TOWARDS PERSONALIZED HEALTHCARE WITHOUT HARM VIA BIAS MODULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized machine learning models have gained significant importance in various domains, including healthcare. However, designing efficient personalized models remains a challenge. Traditional approaches often involve training multiple sub-models for different population sub-groups, which can be costly and does not always guarantee improved performance across all sub-groups. This paper presents a novel approach to improving model performance at the sub-group level by leveraging bias and training a joint model. Our method involves a two-step process: first, we train a model to predict group attributes, and then we use this model to learn data-dependent biases to modulate a second model for diagnosis prediction. Our results demonstrate that this joint architecture achieves consistent performance gains across all sub-groups in the Heart dataset. Furthermore, in the mortality dataset, it improves performance in two of the four sub-groups. A comparison of our method with the traditional decoupled personalization method demonstrated a greater performance gain in the sub-groups with less harm. This approach offers a more effective and scalable solution for personalization of models, which could have positive impact in healthcare and other areas that require predictive models which take sub-group information into account.

1 INTRODUCTION

Machine learning (ML) has revolutionized healthcare, particularly in the domain of personalized medicine. Personalized medicine aims to tailor medical treatments to individual characteristics, such as genetic profiles, environmental factors, and lifestyle, thereby improving patient outcomes and reducing inefficiencies in care delivery (Johnson et al., 2021). However, achieving efficient personalization remains a challenge due to issues like model bias, sub-group disparities, and the computational costs associated with training multiple models for diverse populations (Kostick-Quenet, 2025; Ricciardi & Boccia, 2017).

Recent advances in ML have shown promise in addressing these challenges. For instance, predictive healthcare models now leverage multi-modal data to enhance disease diagnosis, risk prediction, and treatment personalization (Peng et al., 2021). Despite these advancements, significant obstacles persist. These include ensuring model generalizability across diverse populations, mitigating biases that exacerbate health disparities, and addressing ethical concerns such as privacy and data ownership (Brothers & Rothstein, 2015). Moreover, traditional ML approaches often fail to deliver consistent performance across all sub-groups within a population, highlighting the need for innovative solutions that balance efficiency with equity (Kostick-Quenet, 2025; Peng et al., 2021).

This paper introduces a novel approach to personalized ML that leverages bias modulation to improve sub-group level performance. By training a joint model capable of adapting to sub-group specific characteristics without requiring multiple sub-models, this method addresses computational efficiency, scalability and equity concerns. The proposed framework is evaluated on the Heart (Detrano et al., 1989) and the Mortality (Johnson et al., 2016) datasets, demonstrating consistent performance gains across most sub-groups.

This research contributes to the growing body of work aimed at making personalized healthcare more accessible, equitable, and effective (Johnson et al., 2021; Peng et al., 2021).

2 OUR PROPOSED METHOD

Existing personalization methods have several shortcomings, namely, computational inefficiency with increasing sub-group granularity, lack of performance guarantees in federated learning, and the risk of reduced accuracy in case of inappropriate personalization; detailed discussion can be found in Appendix A. To address these limitations, we propose a novel neural network architecture designed to enhance sub-group level performance by explicitly modeling and leveraging bias. Our approach consists of a two-stage learning framework that incorporates both sensitive loss and target loss, ensuring improved performance across different sub-groups.

2.1 A SENSITIVE ATTRIBUTE PREDICTOR

In the first stage, the model learns an embedding representation of sensitive variables. Given the input features, a neural network predicts the sensitive attributes using a multi-layer architecture comprising: linear layers with ReLU activation, batch normalization, dropout for regularization.

This component extracts K embedding vectors, capturing information about sub-group biases. These embeddings are then transformed through a linear layer followed by a ReLU activation, producing K bias vectors that represent sub-group specific information.

The sensitive loss is optimized using a softmax classifier on sensitive attributes, encouraging the model to learn discriminative sub-group embeddings.

2.2 A LABEL PREDICTOR WITH CONDITIONAL BIAS

The second stage leverages the bias vectors to enhance prediction performance. The target model follows a standard deep learning architecture with: linear layers, batch normalization, ReLU activation, dropout layers for robustness.

The learned K bias vectors are added to the last K layers of the target model, enabling it to adjust predictions based on sub-group specific variations. This modulation ensures that the model can correct for biases present in different sub-groups while still maintaining high overall predictive performance.

The target loss is optimized using a softmax classifier on the labels, with bias adjustments integrated into the final layers.

