Using Self-supervised Learning Can Improve Model Fairness

Sofia Yfantidou*
syfantid@csd.auth.gr
Aristotle University of Thessaloniki
Greece

Dimitris Spathis†
dimitrios.spathis
@nokia-bell-labs.com
Nokia Bell Labs
United Kingdom

Marios Constantinides†
marios.constantinides
@nokia-bell-labs.com
Nokia Bell Labs
United Kingdom

Athena Vakali
avakali@csd.auth.gr
Aristotle University of Thessaloniki
Greece

Daniele Quercia
daniele.quercia
@nokia-bell-labs.com
Nokia Bell Labs
United Kingdom

Fahim Kawsar
fahim.kawsar
@nokia-bell-labs.com
Nokia Bell Labs
United Kingdom

arXiv:2406.02361v1 [cs.LG] 4 Jun 2024

ABSTRACT

Self-supervised learning (SSL) has become the de facto training paradigm of large models, where pre-training is followed by supervised fine-tuning using domain-specific data and labels. Despite demonstrating comparable performance with supervised methods, comprehensive efforts to assess SSL's impact on machine learning fairness (i.e., performing equally on different demographic breakdowns) are lacking. Hypothesizing that SSL models would learn more generic, hence less biased representations, this study explores the impact of pre-training and fine-tuning strategies on fairness. We introduce a fairness assessment framework for SSL, comprising five stages: defining dataset requirements, pre-training, fine-tuning with gradual unfreezing, assessing representation similarity conditioned on demographics, and establishing domain-specific evaluation processes. We evaluate our method's generalizability on three real-world human-centric datasets (i.e., MIMIC, MESA, and GLOBEM) by systematically comparing hundreds of SSL and fine-tuned models on various dimensions spanning from the intermediate representations to appropriate evaluation metrics. Our findings demonstrate that SSL can significantly improve model fairness, while maintaining performance on par with supervised methods—exhibiting up to a 30% increase in fairness with minimal loss in performance through self-supervision. We posit that such differences can be attributed to representation dissimilarities found between the best- and the worst-performing demographics across models—up to ×13 greater for protected attributes with larger performance discrepancies between segments.

Code: https://github.com/Nokia-Bell-Labs/SSLfairness

ACM Reference Format:
Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2024. Using Self-supervised Learning Can Improve Model Fairness. In *Proceedings of ACM Knowledge Discovery and Data Mining Conference (KDD '24).* ACM, New York, NY, USA, 12 pages. https://doi.org/X

*Work done at Nokia Bell Labs.
†Also affiliated with the University of Cambridge, UK.

KDD '24, August 25–29, 224, Barcelona, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/X

1 INTRODUCTION

Self-supervised learning (SSL) has emerged as a dominant training paradigm for large models, involving unsupervised pre-training followed by supervised fine-tuning using domain-specific data and labels. SSL has proven its performance robustness and beyond state-of-the-art capabilities mainly in the areas of computer vision (CV) [5] and natural language processing (NLP) [8] research. Inspired by such efforts, which leverage massive amounts of unlabeled data, many research communities that deal with human-centric data have swiftly recognized the potential of self-supervision. SSL exploitation is promising to explore extensive unlabeled data, effectively complementing small, labeled domain datasets [53]. As a case in point, in the healthcare domain, leveraging this wealth of unlabeled information can uncover intricate physiological and behavioral patterns at an unprecedented scale, offering novel insights into personalized and proactive healthcare [36].

Due to the recency of SSL adoption for human-centric, multimodal data, such as time-series, performance metrics, such as accuracy scores, are typically used as the main evaluation criteria. Yet, a performance-centric evaluation approach can result in discriminatory impacts when comparing across different demographics. For instance, in the context of supervised learning, Kamulegeya et al. [23] found that neural network algorithms trained to perform skin lesion classification showed approximately half the original diagnostic accuracy on black patients. At the same time, people of color are consistently misclassified by health sensors such as oximeters as they were validated on predominantly white populations [47].

Preliminary evidence suggests that SSL models may avoid such pitfalls due to their pre-training without (potentially) biased human annotations [38]. Similarly, self-supervision has demonstrated superiority in key aspects of Data-centric Machine Learning (ML), namely a subset of Responsible ML, emphasizing data quality, such as robustness and uncertainty estimation [19]. Yet, comprehensive efforts to compare the fairness of supervised and SSL models are lacking. Note that group fairness assessments typically look for accuracy disparities among diverse protected attributes, namely sensitive personal characteristics, such as race or gender, that are legally safeguarded from discrimination.

In this work, we aim to bridge this gap by introducing a five-stage fairness assessment framework (Figure 1) for SSL. Our framework encompasses dataset requirements definition, modeling, fine-tuning

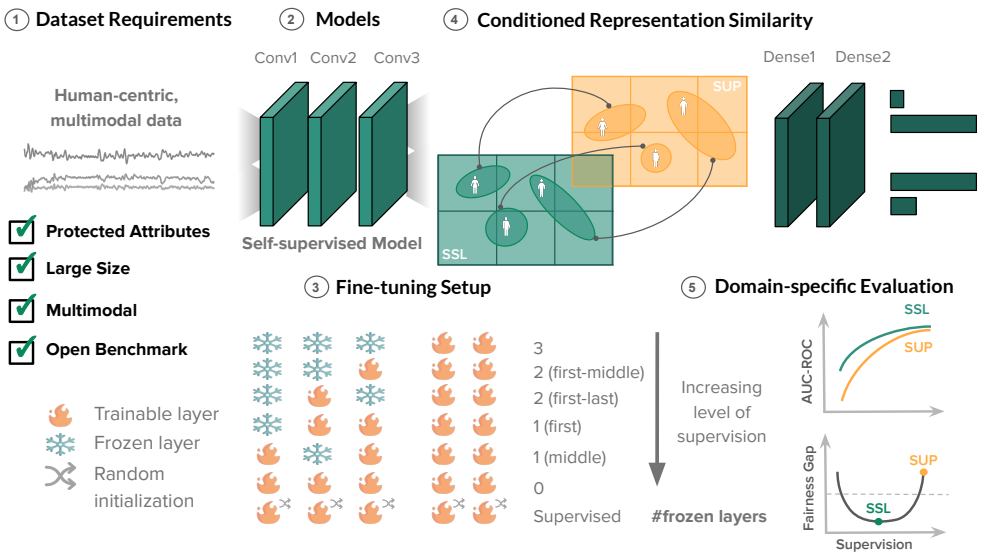


Figure 1: Overview of the proposed fairness assessment framework for SSL. Starting with benchmark selection, we systematically study the impact of fine-tuning on fairness through a novel combination of evaluation and representation learning metrics.

setup, representation similarity, and domain-specific evaluation considerations. As such, it facilitates an examination of how SSL fine-tuning affects fairness compared to supervised alternatives, focusing on both outcomes and representations. We hypothesize, that SSL models will exhibit less bias given that their representations are only partially affected by labels, which may comprise biases steered by the downstream tasks. Using SSL may seem less prone to bias, but concluding it is inherently fair oversimplifies. The pre-training phase can still encode biases from data distributions. Fine-tuning using labeled data can lead to bias amplification. The influence of contrastive objectives on fairness, versus the role of design choices like data augmentation, remains unexplored. These key aspects make the assessment of our hypothesis non-trivial. In detail, we make four contributions:[1]

(1) Moving away from conventional performance-centric assessments, we introduce a fairness assessment framework for SSL, integrating fairness metrics into our methodology to evaluate how fine-grained differences in the layer, model, and metric level between supervised and SSL models affect model outcomes and representations (§3).
(2) We conduct a systematic comparison of more than 100 models with various levels of supervision and fine-tuning on three large real-world benchmarks and tasks (§4). To foster reproducibility we make our code publicly available.[2]
(3) We show that SSL yields smaller performance discrepancies between groups, while performing on par with supervised models across datasets. More notably, we observe up to a 30% increase in fairness, accompanied by only a 2% loss in performance for certain SSL fine-tuning strategies. Similarly, the SSL model shows quicker fairness gains than the supervised one as a function of limited training data (§5.2).
(4) In light of these results, we compare learned representations using the latent similarity between supervised and SSL models, which reveals discrepancies in the latent space across different demographic groups. Specifically, the larger the performance gap between segments the larger the representation similarity gap (up to ×13 greater) between the SSL and the supervised models (§5.3).

