5	19	5	11	5	30
	unhealthy		13		7		14		6	
Train (%)	healthy	60.25%	40.82%	39.75%	51.55%	73.36%	45.25%	26.64%	44.62%	-
	unhealthy		59.18%		48.45%		54.75%		55.38%	
Test (%)	healthy	58.06%	55.56%	41.94%	46.15%	80.65%	44.00%	19.35%	83.33%	-
	unhealthy		44.44%		53.85%		56.00%		16.67%	
Validation (%)	healthy	60.00%	27.78%	40.00%	41.67%	63.33%	26.32%	36.67%	45.45%	-
	unhealthy		72.22%		58.33%		73.68%		54.55%	

and clinical information about the age and gender of the patients and medical test results including refractive error, crystalline lens, IOP of both eyes, corneal thickness, axial length, and mean defect of both eyes. Fundus images are 2576 x 1934 dimensions. Patient diagnoses contain three categories: healthy, glaucoma, and suspicious. Figure. 4 is an example of our input image.

170 patients are labeled as healthy and 74 patients are labeled as glaucoma or suspicious. Among healthy patients, there is a relatively large discrepancy in gender composition, which 66% of patients being female. 52% of patients are older than 60 years old. Among glaucoma and suspicious patients, there is not a large difference between gender where 51% of patients are female. Patients who are older than 60 years old consist of 64% of glaucoma and suspicious patients.

We defined binary labels to be 1 for patients who are diagnosed with glaucoma or suspicious of glaucoma. We only used the images from patients' left eyes for the prediction to avoid the effect of the data concatenation strategy. Glaucoma often affects both eyes, so we assume this experiment design would not worsen the model's prediction ability.

4.2 PAPILA diagnosis prediction

Using the pre-trained ViT model, each image representation is a tensor shaped 197x768. Using the flattened embedding as input, which has a dimensionality as high as about 150,000, we use net structured classifier to implement this prediction task. However, under scaling preprocessing, different batch size training, and different net structures, the net structured classifier failed to give a reasonable prediction result, mostly predicting as healthy. The reason why it failed to give reasonable diagnosis predictions may be due to the high dimensionality of image embedding.

4.3 PAPILA sensitive attribute prediction

To investigate the ability of deep learning models to detect sensitive attributes (gender and age), we developed one model for the detection of each attribute. We extracted a representation for each image from the pre-trained ViT feature extractor. The image representation has the shape of (197x768). Then we flatten the 2-dimensional representation and input the image representation into a neural network with 4 fully-connected layers and a sigmoid activation function. After model training, we evaluated

Figure 4: PAPILA dataset examples

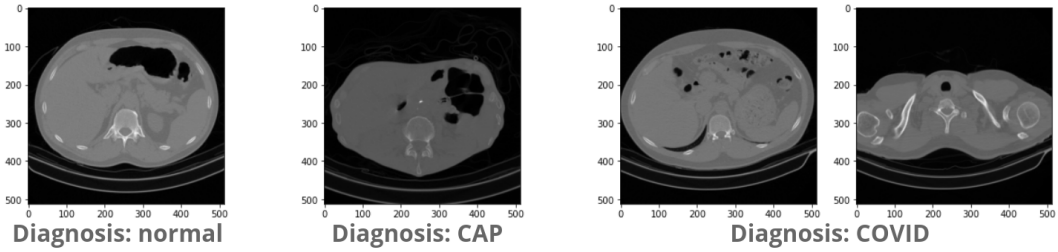

Figure 5: COVID-CT-MD dataset examples

the AUROC score on the test dataset. For the patient gender prediction task, the model achieves 0.66 AUROC. The performance for the patient age prediction task has AUROC=0.58. Though the models do not have a very high ability in predicting, the prediction is not random. Therefore, the experiments demonstrate that image representations contain information about patients' sensitive attributes.

4.4 COVID-CT-MD dataset characteristics

COVID-CT-MD Dataset contains CT scans of the lung of 305 patients. Out of the entire dataset, 169 were diagnosed as COVID-19 positive, 60 patients were diagnosed with Community-Acquired Pneumonia (CAP), and 76 patients were diagnosed as healthy. Figure. 5 shows raw images of COVID-positive cases, CAP cases, and healthy cases, respectively. All images are obtained from one medical imagining center, and the labels are obtained by the common agreements of three radiologists. While lobe-level and slice-level labels are included in the dataset, we use the patient-level dataset for simplicity. It is important to note that although CT imaging is not the gold standard for COVID-19 diagnosis, previous work has found positive correlations between CT scan images and the severity of COVID infections.