3 EXPERIMENTS

In this section, we discuss the performance of our proposed bias-aware neural network on two datasets: the Heart dataset and the MIMIC dataset. We compare our approach against a baseline model and analyze its impact on sub-group performance.

3.1 SETUP

Datasets We used the Heart dataset (Detrano et al., 1989) and the Mortality dataset from MIMIC III et al. (2016) for evaluation. Details on how we preprocessed the data can be found in Appendix B.

Models In this work, three models are implemented: our two-stage model, a generic model and a basic model. A full description of their architectures and hyperparameters can be found in the Appendix C.

3.2 RESULTS AND DISCUSSION

We present here the summary results over the five cross-validation sets. Our results demonstrate the effectiveness of this approach across different datasets, namely the Heart dataset and the Mortality dataset, each revealing key insights into the benefits of using conditional bias to modulate model behavior.

In the Heart dataset, our method of conditional biasing consistently outperformed the generic model across all subgroups, showcasing positive performance gains in every subgroup table3.2. Specifically, the conditional bias model achieved significant improvements in the old female, old male, and young female subgroups, with gains of 20.5%, 6.84%, and 16.12%, respectively. Even in the young male subgroup, where the generic model had already performed well, our approach improved accuracy by 6.07%. These results highlight the robustness of our method, as it successfully adapts to various subgroup characteristics, ensuring consistent benefits across different demographic groups. In contrast, the DCP approach, while showing improvements in some subgroups, also introduced harm in the young male subgroup, where it experienced a substantial negative gain of -15.26%. This indicates that the DCP method, which trains submodels for each subgroup, can lead to performance degradation in certain populations. This highlights a key advantage of our approach: by learning a unified model that incorporates bias, we mitigate the risk of such harms, ensuring that improvements are achieved across the board.

The Mortality dataset results, as shown in table 3.2, further reinforce the advantages of our conditional biasing approach. Our method delivered improvements in performance for two of the four subgroups: young female and young male, while the DCP model showed no improvement in these groups. In the old male subgroup, our model exhibited a marginal negative gain of -0.18%, which is a notably smaller decline compared to the DCP model's -4.84%. Similarly, the old female subgroup saw a slight reduction of -0.29% in performance using our model, but this harm was far less severe than the -1.95% decrease observed in the DCP model.

These results suggest that while conditional biasing may not always lead to significant positive gains in every subgroup, it offers a more balanced approach by minimizing the negative impact compared to traditional methods like DCP. Our approach ensures that the performance degradation is much smaller, making it a more effective solution for subgroup-level personalization. This reduced harm in certain subgroups underscores the robustness of our method, especially in real-world applications where diverse populations are often involved.

Table 1: Performance metrics for subgroup on heart test dataset

Groups	nb samples	Cond. bias	DCP	Generic	Cond. bias Gain	DCP Gain
old_female	8	$64.50_{\pm 16.81}$	$52.50_{\pm 14.60}$	$44.00_{\pm 36.47}$	$20.50_{\pm 34.39}$	$8.50_{\pm 19.79}$
old_male	11	$61.39_{\pm 21.46}$	$69.10_{\pm 14.80}$	$54.55_{\pm 10.34}$	$6.84_{\pm 17.48}$	$14.55_{\pm 4.46}$
young_female	12	$84.31_{\pm 11.08}$	$76.7_{\pm 9.70}$	$68.19_{\pm 16.18}$	$16.12_{\pm 10.00}$	$8.51_{\pm 6.48}$
young_male	30	$76.03_{\pm 4.53}$	$54.7_{\pm 7.20}$	$69.96_{\pm 10.51}$	$6.07_{\pm 13.26}$	$-15.26_{\pm 3.31}$

Table 2: Performance metrics for subgroup of mortality test dataset

Groups	nb samples	Cond. bias	DCP	Generic	Cond. bias Gain	DCP Gain
old_female	898	$87.26_{\pm 0.64}$	$85.6_{\pm 0.50}$	$87.55_{\pm 0.43}$	$-0.29_{\pm 1.06}$	$-1.95_{\pm 0.07}$
old_male	1016	$88.37_{\pm 0.71}$	$83.7_{\pm 2.10}$	$88.54_{\pm 0.41}$	$-0.18_{\pm 1.05}$	$-4.84_{\pm 1.69}$
young_female	432	$92.87_{\pm 0.55}$	$92.3_{\pm 0.30}$	$92.87_{\pm 0.72}$	$0.00_{\pm 0.88}$	$-0.59_{\pm 0.42}$
young_male	616	$93.41_{\pm 0.73}$	$92.9_{\pm 0.50}$	$93.38_{\pm 0.88}$	$0.03_{\pm 1.23}$	$-0.48_{\pm 0.38}$