[1]Note that some results appeared as a workshop paper at the HCRL (AAAI 2024) [62]
[2]Code: `https://github.com/Nokia-Bell-Labs/SSLfairness`

2 BACKGROUND & RELATED WORK

2.1 Bias & Fairness in Machine Learning

There are two opposing perspectives when quantifying group fairness, i.e., statistical parity for individuals belonging to different protected groups [10], in ML research: "We're All Equal" (WAE) and "What You See Is What You Get" (WYSIWYG) [11, 60]. The WAE perspective assumes equal ability across groups to perform a task and is closely related to treating equals equally. On the other hand, the WYSIWYG viewpoint assumes that the data itself reflects a group's ability with respect to the task, and thus, unequals should not be treated equally. Different fairness metrics quantify each perspective [12]; demographic parity metrics, such as disparate impact and statistical parity difference, quantify WAE. Equality of odds metrics, such as average odds and average absolute odds difference, quantify WYSIWYG. However, the choice of metric is often guided by the question "What is the consequence of the predictive outcome?." Equality of opportunity metrics, such as false negative rate, and false positive rate ratios, find common ground between the two perspectives. To capture the different perspectives, in this work, we utilize a combination of metrics, as discussed in §4.

2.2 Fairness in Self-Supervised Learning

While SSL methods (e.g., SimCLR [5], BYOL [15], Masked Autoencoders [18]) have seen widespread use in CV [24], NLP [26], and

audio [42], they have also been validated in multimodal, human-centric data; yet, the area remains under-explored [16, 50].

Existing works have extensively benchmarked SSL algorithms across domains, primarily focusing on performance metrics. However, limited attention has been given to evaluating fairness in SSL methods, particularly for multimodal, human-centric data. For example, in the healthcare setting, SSL has been applied to online patient monitoring [59], Atrial Fibrillation detection [54], mortality or decompensation prediction [17], maternal and fetal stress detection [44], and human-activity recognition [53], among others. Yet, the above works focus on performance-centric assessments.

However, the mere absence of biased annotations in SSL does not guarantee fairness, necessitating evaluations that extend beyond accuracy. Preliminary research efforts show that SSL techniques can incorporate protected attributes into their representations causing potentially unfair predictions on downstream tasks [29]. For example, Steed and Caliskan [51] have demonstrated that image representations learned with unsupervised pre-training exhibit human-like biases. Yet, while studies in CV and audio have found similarities in intermediate representations between SSL and supervised alternatives, it is crucial to emphasize that most comparisons focus on aspects other than fairness [7, 14]. To date, there is a distinct lack of comprehensive investigations specifically addressing fairness considerations in SSL learned representations.

Fairness evaluations in SSL have been more prevalent in CV [38] and NLP, including recent advances in generative models [45]. For instance, while models fine-tuned on top of pre-trained models can inherit their biases [56], Ranjit et al. [40] have shown that supervised models tend to preserve their pre-training biases regardless of the target dataset, in contrast to SSL methods, where the fine-tuning objective and dataset influence the extent of transferred biases. Discussions on SSL's impact on fairness include considerations of training without prior data curation and the effects of fine-tuning [13, 30, 38]. However, in multimodal, human-centric data, fairness evaluations have seen limited exploration, mainly in a supervised setting. For instance, for in-hospital mortality using the MIMIC dataset [31, 41] or keyword spotting for on-device ML [21]. Hence, such efforts are still in their early stages [61].

Research Gap. This paper aims to address the research gap by assessing SSL approaches in real-world human-centric data, considering both performance and fairness aspects. While there are existing works addressing performance, fairness, or learned representations *individually* in SSL (across different domains), evidence that connects all these three aspects, particularly in human-centric, multimodal, data, such as time-series, is still lacking.

3 METHOD

In this section, we introduce the proposed framework to investigate how design choices in SSL affect outputs and representations.

Notation. Let $X = (x_1, \ldots, x_N) \in \mathbb{R}^{N \times T \times M}$ denote an input sequence with N samples of T sample length and M modalities (e.g., multivariate signals), and $Y = (y_1, \ldots, y_N) \in \mathbb{R}^N$ denote the respective binary output sequence. In the context of SSL, let $f(\cdot)$ denote an encoder that maps input samples X into intermediate embeddings $H = (h_1, \ldots, h_D) \in \mathbb{R}^{N \times D}$ where D is the size of the latent dimension. These embeddings are further trained by the fine-tuning strategy ϕ employed for the downstream task resulting in $H_\phi = (h_{1,\phi}, \ldots, h_{D,\phi}) \in \mathbb{R}^{N \times D}$, where $\Phi = \{\phi_1, \ldots, \phi_L\} \in \mathbb{B}$ controls the training status (trainable or frozen) of each layer l in the base encoder, where L is the total number of layers. For conditioning these representations on protected attributes, we denote P as the set of protected attributes, and V_p as the set of possible values for the protected attribute p. Thus, we denote the conitioned representations as $H_{\phi,p=v} = (h_{1,\phi,p=v}, \ldots, h_{D,\phi,p=v}) \in \mathbb{R}^{N' \times D}$, where N' is the subgroup of samples, where the user has a value v for the protected attribute p.

Put simply, SSL models learn representations from data in an unsupervised manner, potentially avoiding biases present in labeled data. In contrast, supervised models can amplify biases in labels into learned representations. SSL's contrastive learning objective encourages invariant representations, aligning with debiasing goals. The subsequent supervised fine-tuning has a limited capacity to bias precomputed representations compared to training a supervised model from scratch. We hypothesise that the information bottleneck theory [46] in SSL acts as a regularizer, potentially mitigating biased signals during pretraining. In the following, we present the five stages of our framework for facilitating fairness assessments in SSL:

1. Dataset Requirements Definition. In the context of fairness analyses and the scope of this work, it is essential to consider certain requirements during benchmark dataset selection, significantly limiting the available dataset choices as follows:

(1) **Protected attributes:** The dataset should provide at least one protected attribute, such as age, gender, or race;
(2) **Size:** The dataset should contain data from "sufficient"[3] users to allow for statistical comparisons (of fairness and performance metrics) between user segments;
(3) **Modality:** The dataset should contain more than one modality, such as different sensor measurements, hence excluding unimodal benchmarks in vision or language;
(4) **Open Benchmark:** To foster reproducibility and allow for comparisons with the literature, we focus on publicly available benchmarks and pre-processing pipelines.

2. Models. We use a SimCLR [5] variant adapted for time-series data [53]. Our design mirrors SimCLR's components: a) a *stochastic data augmentation* module that transforms a data sample x in two correlated views, denoted $\tilde{x}_i$ and $\tilde{x}_j$ (i.e., positive pair) by employing scaling and signal inversion; b) a *base encoder* $f(\cdot)$ for extracting representation embeddings from augmented data samples. We opt for a 3-layer Convolutional Neural Network (CNN) in line with [53] to obtain $h_i = f(\tilde{x}_i) = \text{ConvNet}(\tilde{x}_i)$, where $h_i \in \mathbb{R}^d$ is the output after the max pooling layer (dependent on the fine-tuning setup described in the next section); c) a *projection head* $g(\cdot)$ for mapping representations to the contrastive loss space. We opt for a 2-layer Multi-layer Perceptron (MLP) to obtain $z_i^{[l]} = g(h_i^{[l]}) = W^{[l]}\sigma({}^{[l-1]}h_i^{[l-1]})$, where σ is a ReLU non-linearity and $l = 3$ for 2 hidden and 1 output layers; d) a contrastive loss function, namely a normalized temperature-scaled cross-entropy loss (NT-Xent) [5,

[3]Considering that SSL requires large datasets for pre-training, we focus on large human-centric datasets with thousands of samples.

49]. We define a similar architecture for the supervised baseline, replacing the contrastive loss with categorical cross-entropy.

3. Fine-tuning Setup. To assess the effect of self-supervision on fairness outcomes and representations, we employ a "gradual unfreezing" strategy (Algorithm 1) [20], balancing the impact of the pre-trained encoder and the downstream labels. We start by freezing all three base encoder layers and fine-tuning only the projection head. Then, starting from the last layer containing the least general knowledge [63], we gradually unfreeze layers one by one (or block by block, achieved via a step of $\lfloor \frac{L}{3} \rfloor$ in the pseudocode) until the encoder layers and projection head are fully trainable (similar to full supervision). We also experiment with different freezing configurations, similar to "surgical fine-tuning" [27], where we tune only one (block of) layer(s), and freeze the remaining, as tuning different blocks of layers performs best for different types of distribution shifts. Figure 1 visualizes the described fine-tuning setup.

