Evaluating Fairness in Self-supervised and Supervised Models for Sequential Data

Sofia Yfantidou [2*†] Dimitris Spathis [1] Marios Constantinides [1] Athena Vakali [2] Daniele Quercia [1] Fahim Kawsar [1]

[1]Nokia Bell Labs Cambridge, United Kingdom
[2]Aristotle University of Thessaloniki, Greece

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. 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 (i.e., performing equally on different demographic breakdowns). Motivated by human-centric applications on real-world timeseries data, we interpret inductive biases on the model, layer, and metric levels by systematically comparing SSL models to their supervised counterparts. Our findings demonstrate that SSL has the capacity to achieve performance on par with supervised methods while significantly enhancing fairness—exhibiting up to a 27% increase in fairness with a mere 1% loss in performance through self-supervision. Ultimately, this work underscores SSL's potential in human-centric computing, particularly high-stakes, data-scarce application domains like healthcare.

Introduction

The availability of extensive unlabeled, real-world timeseries data captured through various multimodal sensors presents significant opportunities. As a case in point, in the healthcare domain, leveraging this wealth of information can uncover intricate physiological and behavioral patterns at an unprecedented scale, offering novel insights into personalized and proactive healthcare (**?**). However, the absence of ground-truth labels poses a significant obstacle to progress in supervised machine learning for timeseries tasks (**?**). Self-supervision has proven its performance robustness and beyond state-of-the-art capabilities mainly in the areas of computer vision (CV) (**?**) and natural language processing (NLP) (**?**) research. Inspired by such research, which leverages massive amounts of unlabeled data, many research communities that deal with timeseries data have swiftly recognized the potential of self-supervised learning (SSL). SSL exploitation is promising to explore extensive unlabeled data, effectively complementing small, labeled domain datasets (**?**).

Due to the recency of SSL adoption, in the cases of timeseries, performance metrics, such as accuracy and AUC-ROC score, are typically used as the prime evaluation criteria. Yet, a performance-centric evaluation approach can result in discriminatory or unjust impacts when comparing across different demographics. For instance, **?** 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 (**?**).

While SSL models may avoid such pitfalls due to their pre-training without (potentially) biased human annotations (**?**), comprehensive efforts to compare the fairness of supervised and SSL models are lacking. Note that fairness assessments typically assess accuracy disparities among diverse protected attributes, namely sensitive personal characteristics, such as race or gender, that are legally safeguarded from discrimination. In this study, we aim to address this gap by investigating how fine-tuning SSL models affects fairness in terms of outcomes and latent representations when compared to supervised alternatives. In detail, we make the following contributions:

- For the first time, we look into fine-grained differences in the layer, model, and metric level between supervised and SSL models. Moving away from conventional performance-centric assessments, our approach adopts a human-centric perspective, integrating fairness metrics into our methodology to evaluate how these differences affect model outcomes.
- We train and conduct a systematic comparison of a large number ($>$30) models with various levels of supervision and fine-tuning on multi-year health records timeseries data. We show that SSL performs on par with supervised models, but more notably, we observe up to a 27% increase in fairness, accompanied by only a 1% loss in performance for certain fine-tuning strategies.
- We compare the learned representations using the latent similarity between supervised and SSL models, which reveals discrepancies in latent representations across different demographic groups.

Related Work

While SSL methods such as SimCLR (**?**), BYOL (**?**), and CPC (**?**), have seen widespread use in CV (**?**), NLP (**?**), and audio (**?**), their application in unlabeled, real-world timeseries data has also been successful (**??**).

Existing works have extensively benchmarked SSL algorithms across domains, primarily focusing on performance

[*]Work done at Nokia Bell Labs

[†]Corresponding author: syfantid@csd.auth.gr

metrics. However, limited attention has been given to evaluating fairness in self-supervised methods, particularly for multimodal, timeseries data. For example, in the healthcare setting, representation learning has been applied to patient monitoring (?), Atrial Fibrillation detection (?), mortality or decompensation prediction (?), maternal and fetal stress detection (?), and human-activity recognition (?), among others. However, the above works focus on performance-centric assessments. Yet, the absence of biased annotations in SSL does not ensure fairness, necessitating evaluations beyond accuracy. As a matter of fact, studies in CV and NLP have demonstrated similarities with supervised alternatives in intermediate representations, emphasizing the need for fairness considerations (??).