To preprocess CT images, we first normalize, crop, and resize the images to 244×244, then take the center 80 slices for each 3D image. We then apply a random flip transformation to every image in the dataset. The preprocessed images are fed into a pre-trained 3D ResNet-18 model. We select patient age and gender as the sensitive attribute and constructed binary labels for diagnostic classifiers and fairness evaluation. CAP and healthy patients are classified as "non-COVID cases", as opposed to "COVID cases". Moreover, patients over 60 years old and patients under 60 years old are split into two groups. We apply an 80/10/10 train/val/test split.

Moreover, we provide preliminary sensitive distribution of the COVID-CT-MD dataset to provide a basic understanding of the dataset structure. Overall, we observe 122 (40%) female and 183 (60%) male patients. Out of all 305 patients, 223 (73%) patients are under 60 years old, while 82 (27%) patients are over 60 years old.

4.5 COVID Diagnosis Prediction

To obtain image embeddings, we feed preprocessed images to pre-trained 3d ResNet-18 and output the results from the `AvgPool` layer. The embedding size is (1×512). Then, we develop a 1d convolutional neural network as a diagnostic classifier, which takes in the image embeddings and predict diagnosis. We find the AvgPool layer in ResNet tends to produce similar outputs for lung CT scans, which is likely due to the fact that the model was previously trained on the Kinetics 400 dataset, which does not resemble medical images. Therefore, we again apply standard scaling to image embeddings before training the diagnostic classifier.

We use two 1D convolution layers with ReLU activation and max-pooling, followed by three fully-connected layers with ReLU activation. The last fully-connected layer is activated with a Sigmoid

Table 3: Diagnosis Prediction Performance Summary of COVID-CT-MD

Batch Size	Resampling	ACC	Group Fairness	Subgroup Estimator AUC	
				Male-advantaged	Female-disadvantaged (Max-Min Fairness)
1	None	0.48	**0.26**	0.6	0.33
16	None	**0.65**	0.28	**0.71**	**0.43**
16	Oversample-remain female diagnosis rate	0.55	0.27	0.64	0.38
16	Oversample-change female diagnosis rate to match male	0.58	**0.26**	0.65	0.38

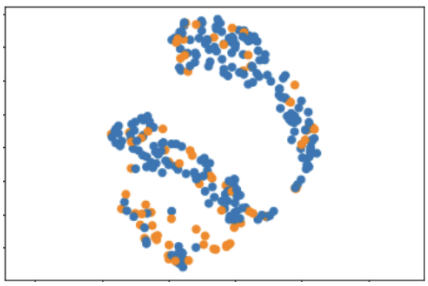

Figure 6: UMAP: binary age

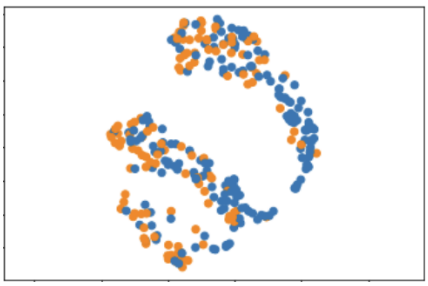

Figure 7: UMAP: gender

layer. We use binary cross entropy loss and Adam optimizer with a learning rate of 1e-4 and the weight decay parameter (L2 weight penalty) is set to 5e-5.

The best result of prediction is from not using resampling and using a larger batch size when training. However, this leads to the biggest unfairness of group fairness. Also, in both the Group Fairness indicator and Max-Min Fairness indicator, we can clearly find unfairness in the diagnosis prediction task. Furthermore, the result is shown in Table 3, which exhibits that only using resampling, not only could not get better fairness but also damage the prediction performance.

4.6 COVID Sensitive Attribute Prediction

To investigate how much information is contained in the image embedding outputs of ResNet-18, we train two additional classifiers using the 1d convolutional network described above. To measure the classifier performance, we use the AUROC metric that describes the model's ability to discriminate between different classes. A higher AUROC score means a better classifier. We obtained a 0.8120 AUROC score in predicting patient gender using representations of CT scans, while we obtained a 0.6467 AUROC score in discriminating patient age. Moreover, Figure. 6-7 shows UMAP visualizations using binary age and gender labels, respectively. We can observe slight separations between the two classes.