Algorithm 1: Gradual Unfreezing

Input: Sequence X and encoder $f(\cdot)$ where the layers' training status is controlled by $\Phi = \{\phi_1, \ldots, \phi_L\}$
Output: Embeddings $H_\phi = (h_{1,\phi}, \ldots, h_{D,\phi})$
for $l \leftarrow L-1$ **to** 0 **do**
  $\Phi[l] \leftarrow 0$; `/* freeze all */`
end
for $l \leftarrow L-1$ **to** 0 **by** $\lfloor \frac{L}{3} \rfloor$ **do**
  $\Phi[l] \leftarrow 1$; `/* unfreeze one-by-one */`
  $H_\phi \leftarrow f(X, \Phi)$; `/* fine-tune trainable */`
end

More formally, we define individual parameters $\Phi = \{\phi_1, \phi_2, \phi_3\}$ to control the training status of each layer in the 3-layer base encoder. The output h_i is then determined by the ConvNet function with parameters Φ, where ϕ_i indicates whether each corresponding layer is frozen (0) or trainable (1). This allows for a flexible downstream task configuration where specific layers can be selectively frozen or trained based on the desired experimental setup, where the representations embeddings H_ϕ are obtained as follows:

$$h_{i,\Phi} = f(\tilde{x}_i, \Phi) = \text{ConvNet}(\tilde{x}_i, \Phi)$$

4. Custom Representation Similarity Function. For assessing the impact of supervision on learned representations, we adopt the linear Centered Kernel Alignment (CKA) method as a similarity index. CKA has proven superior to related methods, such as linear regression, or canonical correlation analysis (CCA), addressing challenges regarding the distributed nature, potential misalignment, and high dimensionality of representations [25]. We propose the conditioning of CKA on protected attributes to identify differences in representation similarity between diverse demographic groups.

More formally, let P represent the set of protected attributes, and V_p represent the set of possible values for the protected attribute p. Given the $H_\phi \in \mathbb{R}^{N \times D}$ activations for the SSL model (i.e., intermediate feature representations) and $J \in \mathbb{R}^{N \times D}$ activations for the supervised model, for the same examples, we can calculate CKA based on a subset of those activations conditioned on the users' protected attributes as $H_{\phi,p=v} = \{h_{i,\Phi} \mid p(\tilde{x}_i) = v,\ p \in \mathcal{P},\ v \in \mathcal{V}_p\}$. Then, the conditioned linear CKA is given by:

$$CKA(H, J, P, V) = \frac{\|H^T_{\phi,p=v} J_{p=v}\|_F^2}{\|H^T_{\phi,p=v} H_{\phi,p=v}\|_F \|J^T_{p=v} J_{p=v}\|_F}$$

5. Domain-specific Evaluation Processes. Enhancing fairness in ML requires a means to quantify biases. Particularly in human-centric settings and high-stakes applications, single evaluation metrics struggle to reflect the success of ML models. As such, monitoring and reporting a multitude of metrics across different protected groups becomes the norm. Fairness trees [43] accompanied by domain expertise can help researchers choose appropriate metrics.

To capture the different fairness perspectives (§2.1), we adopt multiple ratio-based metrics. Specifically, we use the disparate impact (WAE), false omission rate, false discovery rate, false negative rate, and false positive rate (hybrid) ratios. We adopt the above metrics to acknowledge the potential consequences of both false positives and false negatives in the context of healthcare and the implications of prediction disparities among protected attributes (e.g., falsely administered medication or unnecessary financial burden for false positives and life threatening consequences or missed treatment opportunities for false negatives) [3, 9].

Ratio metrics are bounded within the range $[0, +\infty)$, where a value of 1.0 signifies parity across protected attributes. In this work, we define and use a custom fairness meta-metric, the so called parity deviation, as a fairness indicator, which we calculate as follows:

$$Parity\ Deviation_{fairness_metric} = \|1 - fairness_metric\|$$

Here, the term "metric" refers to the specific ratio metrics mentioned earlier. For example, for the disparate impact ratio (DIR) metric:

$$Parity\ Deviation_{DIR} = \left\| 1 - \frac{Pr(Y = 1 \mid V_p = unprivileged)}{Pr(Y = 1 \mid V_p = privileged)} \right\|$$

The ideal deviation lies close to 0.0 (i.e., no deviation from parity), whereas values below 0.2 fall within the acceptable ("fair") range [2]. Instead of focusing on individual fairness metrics that assess fairness for specific groups, a fairness meta-metric combines multiple metrics into a single measure to facilitate comparisons between metrics regardless of output range and interpretation.

To ensure that models are both fair and accurate, we employ the AUC-ROC metric and calculate 95% confidence intervals (CI). Consistent with the benchmark introduction [17], our focus on per-instance accuracy leads to calculating overall performance as the micro-average across all predictions, irrespective of the user.

4 EVALUATION

We follow the protocol below to assess the applicability and generalizability[4] of our framework across datasets.

Overview of Datasets & Tasks. Considering the dataset requirements defined in §3, we exclude certain datasets, such as those typically used for fairness research, due to insufficient size for SSL training (e.g., Adult, COMPAS, German Credit), or widely used SSL benchmark datasets, due to modality mismatch (e.g., CelebA, Equity

[4]The term "generalizability" refers to the broad applicability of our fairness evaluation framework and findings across multiple real-world datasets and tasks, demonstrating the relative fairness improvements of SSL across diverse data modalities involving human subjects–not the generalization abilities of the models themselves.

Table 1: Datasets used in evaluation

Data	# Users	# Samples	Downstream Task	Modalities	Protected Attributes
MIMIC	18.1K	21.1K	Mortality Prediction	Multivariate clinical measurements, e.g, weight, heart rate, blood pressure (M=76)	Age, Race, Gender, Language, Insurance
MESA	1.8K	2.2M	Sleep-Wake Classification	Activity count and white, red, green, blue light measurements (M=5)	Age, Race, Sex
GLOBEM	0.7K	8.1K	Depression Detection	Multivariate behavioral signals, e.g., phone use, sleep, location (M=1390)	Race, Gender, Disability

Evaluation Corpus). We also exclude benchmarks used for human-centric tasks such as human-activity recognition due to small sample size, lack of protected attributes, or both (e.g., PAMAP2, MotionSense, UCI-HAR). We select three multimodal, human-centric datasets (Table 1) spanning the following use cases: in-hospital mortality prediction based on health records and physiological signals, sleep-wake classification based on actigraphy signals, and depression detection based on behavioral signals. Selected datasets contain different levels of representation bias (Appendix A) to evaluate our hypothesis under different scenarios. For simplicity and readability, we refer to the participating individuals as "users".

1. **MIMIC:** the MIMIC-III Clinical Database [22] contains more than 31 million clinical events that correspond to 17 clinical variables (e.g., heart rate, oxygen saturation, temperature). Our task involves prediction of in-hospital mortality from observations recorded within 48 hours of an intensive care unit (ICU) admission–a primary outcome of interest in acute care. Following the benchmark preparation workflow by Harutyunyan et al. [17], we proceed with a total of 18.1K users, forming 21.1K windows, each with 48 timestamps, 76 channels, and no overlap.
2. **MESA:** the Multi-Ethnic Study of Atherosclerosis (MESA) [6], contains polysomnography (PSG) and actigraphy data for 1817 out of the initial 2.2K users in the MESA sleep study, based on the benchmark by Palotti et al. [34]. Our task involves the classification of sleep-wake stages over overnight experiments split into 30-s epochs, forming a total of more than 2.2M windows, each with 101 timestamps, 5 channels, and maximum overlap.
3. **GLOBEM:** the multi-year sensing dataset, GLOBEM [58] contains a rich collection of survey and behavioral data, including location, phone usage, physical activity, and sleep, for 497 unique users monitored over four consecutive years for 3-month periods at a time. Our task involves depression detection (self-reported), given a feature matrix including daily feature vectors for the past four weeks. Following the benchmark preparation workflow by Xu et al. [57], we proceed with a total of more than 8K windows, each with 28 timestamps, and 1390 channels.