Fairness evaluations in SSL have been more prevalent in CV (?) and NLP, including advances in generative models (?). Discussions on SSL's impact on fairness include training without prior data curation and the effects of fine-tuning (??). However, the examination of fine-tuning considerations in SSL primarily refers to updating Batch Normalization statistics and training residual skip connections, rather than assessing the impact of supervision level. In human-centric timeseries data, fairness evaluations have seen limited exploration, mainly in a supervised setting. For instance, for in-hospital mortality using the MIMIC-III or MIMIC-IV dataset (??) or keyword spotting for on-device ML (?). All in all, such efforts are still in their early stages (?), and an exploration of fairness in SSL for timeseries data is still lacking.

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 timeseries data, is still lacking.

Method

Having established that prior SSL literature has not adequately studied the impact of the level of supervision on fairness for human-centric timeseries, we conduct systematic experiments to investigate the extent to which fine-tuning strategies in SSL (Figure 1) affect fairness and learned representations. Based on prior work on using contrastive learning in timeseries (?), we aim to answer the following research questions within the context of health timeseries tasks: How do supervision and fine-tuning levels affect fairness? What are the differences in learned representations between supervised learning and SSL?

Models. We use a SimCLR (?) variant by **?**, adapted for (health) timeseries data. Our design mirrors SimCLR's key components: a) a *stochastic data augmentation* module employing scaling and signal inversion; b) a 3-layer Convolutional Neural Network (CNN) as the *base encoder*; c) a *projection head* using a 2-layer Multi-layer Perceptron (MLP) for mapping representations; and d) a contrastive loss function, namely a normalized temperature-scaled cross-entropy loss (NT-Xent) (??), as per **?**. We define a similar architecture for the supervised baseline, replacing the contrastive loss with categorical cross-entropy.

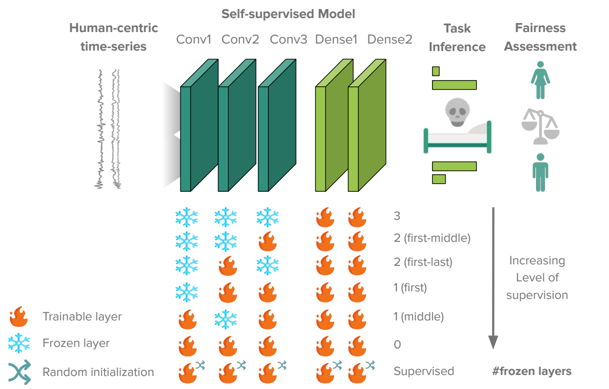


Figure 1: **Fine-tuning strategies for SSL workflows.** We experiment with various levels of supervision to optimize both performance and fairness.

Fine-tuning Setup. To assess the effect of self-supervision on fairness outcomes and representations, we employ a gradual freezing strategy, balancing the impact of the pre-trained encoder and the downstream labels (i.e., by increasing the #trainable parameters). We start by freezing all three base encoder layers and fine-tuning only the projection head. Then, gradually, we unfreeze layers one by one until both the encoder layers and projection head are fully trainable (similar to a fully supervised approach). We experiment with different freezing combinations, as seen in Figure 1.

Representation Similarity. For assessing the impact of supervision on learned representations, we utilize recent advances in comparing neural representations. Specifically, we utilize Centered Kernel Alignment (CKA) as a similarity index, addressing challenges regarding the distributed nature, potential misalignment, and high dimensionality of representations (?). Note that we also condition CKA on protected attributes to identify differences in representation similarity across demographic groups.

Evaluation

Dataset. We employ a real-world timeseries dataset, the MIMIC-III Clinical Database (?), in our evaluation. It contains over 31 million clinical events corresponding to 17 clinical variables. These events cover 42.2K intensive care unit (ICU) stays of 33.8K unique subjects who were treated between 2001 and 2012. Our task involves the prediction of in-hospital mortality from observations recorded within 48 hours of an ICU admission—a primary outcome of interest in acute care. Following the benchmark by **?**, we proceed with a total of 17,903 ICU stays and 3,236 subjects, forming 21.1K windows, each with 48 timestamps, 76 channels, and no overlap.