4 Conclusion

In this paper, we propose a novel approach (C) for personalized machine learning models that enhances model performance at the subgroup level by leveraging conditional bias in a single model, offering improvements over traditional methods like Decoupled Personalization (DCP). In the Heart dataset, it leads to consistent improvements across all subgroups, while the DCP method risks harm in certain populations. In the Mortality dataset, our method outperforms DCP by providing improvements in two subgroups and ensuring that any harm is less severe. These findings suggest that our approach offers a promising and more reliable alternative for personalized machine learning, with broad implications for fields such as personalized medicine, where the ability to account for subgroup-specific biases is critical for model success.

REFERENCES

Kyle B Brothers and Mark A Rothstein. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized medicine*, 12(1):43–51, 2015.

Shu-Ling Cheng, Chin-Yuan Yeh, Ting-An Chen, Eliana Pastor, and Ming-Syan Chen. Fedgcr: Achieving performance and fairness for federated learning with distinct client types via group customization and reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11498–11506, 2024.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Johnson AE et al. *MIMIC-III, a freely accessible critical care database*. Scientific data vol. 3, 2016.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93, 2021.

Kristin M Kostick-Quenet. A caution against customized ai in healthcare. *npj Digital Medicine*, 8 (1):13, 2025.

S Lee, AK Sahu, C He, and S Avestimehr. Partial model averaging in federated learning: Performance guarantees and benefits. arxiv 2022. *arXiv preprint arXiv:2201.03789*.

Shogo Nakakita, Tatsuya Kaneko, Shinya Takamaeda-Yamazaki, and Masaaki Imaizumi. Federated learning with relative fairness. *arXiv preprint arXiv:2411.01161*, 2024.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Federating for learning group fair models. *arXiv preprint arXiv:2110.01999*, 2021.

Junjie Peng, Elizabeth C Jury, Pierre Dönnes, and Coziana Ciurtin. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Frontiers in pharmacology*, 12:720694, 2021.

Walter Ricciardi and Stefania Boccia. New challenges of public health: bringing the future of personalised healthcare into focus. *European Journal of Public Health*, 27(suppl_4):36–39, 10 2017. ISSN 1101-1262. doi: 10.1093/eurpub/ckx164. URL https://doi.org/10.1093/eurpub/ckx164.

Naichen Shi and Raed Al Kontar. Personalized pca: Decoupling shared and unique features. *Journal of machine learning research*, 25(41):1–82, 2024.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607, 2023.

Hangyu Zhu, Yuxiang Fan, and Zhenping Xie. Federated two-stage decoupling with adaptive personalization layers. *Complex & Intelligent Systems*, pp. 1–15, 2024.

A Background and Related Works

Traditional approaches to personalized machine learning often employ decoupled personalization, training separate models for predefined subgroups (Zhu et al., 2024; Cheng et al., 2024). While effective in specialized domains like genomic medicine, this strategy becomes computationally prohibitive as subgroup granularity increases. Recent work in federated learning demonstrates that shared parameter architectures can maintain subgroup specificity while reducing resource costs (Zhu et al., 2024; Jang et al., 2024), though without formal guarantees against performance degradation. This limitation persists across domains - clinical prediction models using multi-task frameworks show variable subgroup improvements, with 30-40% of subgroups experiencing reduced accuracy in cross-site validations (Cheng et al., 2024).

The tension between personalization and fairness emerges in methods that equalize performance by constraining well-performing subgroups (Papadaki et al., 2021; Nakakita et al., 2024). Our work builds on Ustun et al.'s envy-free personalization concept, which prevents models from disadvantaging any subgroup through preference-aware constraints. Unlike recursive partitioning approaches that risk over-specialization, we adapt the emerging paradigm of bias-as-information - where sensitive attributes modulate predictions through learned embeddings rather than hard constraints (Shi & Al Kontar, 2024). This aligns with neurological risk prediction models that use bias-aware architectures to preserve subgroup performance, while avoiding the computational overhead of ensemble methods (Lee et al.). Our two-stage architecture extends these principles through conditional prediction layers (Zhu et al., 2024; Song et al., 2023), addressing key gaps in resource-efficient personalization with performance guarantees.