Establishing Protected Attributes. Human activities data exhibit variability based on the user's attributes [50]. A starting point for investigating bias is thus to investigate test-time performance for protected attribute groups with different socio-demographic attributes. Figure 7 (Appendix A) shows the (highly imbalanced) distribution of users based on protected attributes. Specifically, the *MIMIC* dataset contains a multitude of protected attributes relevant to the in-hospital mortality task: gender, age, ethnicity, religion, language, and insurance type (a proxy for socioeconomic status). Prior work has revealed disparate treatment in prescribing mechanical ventilation among user groups across ethnicity, gender, and age [31], and voiced general fairness concerns for Black and publicly insured users [41]. Containing fewer protected attributes, the *MESA* dataset includes age, gender, and ethnicity–highly relevant for sleep classification. Specifically, studies have shown that sleep disorders are more prevalent among older adults, and Black populations and vary with gender and obesity status [6]. Finally, *GLOBEM* provides access to gender, race, and disability data upon request. These attributes are highly relevant for depression prediction, as depression rates are higher in women, people with physical disabilities, and untreated racial minority populations [1, 35, 55].

Training Setup and Hyper-parameter Tuning. Following Tang et al. [53]'s recommended architecture for contrastive learning on signals, our model comprises a base encoder featuring three temporal (1D) convolutional layers with kernel sizes of 24, 16, 8, and 32, 64, 96 filters, ReLU activation, a dropout rate of 0.1 (0.4 for GLOBEM), and a concluding global maximum pooling layer. For pre-training, a projection head with three fully-connected layers (256, 128, and 50 units) is utilized, while the fine-tuned evaluation incorporates a classification head with two fully-connected layers (128 and 2 units). Pre-training employs the SGD optimizer with cosine decay of the learning rate over 200 epochs and a batch size of 128. Linear evaluation involves training for 100 epochs with the Adadelta optimizer and a learning rate of 0.03.

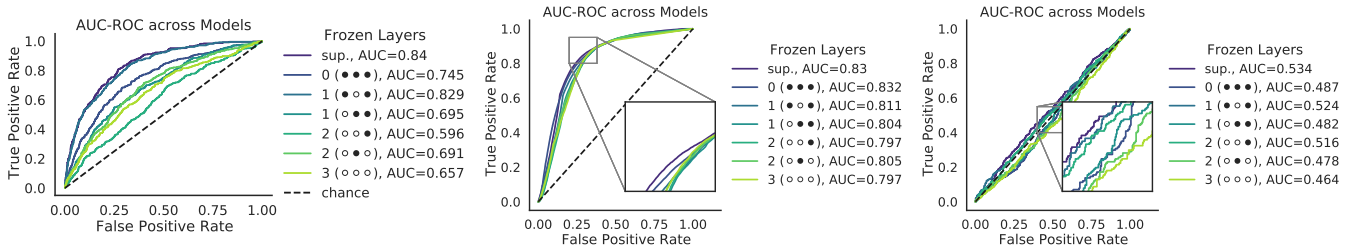
Hyperparameters have been finetuned through grid search across ranges of layer numbers (projection head) [2, 3], batch size [64, 128], epochs with or without early stopping [100, 200], learning rates [0.1, 0.01, 0.03, 0.001] with and without decay [1000, 2000 steps], optimizers [SGD, Adam, Adadelta], and dropout [0.1, 0.3, 0.4, 0.5], and their impact was evaluated on the validation set. A full grid search would yield over 700 models, but we estimate 100 models conservatively due to untested combinations through pruning.

5 RESULTS

To assess the truthfulness of our hypothesis, we explore the impact of fine-tuning in SSL on performance, fairness, and representations.

5.1 Impact of Supervision on Performance

SSL performs on par with supervised alternatives. Figure 2 presents the ROC Curves and the AUC-ROC scores for both supervised and SSL models with various levels of fine-tuning across datasets. We notice that for the all datasets, the fully supervised model performs the best in terms of AUC-ROC with a score of 0.84 (CI 0.82-0.86) for MIMIC, 0.83 (CI 0.83-0.83) for MESA, and 0.534 (CI 0.49-0.57) for GLOBEM. In every case, it is closely followed by the SSL model with a single frozen layer (middle) during fine-tuning, i.e., 1 (• ∘ •) with an AUC-ROC score of 0.829 (CI 0.81-0.85) for MIMIC, 0.811 (CI 0.81-0.81) for MESA, and 0.524 (0.49-0.56) for GLOBEM—a mere 1-2% loss in overall performance. Such results are in line with prior benchmarking efforts in SSL for visual tasks [32] and, closer to our work, human activity recognition [16].



(a) MIMIC **(b) MESA** **(c) GLOBEM**

Figure 2: AUC-ROC curves across datasets and fine-tuning strategies. The supervised models show superior performance, but are closely followed by SSL alternatives, e.g., 1 (● ○ ●). The level of fine-tuning in SSL greatly affects the observed performance.

Models perform inequitably across protected attributes. To condition performance on protected attributes, Table 2 presents AUC-ROC scores per segment for the supervised and the best-performing self-supervised model. We notice that the models do not perform equitably for all segments. Specifically, for MIMIC, there exists a considerable performance gap experienced by black patients, registering a deviation of nearly −8% in AUC-ROC, followed by Medicaid-insured patients with deviations exceeding −5%. Conversely, patients with self-insurance show the best performance with deviations up to +14%, trailed by Hispanics with deviations over +11%. These findings align with previous studies involving supervised models for MIMIC-III mortality prediction. Notably, Medicaid patients consistently receive inferior predictions despite sharing comparable mortality rates with privately patients. Similarly, black patients consistently underperform compared to white patients, even in the presence of lower mortality rates in the dataset. On the other hand, Hispanic patients exhibit elevated performance attributable to their significantly lower mortality rates compared to other demographic groups [41] (for more details on mortality rates see Table 3 in Appendix A). Similarly, for GLOBEM, there also exists a significant performance gap experienced by users identifying with gender identities other than the ones included, registering deviations of almost −30% in AUC-ROC for the supervised model, while users with disabilities register deviations of −20% for the SSL model. Conversely, White users show the best performance for the supervised model with deviations up to +12%. Lastly, we notice smaller performance discrepancies for MESA, where younger study participants (< 65 years old) show slightly superior performance (+3%), while Asian participants show slightly declined AUC-ROC scores (−2%).

SSL is "fairer" for smaller segments. Overall, performance discrepancies are similar between the two models if we focus on the $\Delta_{Supervised-SSL}$ on Table 2, with a slight fairness benefit for the SSL model (indicated by green color). This is more prevalent on the GLOBEM dataset, with an exception of users with disability. However, we should consider the impact of sample size on performance. Figure 3 shows the correlation between segment size and performance gap across datasets and protected attributes. The closer the points to the "fair" (dashed) line, the smaller the performance gap for this segment with the general population. For segments > 35% of the population, points are closer to 0.0, whereas smaller segments have much wider gaps. Note that in both cases (i.e., supervised and

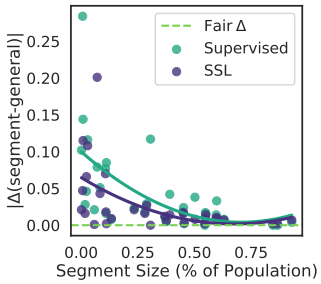


Figure 3: The relationship between segment size and performance (AUC-ROC) across datasets. The smaller the segment the larger the performance discrepancies. Fitted lowess curves show that SSL lies closer to the "fair" (dashed) line.

SSL model), there is a strong negative correlation between segment size and performance gap, namely, the smaller the size, the larger the performance gap. Nevertheless, the lowess curve for the SSL model lies closer to the "fair" line, indicating smaller discrepancies between groups.

Nevertheless, performance metrics are not always the best indicator of fairness. Even if a model performs well on average, it might exhibit significant differences in error rates across different groups. For instance, false positives or false negatives may disproportionately affect certain demographic groups, leading to unfair outcomes–a prospect we explore in the following section.

5.2 Impact of Supervision on Fairness

SSL decreases deviation from fairness parity. Figure 4 shows the deviation of each model's ratio metrics from parity. Deviations greater than 0.2 (dashed line) indicate bias towards a protected attribute, irrespective of privilege. Despite the supervised model having slightly superior performance, it has significantly greater deviation from parity compared to the best-performing SSL model (i.e., 1 ● ○ ●) for the MIMIC and GLOBEM datasets. Specifically, MIMIC's supervised model has on average a 0.24 deviation from parity, while the SSL a 0.21–a 13% decrease, while GLOBEM's supervised model has a 0.23 deviation from parity, while the SSL a 0.17–a 30% decrease. More importantly, SSL models with a balanced level of unfreezing lie mostly within the acceptable "fairness" limits, opposite to the supervised alternative. Note that for the MESA dataset we did not identify any significant differences in parity deviation.