Protected attributes. To investigate fairness, we assess test-time performance for protected attribute groups, considering the documented variability in health timeseries based on physiological attributes (?). 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). For instance, prior work with MIMIC has revealed disparate treatment in prescribing mechanical ventilation among patient groups across ethnicity, gender, and age (**?**), and voiced general fairness concerns for Black and publicly insured ICU patients (**?**), indicating the importance of considering a diverse range of protected attributes.

Evaluation and fairness metrics. To assess the performance of our models, we employ the AUC-ROC metric and calculate 95% confidence intervals (CI). Consistent with the benchmark by **?**, our focus on per-instance accuracy for mortality prediction leads to calculating overall performance as the micro-average across all predictions, irrespective of the patient. Regarding fairness, in line with prior work in high-stakes scenarios (**?**), we adopt the Error Rate Ratio (ERR) metric, namely the ratio of error rates for unprivileged and privileged groups computed as $\frac{ER_{D=unprivileged}}{ER_{D=privileged}}$, where $ER = \frac{FP+FN}{P+N}$ (**?**). Values between 0.8 and 1.25 are considered fair, with 1.0 indicating parity.

We base our decision to use Error Rate Ratio on the worldviews it represents. From a technical perspective, Error Rate Ratio requires that the false positive and false negative rates be the same across protected user groups. Contextually, this means that given an outcome for a patient's mortality prediction, Error Rate Ratio requires that the probability of a corresponding incorrect prediction label be the same across all protected groups. More practically, this means Error Rate Ratio discourages using the protected attribute as a proxy for the outcome, promoting predictions independent of demographic factors (**?**). Conceptually, Error Rate Ratio aims to address population inequity, striving for risk predictions that avoid disproportionate harm to any group, irrespective of underlying risk distributions (**?**). In the context of healthcare, balancing risk predictions is crucial, given the potential consequences of false positives and false negatives and the ethical implications of risk prediction disparities (**?**).

Training Setup. Following **?**'s recommended architecture for contrastive learning on health timeseries, 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, 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. All hyperparameters have been fine-tuned through manual adjustment.

Results & Discussion

This section discusses the impact of supervision and fine-tuning in SSL on performance, fairness, and representations.

Table 1: **Comparison of supervised and SSL models, conditioned on protected attributes.** The Δ columns illustrate differences between a specific segment and the general population, where yellow indicates disadvantaged and green indicates advantaged segments. The most disadvantaged segment is underlined, and the most advantaged one is bolded.

		Models				
Prot. Attribute	**Segments**	*Supervised*	$\Delta_{gen.-segm.}$	*SSL (1 ● ○ ●)*	$\Delta_{gen.-segm.}$	$\mathbf{\Delta_{Sup.-SSL}}$
Gen. 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	White	0.839 (0.82-0.86)	0,000	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)	-0.070	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)	-0.010	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)	0.000	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

Impact of supervision on performance

Figure 2 presents the ROC Curve and the AUC-ROC scores for both supervised and SSL models with various levels of fine-tuning. We notice that for the MIMIC dataset, the fully supervised model performs the best in terms of AUC-ROC with a score of 0.839 (CI 0.82-0.86). 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), a mere 1% loss in overall performance.

Table 1 presents AUC-ROC scores per protected group for the supervised and the best-performing self-supervised model. We notice that the models do not perform equitably for all segments. Specifically, there exists a considerable performance gap for 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 and self-insured patients. Similarly, black patients

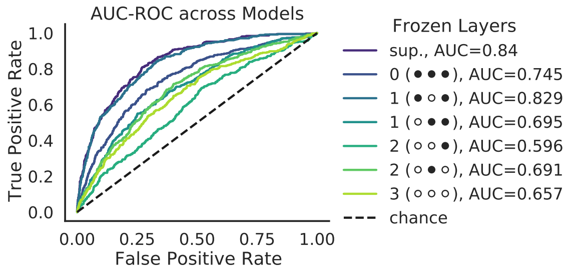
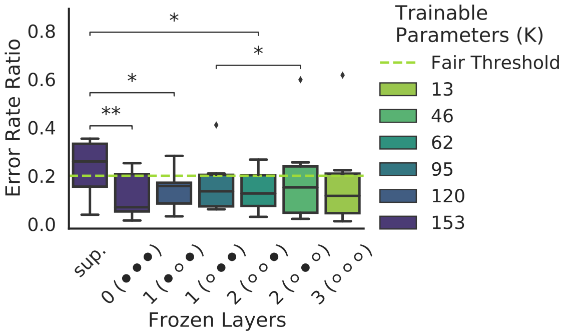


Figure 2: **AUC-ROC curves across models.** Depending on the fine-tuning strategy, SSL models achieve comparable performance to their supervised counterparts.