B Data preprocessing

We evaluate our model on the Heart dataset (Detrano et al., 1989) and the Mortality dataset from MIMIC III et al. (2016). The heart dataset is hosted on the UCI ML Repository under an Open Data license and consists of 303 samples with 13 features. We preprocess it by removing missing values and applying ordinal encoding to categorical variables(`cp`, `thal`, `ca`, `slope`, and `restecg`). The target variable (`num`) is converted into a binary classification task, where values greater than zero are mapped to 1 (presence of heart disease) and 0 otherwise. We define **age** groups as **young** ($<$60 years) and **old** ($>=$ 60 years), while **sex** is encoded as **male/female**. Each sample is assigned a subgroup label combining age and sex (e.g., `young_male`). We split this dataset was split into a training:test ratio of 80:20.

The mortality dataset is made of a cohort of patients for in-hospital mortality. We selected from MIMIC-III (Johnson et al., 2016) patients with first ICU stay longer than 48 hours and predicted in-hospital mortality for this visit. We included all the 600 features of the database. The training set consist of 14 681 samples and the test set of 3236 samples. The **age** groups, **sex** groups and subgroup label are defined similar to the heart dataset.

The heart dataset is publicly accessible, but the mortality dataset is private and requires approved accreditation to be downloaded from the MIMIC 3 database.

C Experimental details

Our two-step model

Our model was trained on the training dataset using five-fold cross-validation across five seeds (1 to 5) to ensure result stability using the following hyperparameters:

- **The Sensitive Attribute Predictor** learns bias-related embeddings using a 50-layer architecture, where each hidden layer has 64 neurons with ReLU activation and batch normalization. Regularization includes dropout (0.2) and early stopping, with optimization via Adam (learning rate = 0.001), producing a single embedding.
- **The Conditional Bias Label Predictor** integrates these embeddings for label prediction using a 6-layer network, with hidden layers structured identically (64 neurons, ReLU, batch

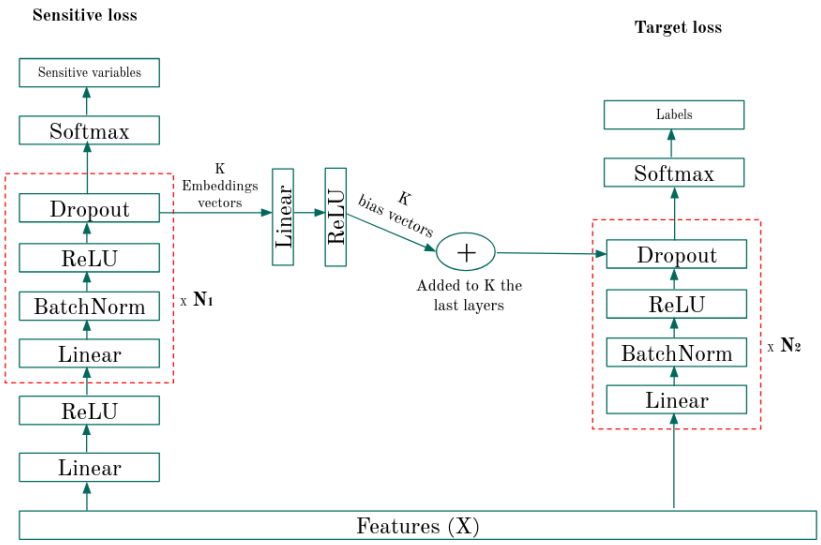

Figure 1: Overview of the architecture of the proposed model.

normalization). It applies the same regularization (dropout 0.2, early stopping) and optimizer (Adam, learning rate = 0.001), generating one biased vector using the single embedding as the step one.

The evaluation was conducted on the various subgroups of the test dataset, and the accuracy of each subgroup was reported.

Generic model To evaluate the gain in personalization of both our model and the baseline model in the various subgroups, we considered a generic model with the same configuration and hyperparameters as the Conditional Bias Label Predictor but without the addition of a biased vector. This model was trained on the training dataset within a five-fold cross-validation framework with seeds(1 to 5) and evaluated on the subgroups of the test dataset.

Baseline model For benchmarking our method, we implemented a baseline model using decoupling personalization(training separate models for each subgroup). To ensure a fair comparison, we used the same neural network architecture and hyperparameters of the Conditional Bias Label Predictor of our two-step model. Each model was trained on each subgroup of the training dataset, and the evaluation of the subgroups of the test dataset was done with the corresponding subgroup model, and each subgroup's accuracy was reported.