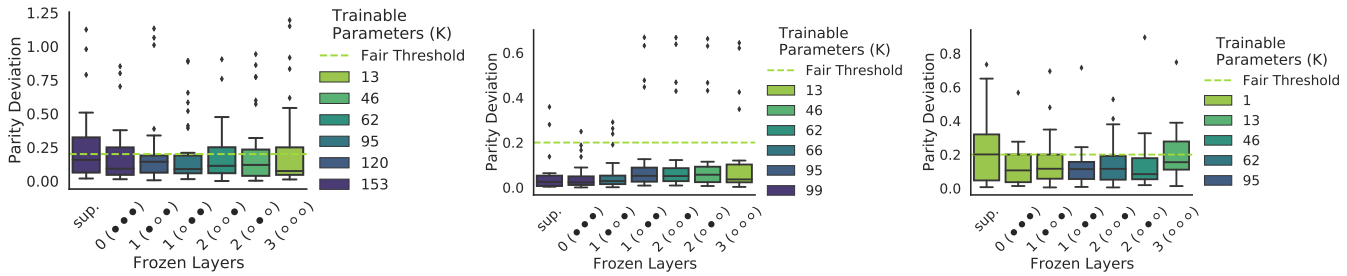
Table 2: Comparison of AUC-ROC between the fully-supervised and the best-performing SSL model, conditioned on protected attributes. Numbers in parentheses indicate 95% Confidence Intervals. The Δ columns show differences between a segment and the general population, where yellow indicates disadvantaged and green advantaged segments. The most disadvantaged segment is underlined, and the most advantaged is in bold.

			Models						
Datasets	**Protected Attribute**	**Segments**	*Supervised*		$\Delta_{segment-general}$	*SSL (1 • ○ •)*		$\Delta_{segment-general}$	$\Delta_{Supervised-SSL}$
MIMIC	General Population		*0.839*	(0.82-0.86)		0.829	(0.81-0.85)		
	Age	< 65	0.863	(0.83-0.89)	0.024	0.845	(0.8-0.88)	0.016	0.008
		≥ 65	0.822	(0.8-0.85)	-0.017	0.82	(0.79-0.85)	-0.009	0.008
	Ethnicity/Race	White	0.839	(0.82-0.86)	0	0.831	(0.81-0.86)	0.002	-0.002
		Black	0.762	(0.65-0.85)	-0.077	0.759	(0.63-0.85)	-0.07	0.007
		Asian	0.811	(0.68-0.92)	-0.028	0.813	(0.63-0.94)	-0.016	0.012
		Hispanic	0.955	(0.9-0.99)	0.116	0.937	(0.86-0.98)	0.108	0.008
	Gender	Male	0.855	(0.83-0.88)	0.016	0.843	(0.81-0.87)	0.014	0.002
		Female	0.821	(0.79-0.85)	-0.018	0.812	(0.78-0.84)	-0.017	0.001
	Insurance	Medicare	0.825	(0.8-0.85)	-0.014	0.819	(0.79-0.84)	-0.01	0.004
		Private	0.868	(0.83-0.9)	0.029	0.856	(0.81-0.9)	0.027	0.002
		Medicaid	0.788	(0.67-0.88)	-0.051	0.786	(0.68-0.87)	-0.043	0.008
		Government	0.885	(0.77-0.99)	0.046	0.895	(0.8-0.98)	0.066	-0.020
		Self Pay	**0.983**	**(0.93-1.0)**	**0.144**	**0.944**	**(0.84-1.0)**	**0.115**	0.029
	Language	English	0.839	(0.81-0.87)	0	0.831	(0.79-0.86)	0.002	-0.002
		Other	0.831	(0.8-0.86)	-0.008	0.82	(0.79-0.84)	-0.009	-0.001
MESA	General Population		0.83	(0.83-0.83)		0.811	(0.81-0.81)		
	Age	**< 65**	**0.856**	**(0.85-0.86)**	**0.026**	**0.838**	**(0.83-0.84)**	**0.027**	-0.001
		≥ 65	0.813	(0.81-0.81)	-0.017	0.794	(0.79-0.80)	-0.017	0.000
	Ethnicity/Race	White	0.838	(0.83-0.84)	0.008	0.82	(0.82-0.82)	0.009	-0.001
		Black	0.833	(0.83-0.84)	0.003	0.808	(0.80-0.81)	-0.003	0.000
		Asian	0.81	(0.81-0.81)	-0.020	0.8	(0.8-0.8)	-0.011	0.009
		Hispanic	0.819	(0.81-0.82)	-0.011	0.801	(0.8-0.8)	-0.01	0.001
	Gender	Male	0.831	(0.83-0.83)	0.001	0.813	(0.81-0.82)	0.002	-0.001
		Female	0.829	(0.82-0.83)	-0.001	0.81	(0.81-0.81)	-0.001	0.000
GLOBEM	General Population		0.534	(0.49-0.57)		0.524	(0.49-0.56)		
	Disability	No	0.530	(0.5-0.57)	-0.004	0.531	(0.49-0.57)	0.007	-0.003
		Yes	0.613	(0.41-0.8)	.079	0.323	(0.17-0.51)	-0.201	-0.122
	Ethnicity/Race	**White**	**0.651**	**(0.57-0.72)**	**0.117**	0.524	(0.44-0.6)	0	0.117
		Black	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		Asian	0.496	(0.45-0.54)	-0.038	0.516	(0.47-0.56)	-0.008	0.030
		Hispanic/Latinx	0.555	(0.41-0.7)	0.021	0.525	(0.39-0.67)	0.001	0.020
		Biracial	0.449	(0.33-0.56)	-0.085	0.526	(0.43-0.63)	0.002	0.083
	Gender	Male	0.576	(0.52-0.63)	0.042	0.538	(0.48-0.6)	0.014	0.028
		Female	0.501	(0.45-0.55)	-0.033	0.514	(0.46-0.56)	-0.01	0.023
		Transgender	0.636	(0.33-0.9)	0.102	0.545	(0.2-0.82)	0.021	0.081
		Other	0.25	(0.0-0.6)	-0.284	**0.571**	**(0.1-1.0)**	**0.047**	0.237

This lack of discernible differences could be attributed to the inherent simplicity of the task at hand, i.e., sleep-wake classification, or the more balanced distribution of subjects. Detailed experimental results on individual fairness metrics comparisons between the SSL model, its linear probing alternative, and the supervised model can be found in Table 5 on Appendix C.

Middle unfreezing balances performance and fairness. Additionally, prior work in other domains supports that fine-tuning has an important impact on fairness [38, 39]. Indeed, our findings illustrate this point for human-centric, multimodal data, too, with statistically significant differences in fairness ratios between SSL models with different levels of fine-tuning (e.g., 1 • ○ • and 3 ○ ○ ○). This is better illustrated by the observed"U-shape" patterns in the MIMIC and GLOBEM datasets, suggesting an optimal level of supervision—a sweet spot at middle unfreezing that balances trainable parameters and frozen layers in SSL (Figure 4).

SSL's fairness gain is more data-efficient. Following prior work supporting that SSL model can achieve high performance with significantly less training data [52], we also assess algorithmic bias (expressed via parity deviation) as a function of limited training data. This evaluation is designed to simulate scenarios where the resources for collecting labeled data are very limited, which might arise in small-scale or academic data collection studies, resulting in limited samples per protected attribute. In this evaluation protocol,



(a) MIMIC (b) MESA (c) GLOBEM

Figure 4: Relationship between fairness (deviation from parity) and fine-tuning strategies, as a function of model size. The supervised model has a greater deviation from parity, i.e., increased bias, (dashed line) compared to the best-performing SSL model (i.e., 1 • ○ •). The observed "U-shape" patterns in MIMIC and GLOBEM datasets suggest an optimal level of fine-tuning.

