TOWARDS PERSONALIZED HEALTHCARE WITHOUT HARM VIA BIAS MODULATION

Frank Kwamou Ngaha[†], Patrik Kenfack[*§], Ulrich Aïvodji[*§] & Samira Ebrahimi Kahou[‡§$]
[†]University of Grenoble Alpes, [*]ÉTS Montréal, [‡]University of Calgary, [§]Mila, [$]CIFAR

ABSTRACT

Clinical prediction models are often personalized to target heterogeneous sub-groups by using demographic attributes such as race and gender to train the model. Traditional personalization approaches involve using demographic attributes in input features or training multiple sub-models for different population subgroups (decoupling model). However, these methods often *harm* the performance at the subgroup level compared to non-personalized models. This paper presents a novel personalization method to improve model performance at the sub-group level. 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 personalized models, which could have a positive impact in healthcare and other areas that require predictive models that take sub-group information into account.

1 INTRODUCTION

Machine learning (ML) has revolutionized healthcare, particularly in 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 BACKGROUND AND RELATED WORKS

Personalization involves various techniques that utilize personal data. Here, we use the term to refer specifically to methods that target groups rather than individuals. Group attributes in modern personalization approaches help enhance population-level performance by, for instance, integrating higher-order interaction effects (Bien et al. (2013)) or recursively partitioning data (Elmachtoub et al. (2021)). However, research rarely quantifies the benefits of personalization, and when it does, the focus is usually on population-level improvements rather than the specific groups providing personal data

Traditional approaches to personalized machine learning often employ decoupling personalization, training separate models for predefined sub-groups (Zhu et al., 2024; Cheng et al., 2024). While effective in specialized domains like genomic medicine, this strategy becomes computationally prohibitive as sub-group granularity increases. A recent work in federated learning demonstrates that shared parameter architectures can maintain sub-group specificity while reducing resource costs (Zhu et al., 2024; Jang et al., 2024), though without formal guarantees against performance degradation.

Some personalization methods achieve subgroup-level performance equalization by imposing constraints on well-performing subgroups (Papadaki et al., 2021; Nakakita et al., 2024). However, this approach highlights the inherent tension between personalization and fairness.

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 sub-group 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.

3 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 (Suriyakumar et al., 2023); detailed discussion can be found in Section 2. To address these limitations, we propose a novel two-stage learning framework in which, in the first step, we train a group attribute predictor to obtain a bias embedding vector. In the second step, we utilize the bias embedding vector to modulate the training of the label predictor. Figure 3 shows an overview of the proposed architecture.

3.1 THE GROUP ATTRIBUTE PREDICTOR

In the first stage, we train an embedding representation of the group attributes using the input features. The embedding model uses a multi-layer perception architecture comprising linear layers with ReLU activation, batch normalization, and dropout for regularization. This the attribute predictor 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 group-specific information. The sensitive loss is optimized using a cross-entropy loss on sensitive attributes, encouraging the model to learn discriminative sub-group embeddings.

3.2 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, and 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 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 cross-entropy loss classifier on the labels, with bias adjustments integrated into the final layers.

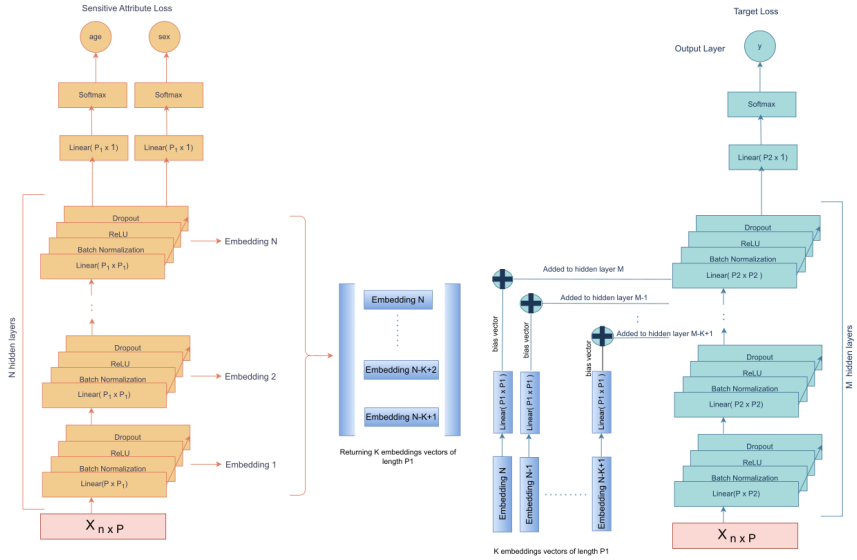

Figure 1: Overview of the architecture of the proposed model.

4 Experiments

4.1 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 A. On both datasets, we use *gender* ({`male`, `female`}) and *age* ({`old`, `young`}) as group attributes (Suriyakumar et al., 2023) and subgroups are defined based on their intersection, i.e., {`male`, `female`} × {`old`, `young`}.