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 (**?**).

Overall performance discrepancies seem similar between the two models if we focus on the $\Delta_{Supervised-SSL}$, with a slight fairness benefit for the SSL model. Yet, performance metrics are not always the best indicator for 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.



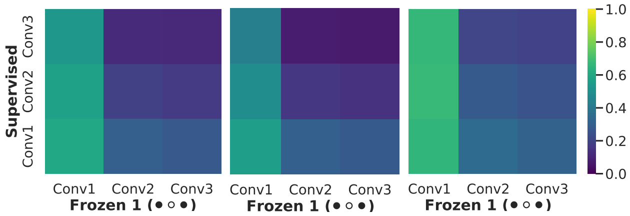
While the supervised model has superior performance, it has a greater deviation from parity (dashed line) in terms of error rate ratio compared to the best-performing SSL model (i.e., 1 ● ○ ●).

Impact of supervision and fine-tuning on fairness

Figure visualizes the deviation of each model's error rate ratio –our fairness metric of choice– from parity, i.e., $ERR = 1.0$ (dashed line). Deviations greater than 0.2 indicate bias towards a protected attribute, irrespective of priviledge. While the supervised model has superior performance, it has a statistically significant ($p < 0.05$) greater deviation from parity in terms of error rate ratio compared to the best-performing SSL model (i.e., 1 ● ○ ●). Specifically, the supervised model has a 0.125 deviation from parity, while the SSL one a 0.095–a 27% decrease. lie mostly within the acceptable limits, opposite to the supervised alternative. Note that prior work in other domains supports that model fine-tuning has an important impact on fairness (**??**). Indeed, our findings illustrate this point for timeseries data, too, as there exist differencies in error rate ratios between SSL models, even statistically significant ones (e.g., 1 ○ ● ● and 2 ○ ● ○).

Interplay of representation similarity and fairness

We compare representation similarity between the supervised and the best-performing SSL model, i.e., 1 ● ○ ●, across protected attributes through CKA. Our findings regarding the impact of supervision on representation learning for timeseries data align with prior work on CV. Specifically, **?** 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 3, 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 gender as a case in point, we notice greater similarity in learned representations for males than for females, which could partially explain the superiority of the SSL models during the fairness assessment.



(a) Random (b) Female (c) Male

Figure 3: **Representation similarity through CKA conditioned on gender between the supervised and the (best) SSL models.** The random data subset is balanced to represent both genders equally. The similarity of the SSL and supervised models is higher for male than female users.

Conclusions

supervised models' performance while demonstrating enhanced fairness, depending on fine-tuning. Specifically, moderate supervision during fine-tuning leads to self-supervised models exhibiting greater fairness across protected attributes. The impact on learned representations, measured through CKA, suggests that self-supervised models capture different data aspects, contributing to observed fairness variations. While our evaluation focuses on the large multi-year MIMIC dataset, future work should replicate these findings in other tasks beyond critical care. Nevertheless, our study underscores SSL's potential benefits in harnessing scarce labeled data in real-world, timeseries-rich application domains, such as healthcare. Ultimately, it stresses the need for assessing fairness alongside performance metrics when considering machine learning models for real-world applications.

Acknowledgments

acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. 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. Results presented in this work have been produced using the Nokia Bell Labs and the Aristotle University of Thessaloniki Compute Infrastructure and Resources.

minus deserunt inventore ex vitae possimus incidunt laborum fugit, eligendi mollitia aspernatur quos porro illo, labore omnis incidunt blanditiis neque fugit recusandae?Natus dicta accusamus sit culpa id totam, fugiat quam excepturi omnis laudantium sequi.