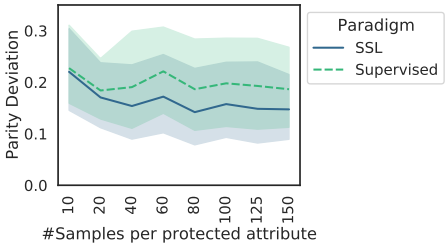
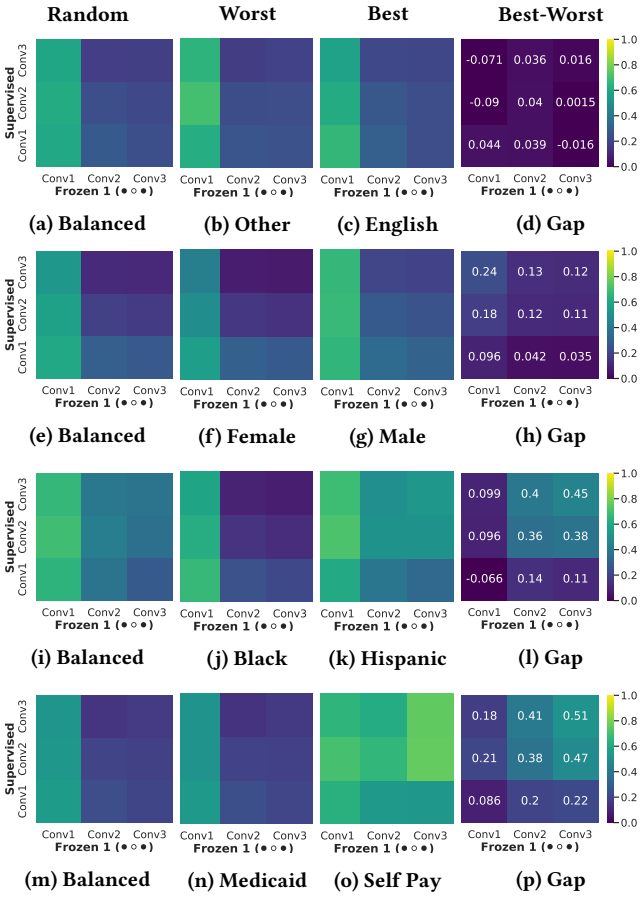


Figure 5: Assessing fairness as a function of fine-tuning labelled data in the MIMIC dataset. The SSL model achieves increased fairness with less training data ($\geq$ 40 samples per attribute). The shaded error bands represent the range of parity deviation values across fairness metrics and attributes.

a fixed number of labeled samples per ethnicity (i.e., the protected attribute) are extracted from the labeled datasets, and they are the only labeled training data that the models are trained or fine-tuned on. We extract 10-150 samples per ethnicity segment to simulate the different degrees of availability. Figure 5 illustrates the parity deviation of models trained on an increasing number of labeled data per protected attribute for the MIMIC dataset, as a case in point. We notice that while the deviation is similar for very limited data ($\leq$ 20), the SSL model shows a quicker fairness gain than the supervised alternative (sample size $\geq$ 40 per attribute).

5.3 Interplay of Representations and Fairness

The larger the performance gap between protected attributes, the greater the fairness deviation. We compare representation similarity between the supervised and the best-performing SSL model across protected attributes (language, gender, ethnicity, insurance) through CKA. Our findings regarding the impact of supervision on representation learning for time-series data align with prior work on CV. Specifically, Grigg et al. [14] illustrate how self-supervised and supervised methods learn similar visual representations through dissimilar means and that the learned representations diverge rapidly in the final few layers. Indeed, as illustrated in Figure 6, the initial layer representations are similar, indicating a shared set of primitives. However, we notice discrepancies at the level of similarity for certain protected attributes. Taking ethnicity as a case



(a) Balanced (b) Other (c) English (d) Gap
(e) Balanced (f) Female (g) Male (h) Gap
(i) Balanced (j) Black (k) Hispanic (l) Gap
(m) Balanced (n) Medicaid (o) Self Pay (p) Gap

Figure 6: Conditioned representation similarity between the supervised and the SSL model through CKA (MIMIC). Rows correspond to language, gender, race, and insurance. The random subset (first column) is balanced per segment. The similarity is lower for the worst- (second) than the best-performing (third) segment. The higher the performance gap between segments the larger the representation gap (fourth).

in point, we observe a greater similarity in learned representations for Hispanic patients compared to Black patients. This dissimilarity could contribute to the performance gap between these two demographic groups. Notably, we find a correlation between the magnitude of the performance gap across patient segments (Table 4) and the representation similarity gap (Figure 6). For instance, the performance gap is minimal for language (∼0%), slightly larger for gender (∼3%), and more pronounced for ethnicity and insurance (∼19%), mirroring the same trend in the similarity gap. For instance, the representation similarity for the best-performing segment is up to ×13 greater for the insurance attribute compared to language (median CKA 0.22 to 0.016, respectively).

Interestingly, for the best-performing groups, such as Hispanic or self pay patients, both the SSL and supervised models not only excel in performance but also exhibit strikingly similar representations. This suggests a shared capability in capturing and encoding the underlying data patterns, leading to coherent model outputs. Conversely, when confronted with the worst-performing groups, like Black or Medicaid patients, both models struggle in terms of performance, yet their learned representations diverge notably. Such dissimilarity in their representations implies a focus on different data aspects. The SSL model may be capturing patterns or features not effectively recognized by the supervised model, possibly due to the former's limited reliance on labels. Conversely, the supervised model may emphasize features aligned with labeled data but could face challenges in generalizing due to the inherent complexity of the worst-performing groups (e.g., limited data, class imbalance). A preliminary exploration of the correlation between features learned and protected attributes is given in Appendix B. Yet, an in-depth understanding of why the models differ in learned representations for the worst-performing groups is crucial to figuring out the challenges of each learning paradigm in terms of bias.

6 DISCUSSION & CONCLUSIONS

In conclusion, our investigation into the application of self-supervision in human-centric, multimodal data revealed that SSL models, particularly when fine-tuned with middle unfreezing, can achieve improved fairness compared to supervised models while preserving performance. Our intuition is that this fine-tuning strategy strikes a balance between retaining knowledge from raw data representations and leveraging information from labeled data. Interestingly, the SSL's learned representations showcased both similarities and differences with their supervised counterparts, indicating nuanced patterns in capturing and encoding information.

Broadly, the SSL models exhibited smaller deviations from parity across protected attributes, indicating potential effectiveness in mitigating biases associated with downstream labels. Yet, the focus of this work is on evaluating how design choices in SSL impact fairness, rather than proposing new fairness mitigation algorithms. However, our SSL framework parallels implicit fairness mitigation methods. For instance, the pre-training phase acts akin to pre-processing, removing discriminatory signals by learning from unlabeled data. The subsequent fine-tuning phase operates like an in-processing method, controlling the regularization effect on the model's accuracy. However, it is essential to acknowledge that SSL alone may not eliminate all disparities, especially when trained on poor-quality or biased data, as seen in cases from other domains [33]. Future research should explore additional strategies for bias mitigation, and comparative studies with supervised models designed explicitly for bias reduction [28] are warranted.

Overall, while SSL presents a positive step towards fairness in real-world, human-centric tasks, it should be considered as part of a broader strategy for addressing bias in ML models, taking into account task-specific nuances and the quality of training data. The assessment of prediction fairness should consider the data context, and any unfairness arising from insufficient sample sizes or unmeasured predictive variables should be rectified through additional data collection rather than restricting the model.

ACKNOWLEDGMENTS

The authors affiliated with the Aristotle University of Thessaloniki acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162 and the Hellenic Artificial Intelligence Society. The content of this paper reflects only the authors' view and the Agency and the Commission are not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Rahn Kennedy Bailey, Josephine Mokonogho, and Alok Kumar. 2019. Racial and ethnic differences in depression: current perspectives. *Neuropsychiatric disease and treatment* (2019), 603–609.

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

[3] Tal Burt, KS Button, HHZ Thom, RJ Noveck, and Marcus R Munafò. 2017. The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures. *Clinical and translational science* 10, 6 (2017), 470–479.

[4] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems* 31 (2018).

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. 2015. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* 38, 6 (2015), 877–888.

[7] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240* (2019).

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

[9] TR Dresselhaus, J Luck, and JW Peabody. 2002. The ethical problem of false positives: a prospective evaluation of physician reporting in the medical record. *Journal of medical ethics* 28, 5 (2002), 291–294.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[11] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[12] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 3662–3666.

[13] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360* (2022).

[14] Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. 2021. Do Self-Supervised and Supervised Methods Learn Similar Visual Representations? *arXiv preprint arXiv:2110.00528* (2021).

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[16] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2022. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–47.

[17] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6, 1 (2019), 96.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems* 32 (2019).