4.7 Discussion

In this section, we want to further discuss and reiterate our motivation for the sensitive attribute prediction task. The experiments from [28] show that the performance gap widely exists for patients' age and gender subgroups. As discussed in the previous sections, the image representation contains information on patients' age and gender. This result does not show unfairness since people's eye structure and texture could change as they age. Our reason for doing this prediction task is to investigate which part of the AI algorithm contains patients' sensitive attributes and how much information is pertained without explicitly using the sensitive attributes as input to the downstream prediction tasks.

5 Related Work

5.1 Unfairness Measurement and Quantification

Researchers have developed systems for fairness analysis. TPR Disparity Consider a classification problem, one metric is to compare the *TPR disparity* between subgroups [23]. Let $G_1, G_2, ..., G_m$ denote m subgroups from N samples. For a single group G_i, TPR_{G_i} is calculated as $\frac{TP_i}{TP_i+FN_i}$. Let $\widetilde{TPR} = Median(G_1, G_2, ..., G_m)$ denote the median TPR of all subgroups, the TPR disparity for G_i is $TPR_d = TPR_{G_i} - \widetilde{TPR}$. Demographic Parity Consider a classifier f_θ trained on examples in the following format (X, Y, A) where X is the input feature, Y is the ground-truth label, and A is the protected attribute that can either be included in X or not, demographic parity (DP) measures statistical dependency of prediction $f_\theta(X)$ on A. Formally, we say the classifier f_θ is *fair* if $P[f_\theta = \hat{y} \mid A = a] = P[f_\theta = \hat{y}]$ [2]. However, in our task, although those metrics are applicable,

our methods are still considered better since they are more intuitive and convincing while dealing with real medical diagnosis problems.

5.2 Unfairness Mitigation and Removal

Unfairness removal (or mitigation) tricks like [7, 4, 11] modify the input values to lower the unfairness risk in downstream tasks. However, in medical diagnosis, editing image input could lead to misdiagnosis which further introduces health crises to patients. Other tricks like [27, 8, 24] apply auto-encoders to encode the image, which is similar to our idea, however, we consider ours computationally easier, since no new upstream representation models need to be trained.

6 Conclusion

In conclusion, our work used two public datasets from two different domains. We utilized effective/state-of-the-art models to learn image representations and use these representations for downstream diagnosis prediction tasks and sensitive attribute prediction tasks. We found performance differences between patient subgroups and demonstrated that sensitive attributes are encoded in the representation phases. It is important for future work to evaluate representation fairness to lessen the performance gaps between patient subgroups.

References

[1] Parnian Afshar, Shahin Heidarian, Nastaran Enshaei, Farnoosh Naderkhani, Moezedin Javad Rafiee, Anastasia Oikonomou, Faranak Babaki Fard, Kaveh Samimi, Konstantinos N Plataniotis, and Arash Mohammadi. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Scientific Data*, 8(1):121, 2021. 4

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018. 7

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019. 1

[4] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017. 8

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015. 8

[8] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 8

[9] Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*, 2022. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012. 8

[12] Yannik Keller, Jan Mackensen, and Steffen Eger. Bert-defense: A probabilistic model based on bert to combat cognitively inspired orthographic adversarial attacks. *arXiv preprint arXiv:2106.01452*, 2021. 2

[13] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017. 1

[14] Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific data*, 9(1):291, 2022. 4

[15] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[16] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 1, 2

[17] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 1

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[21] Mark P. Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin M Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. 2020. 1

[22] Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O'Brien. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med Inform*, 8(7):e15182, Jul 2020. 1

[23] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew B. A. McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *CoRR*, abs/2003.00827, 2020. 1, 7

[24] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019. 3, 8

[25] Shufan Wang, Laure Thompson, and Mohit Iyyer. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *arXiv preprint arXiv:2109.06304*, 2021. 3

[26] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *medRxiv*, 2022. 1

[27] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. 8

[28] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022. 1, 2, 7

A Appendix

In this work, we also have some lessons gained from the process:

1. Pre-trained embeddings are very similar in both datasets. For instance, `torch.cosine_similarity()` shows the similarity of embeddings of COVID_CT_MD dataset could be as high as 0.9996. Using embedding scaling could help improve the performance of the prediction tasks.
2. Pre-trained ViT model gives the size of the image embedding as `torch.Size([1, 197, 768])`, using flatten method and simply using an MLP model could not give a usable diagnosis prediction and the training processing was very time-consuming.
3. We can clearly find unfairness in both diagnosis prediction and sensitive attribute prediction tasks, which points out that it should be cautious to use the pre-trained model to get embedding to do some downstream tasks, cause it can easily cause unfairness. However, improving fairness is hard. For example, we failed when only using a simple resampling method.