[20] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 328–339. https://doi.org/10.18653/v1/P18-1031

[21] Wiebke (Toussaint) Hutiri, Aaron Yi Ding, Fahim Kawsar, and Akhil Mathur. 2023. Tiny, Always-on, and Fragile: Bias Propagation through Design Choices in On-Device Machine Learning Workflows. *ACM Trans. Softw. Eng. Methodol.* 32, 6, Article 155 (sep 2023), 37 pages. https://doi.org/10.1145/3591867

[22] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[23] Louis Henry Kamulegeya, Mark Okello, John Mark Bwanika, Davis Musinguzi, William Lubega, Davis Rusoke, Faith Nassiwa, and Alexander Börve. 2019. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *BioRxiv* (2019), 826057.

[24] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1920–1929.

[25] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*. PMLR, 3519–3529.

[26] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[27] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2022. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466* (2022).

[28] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.

[29] Martin Q Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Conditional Contrastive Learning for Improving Fairness in Self-Supervised Learning. *arXiv preprint arXiv:2106.02866* (2021).

[30] Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. 2023. Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks. *arXiv preprint arXiv:2304.03935* (2023).

[31] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports* 12, 1 (2022), 7166.

[32] Alejandro Newell and Jia Deng. 2020. How useful is self-supervised pretraining for visual tasks?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7345–7354.

[33] Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine* 6, 1 (2023), 195.

[34] Joao Palotti, Raghvendra Mall, Michael Aupetit, Michael Rueschman, Meghna Singh, Aarti Sathyanarayana, Shahrad Taheri, and Luis Fernandez-Luque. 2019. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ digital medicine* 2, 1 (2019), 50.

[35] Gordon Parker and Heather Brotchie. 2010. Gender differences in depression. *International review of psychiatry* 22, 5 (2010), 429–436.

[36] Ignacio Perez-Pozuelo, Dimitris Spathis, Emma AD Clifton, and Cecilia Mascolo. 2021. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. In *Digital Health*. Elsevier, 33–54.

[37] TS Pias, Y Su, X Tang, H Wang, D Yao, et al. 2023. Undersampling for Fairness: Achieving More Equitable Predictions in Diabetes and Prediabetes. (2023).

[38] Jason Ramapuram, Dan Busbridge, and Russ Webb. 2021. Evaluating the fairness of fine-tuning strategies in self-supervised learning. *arXiv preprint arXiv:2110.00538* (2021).

[39] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. 2023. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering* 30, 4 (2023), 2761–2775.

[40] Jaspreet Ranjit, Tianlu Wang, Baishakhi Ray, and Vicente Ordonez. 2023. Variation of Gender Biases in Visual Recognition Models Before and After Finetuning. *arXiv preprint arXiv:2303.07615* (2023).

[41] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. 2022. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Scientific Data* 9, 1 (2022), 24.

[42] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.

[43] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[44] Pritam Sarkar, Silvia Lobmaier, Bibiana Fabre, Diego González, Alexander Mueller, Martin G Frasch, Marta C Antonelli, and Ali Etemad. 2021. Detection of maternal and fetal stress from the electrocardiogram with self-supervised representation learning. *Scientific reports* 11, 1 (2021), 24146.

[45] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).

[46] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* (2017).

[47] Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. 2020. Racial bias in pulse oximetry measurement. *New England Journal of Medicine* 383, 25 (2020), 2477–2478.

[48] Gizem Sogancioglu and Heysem Kaya. 2022. The effects of gender bias in word embeddings on depression prediction. *arXiv preprint arXiv:2212.07852* (2022).

[49] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems* 29 (2016).

[50] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. 2021. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*. 69–78.

[51] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 701–713.

[52] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 1 (2021), 1–30.

[53] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. 2020. Exploring Contrastive Learning in Human Activity Recognition for Healthcare. *arXiv preprint arXiv:2011.11542* (2020).

[54] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. 2021. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750* (2021).

[55] R Jay Turner and Samuel Noh. 1988. Physical disability and depression: A longitudinal analysis. *Journal of health and social behavior* (1988), 23–37.

[56] Angelina Wang and Olga Russakovsky. 2023. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3957–3968.

[57] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subgiya Nepal, Kevin S Kuehn, Jeremy Huckins, Margaret E Morris, Paula S Nurius, Eve A Riskin, Shwetak Patel, Tim Althoff, Andrew Campell, Anind K Dey, and Jennifer Mankoff. 2022. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2022).

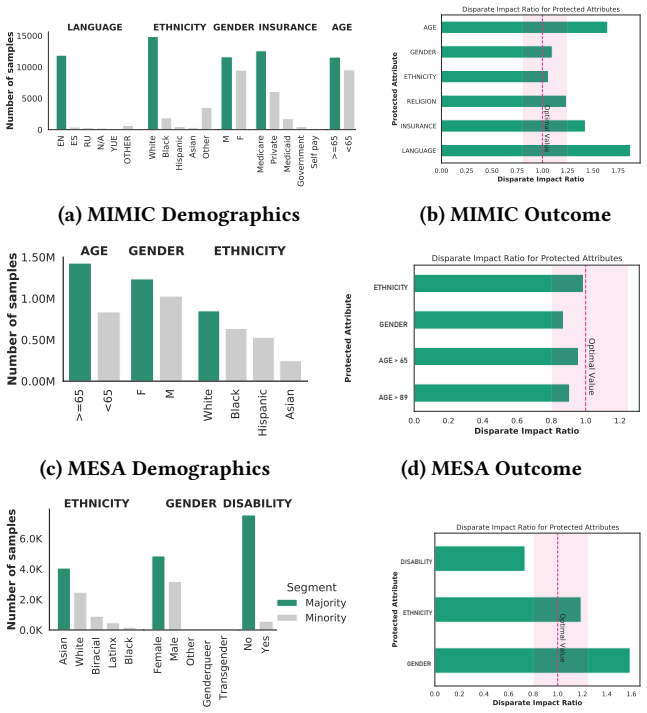
[58] Xuhai Xu, Han Zhang, Yasaman S Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Scott Kuehn, Mike A Merrill, Paula S Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve A. Riskin, Jennifer Mankoff, and Anind Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://arxiv.org/abs/2211.02733

[59] Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. 2021. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*. PMLR, 11964–11974.

[60] Samuel Yeom and Michael Carl Tschantz. 2021. Avoiding Disparity Amplification under Different Worldviews. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 273–283. https://doi.org/10.1145/3442188.3445892

[61] Sofia Yfantidou, Marios Constantinides, Dimitris Spathis, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2023. Beyond Accuracy: A Critical Review of Fairness in Machine Learning for Mobile and Wearable Computing. *arXiv preprint arXiv:2303.15585* (2023).
[62] Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2024. Evaluating Fairness in Self-supervised and Supervised Models for Sequential Data. In *A collection of the accepted papers for the Human-Centric Representation Learning workshop at AAAI 2024*.
[63] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems* 27 (2014).

A REPRESENTATION DIFFERENCES ACROSS DATASETS



(a) MIMIC Demographics (b) MIMIC Outcome

(c) MESA Demographics (d) MESA Outcome

(e) GLOBEM Demographics (f) GLOBEM Outcome

Figure 7: Distribution of subjects (left) and outcomes (right) based on protected attributes. MIMIC and GLOBEM datasets are highly imbalanced in terms of demographics and outcomes -captured via the Disparate Impact Ratio (DIR) metric- opposite to MESA which shows smaller discrepancies.

Figure 7 shows the distribution of demographic groups and outcomes per dataset. GLOBEM and MIMIC exhibit highly imbalanced distributions, whereas MESA presents a more balanced picture. Notably, in the MIMIC dataset, the majority group constitutes 86.6% of data points for the language attribute (English speakers), 70.3% for ethnicity/race (White), 55% for gender (male), 59.5% for insurance (Medicare), and 54.8% for age (≥ 65). Simultaneously, demographic groups exhibit distinct mortality rates, as illustrated in Table 3. In the GLOBEM dataset, the majority group constitutes 50.2% for the ethnicity/race attribute (Asian), 59.8% for gender (Female), and 92.9% for disability (no disability). Finally, in the MESA dataset, the majority group constitutes 63% for the age attribute (≥ 65), 54.5% for gender (Female), and 37.5% for race/ethnicity (White). Regarding outcomes, DIR values outside the shaded region indicate uneven label sampling, which is the case for several attributes in MIMIC and GLOBEM; less so for MESA. Such representation differences help put our findings into context, as prior work supports that the fairness of predictions should be evaluated in context of the data, and that unfairness can be induced by inadequate samples sizes [4].

Table 3: Mortality rates for the MIMIC dataset.

Dataset	Protected Attribute	Segment	Mortality Rate
MIMIC	Age	< 65	9.8%
		≥ 65	16.0%
	Ethnicity/Race	White	12.9%
		Black	9.2%
		Asian	13.8%
		Hispanic	8.1%
	Gender	Female	13.5%
		Male	13.0%
	Insurance	Medicare	14.9%
		Private	10.7%
		Medicaid	10.5%
		Government	9.9%
		Self Pay	17.0%
	Language	English	9.9%
		Other	17.5%

B INTRA- AND INTER-GROUP DISTANCES IN INTERMEDIATE REPRESENTATIONS

To investigate the correlation between features learned by the SSL model and protected attributes, we first determine the medoids, representing the most representative patients, for each demographic segment, and then, we compute the average distances between these medoids. Within the SSL segments, we observe a significant increase in separability, with distances being 70% larger on average ($L1\text{-}norm_{sup} = 4.32$, $L1\text{-}norm_{ssl} = 7.34$). This implies that SSL's decision-making process is, in part, influenced by representations specific to protected attributes. This tendency is further illustrated in Figure 8, using the insurance attribute as a case in point. Specifically, in the SSL model, intra-distances within the worst-performing segment (Medicaid patients) are smaller than inter-distances between the worst-performing and the best-performing (self pay patients) segments. Such a distinctive pattern is notably absent in the supervised model, emphasizing the potential role of protected-attribute-specific representations in SSL's learning process. For comparison, Table 4 illustrates the performance discrepancy between the worst- and best-performing group (in AUC-ROC), e.g., Medicaid *vs.* self-pay for the insurance attribute.

C FAIRNESS METRICS

Apart from the AUC-ROC performance metric, we utilize six popular fairness metrics for our evaluation:

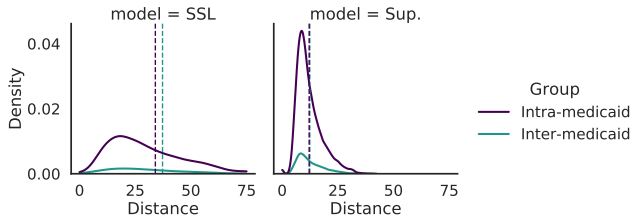


Figure 8: Distribution of intra-group and inter-group distances in intermediate representations between the best-performing and the worst-performing segment. Dashed lines represent the group mean. In the SSL model, intra-distances within the worst-performing segment are smaller compared to inter-distances with the best-performing segments.

Table 4: The disparity between the best and worst-performing groups for the MIMIC dataset is smaller for the SSL model.

Dataset	Protected Attribute	Model	
		Sup	*SSL*
MIMIC	Age	0.04	**0.03**
	Race	0.19	**0.18**
	Gender	0.03	**0.03**
	Insurance	0.20	**0.16**
	Language	0.01	0.01

- Disparate Impact Ratio (DIR): Ratio of selection rates.

$$\frac{Pr(\hat{Y}=1|D=\text{unprivileged})}{Pr(\hat{Y}=1|D=\text{privileged})}$$

- False Discovery Rate Ratio (FDR): Ratio of the proportion of false positives (incorrectly predicted positive cases) to the number of total positive results between different demographic groups, providing a measure to evaluate disparities in model errors across those groups.

$$\frac{FDR_{D=\text{unprivileged}}}{FDR_{D=\text{privileged}}} \text{ where } FDR = FP/(TP+FP)$$

- False Negative Rate Ratio (FNR): Ratio of the proportion of actual positive cases incorrectly predicted as negative between different demographic groups, serving as a measure to assess disparities in model performance across those groups.

$$\frac{FNR_{D=\text{unprivileged}}}{FNR_{D=\text{privileged}}} \text{ where } FNR = FN/P$$

- False Omission Rate Ratio (FOR): Ratio of the proportion of false negatives to the number of total negative results between different demographic groups, offering a metric to assess disparities in model omissions across those groups.

$$\frac{FOR_{D=\text{unprivileged}}}{FOR_{D=\text{privileged}}} \text{ where } FOR = FN/(TN+FN)$$

- False Positive Rate Ratio (FPR): Ratio of the proportion of false positives (incorrectly predicted positive cases) between different demographic groups, serving as a metric to evaluate disparities in model errors across those groups.

$$\frac{FPR_{D=\text{unprivileged}}}{FPR_{D=\text{privileged}}} \text{ where } FPR = FP/N$$

Table 5: Fairness metrics by dataset and protected attribute. Values outside the accepted range (>=0.2 parity deviation) are colored in purple. For those cases, the SSL model performs better or same to the supervised (Sup.) or linear probing (LP) model for all protected attributes in MIMIC and GLOBEM.

Dataset	Protected Attribute	Model	DIR	FDR	FNR	FOR	FPR
MIMIC	Age	SSL	1.19391	0.897846	1.187938	2.132436	1.142868
		Sup.	1.34903	0.917276	1.093168	1.979175	1.319301
		LP	1.047373	0.918621	1.083654	1.832518	1.025794
	Ethnicity	SSL	1.078457	0.977301	0.847877	0.93395	1.063312
		Sup.	1.029786	0.940589	0.877699	0.945341	0.977184
		LP	0.984655	0.964173	0.982594	1.047049	0.957786
	Gender	SSL	1.156379	1.005544	0.959478	1.140239	1.180523
		Sup.	1.26365	1.04365	0.934915	1.116688	1.338922
		LP	1.916451	1.127964	0.938095	1.19075	2.194655
	Insurance	SSL	1.151525	0.949581	1.313889	2.009774	1.143114
		Sup.	1.221075	0.956821	1.173115	1.788963	1.221397
		LP	1.055201	0.944077	1.052997	1.542647	1.041421
	Language	SSL	1.387281	0.873203	0.952534	2.064516	1.318247
		Sup.	1.410703	0.85925	1.018682	2.124872	1.319083
		LP	0.646858	0.807558	1.222419	2.149763	0.568459
MESA	Age	SSL	0.998046	0.999164	0.992309	0.948194	0.983189
		Sup.	0.993162	1.000327	0.995877	0.944088	0.979517
		LP	1.009543	0.997778	0.978821	0.954434	0.993137
	Ethnicity	SSL	0.990607	0.997288	0.992428	0.965692	0.98379
		Sup.	0.992224	0.998345	0.99491	0.970256	0.98644
		LP	1.017328	0.990076	0.962152	0.98008	1.00302
	Gender	SSL	1.081529	1.014118	1.007385	0.991359	1.047817
		Sup.	1.071921	1.015271	1.011415	0.984422	1.039689
		LP	1.033791	1.021893	1.017314	0.941302	1.009246
GLOBEM	Disability	SSL	0.935372	1.348379	1.479401	1.015322	1.084999
		Sup.	0.767811	1.007267	0.95688	0.55	0.665323
		LP	0.635082	1.27369	1.388889	0.793521	0.695868
	Ethnicity	SSL	1.177066	0.911646	0.856541	1.200049	1.183419
		Sup.	1.14846	1.057243	1.188105	1.650568	1.339067
		LP	0.877829	0.892041	1.042793	1.118649	0.863588
	Gender	SSL	0.993029	0.780661	1.07721	1.694643	1.005157
		Sup.	1.006579	0.786325	1.085652	1.734266	1.02626
		LP	1.159948	0.850333	0.986795	1.747962	1.278902

We utilize the parity deviation meta-metric (Section 3), as a means to facilitate the comparison between multiple ratio-based fairness metrics. The usage of multiple fairness metrics, (e.g., false positive rate ratio, false negative rate ratio, etc.) in the healthcare and well-being setting is not uncommon [37, 48]. On the contrary, the usage of multiple metrics is recommended, to capture diverse fairness perspectives in human-centric applications [61]. Beyond the provided meta-metric, Table 5 presents the values of individual fairness metrics, where we compare the best-performing SSL model, with the linear probing model (i.e., the model where we freeze all layers of the pre-trained SSL model and only add a small classification head on top to predict the target labels) as an SSL baseline, and the supervised model (Sup). Partially freezing some layers while fine-tuning others allows for preserving the debiased pre-trained representations to an extent, while allowing specialization of some layers to the target distribution to maintain competitive accuracy. We see that the SSL model (with some level of judicious fine-tuning) shows superior performance for the MIMIC and GLOBEM datasets, having the maximum or equal number of within-range values for the studied fairness metrics for all protected attributes.