Models We compare our personalized model (Cond. bias) to the Decoupled Personalization (DCP) model, which trains a different model for each sub-group (Suriyakumar et al., 2023; Ustun et al.). We measure the population level performance gain against the *generic model* trained without group attributes. A full description of model architectures and hyperparameters can be found in Appendix B. We evaluate performance in term accuracy and measure personalization gains in terms subgroup level accuracy improvement between the generic and the personalized model.

4.2 Results and Discussion

Tables 4.2 and 4.2 summarize the main experimental results over a five-fold cross-validation on the Heart the Mortality datasets, respectively. These results demonstrate the effectiveness of the proposed approach in reducing harm and improving performance across subgroups.

Table 1: Performance metrics for subgroup on heart test dataset

Groups	**nb samples**	**Generic**	**DCP**		**Cond. bias (OURS)**	
		Accuracy	Accuracy	Gain	Accuracy	Gain
`old,female`	8	$44.00_{\pm 36.47}$	$52.50_{\pm 14.60}$	$8.50_{\pm 19.79}$	$64.50_{\pm 16.81}$	$20.50_{\pm 34.39}$
`old,male`	11	$54.55_{\pm 10.34}$	$69.10_{\pm 14.80}$	$14.55_{\pm 4.46}$	$61.39_{\pm 21.46}$	$6.84_{\pm 17.48}$
`young,female`	12	$68.19_{\pm 16.18}$	$76.70_{\pm 9.70}$	$8.51_{\pm 6.48}$	$84.31_{\pm 11.08}$	$16.12_{\pm 10.00}$
`young,male`	30	$69.96_{\pm 10.51}$	$54.70_{\pm 7.20}$	$-15.26_{\pm 3.31}$	$76.03_{\pm 4.53}$	$6.07_{\pm 13.26}$

Table 2: Performance metrics for subgroup of mortality test dataset

Groups	nb samples	Generic	DCP		Cond. bias (OURS)	
		Accuracy	Accuracy	Gain	Accuracy	Gain
old, female	898	$87.55_{\pm 0.43}$	$85.60_{\pm 0.50}$	$-1.95_{\pm 0.07}$	$87.26_{\pm 0.64}$	$-0.29_{\pm 1.06}$
old, male	1016	$88.54_{\pm 0.41}$	$83.70_{\pm 2.10}$	$-4.84_{\pm 1.69}$	$88.37_{\pm 0.71}$	$-0.18_{\pm 1.05}$
young, female	432	$92.87_{\pm 0.72}$	$92.30_{\pm 0.30}$	$-0.59_{\pm 0.42}$	$92.87_{\pm 0.55}$	$0.00_{\pm 0.88}$
young, male	616	$93.38_{\pm 0.88}$	$92.90_{\pm 0.50}$	$-0.48_{\pm 0.38}$	$93.41_{\pm 0.73}$	$0.03_{\pm 1.23}$

In the Heart dataset (Table 4.2), our personalization method consistently outperformed the *generic* and decoupled (DCP) model across all sub-groups, providing a positive accuracy gain for all subgroups. More specifically, our method provides an accuracy gain of **20.5**% for the worst-performing subgroup (old, female) from a previous accuracy of less than 44% on the generic model. We also observe that the best-performing subgroup (young, male) in the generic model experiences an accuracy gain of **6.07**% from a previous accuracy of 69.96%. On the other hand, the DCP model, while improving the worst-performing subgroup by 8.5%, harms the best-performing subgroup with a decrease in accuracy of -15.26%. This shows that the decoupled model can harm certain subgroups, thereby hindering the expected benefits of model personalization for all subgroups (Ustun et al.). These results demonstrate 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 all subgroups.

The results on the mortality dataset are, however, more nuanced in terms of performance gains (Table 4.2). Our method delivered accuracy gain for two of the four sub-groups: (young, female) and (young, male), while the DCP model showed accuracy decrease in all sub-groups.

On this dataset, most subgroups have sufficiently good performance, with the worst-performing subgroup (old, female) having 87.55% accuracy. The results align with recent studies suggesting personalization can not always benefit all subgroups (Suriyakumar et al., 2023), yet, our method minimizes the harm compared to the DCP model. More specifically, in the worst case, the DCP model incurs a -4.84% performance drop on (old, male) subgroup while our method only reduces it by -0.18%. To avoid harm, the generic model should be used to make predictions for the subgroups that do not benefit from personalization.

In sum, these results suggest that while our conditional biasing may not always lead to positive subgroup-level accuracy gain in all tasks, it offers a more balanced approach by minimizing the negative impact compared to traditional methods like DCP. In challenging tasks, our approach can ensure that the performance degradation is much smaller, making it a more effective solution for subgroup-level personalization. This reduced harm in certain sub-groups underscores the robustness of our method, especially in real-world applications where diverse populations are often involved. Investigating the failure cases of personalization remains an open and under-explored research direction.

5 Conclusion

In this paper, we propose a novel approach for personalized machine learning models that enhances model performance at the sub-group level by leveraging a conditional bias model, offering improvements over traditional methods like Decoupled Personalization (DCP). In the Heart dataset, our proposal provides consistent improvements across all sub-groups, while the DCP method harms certain demographic groups. In the Mortality dataset, our method outperforms DCP by improving two subgroups and minimizing harm in the two other groups. 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 group-specific biases is critical for model success.

ACKNOWLEDGEMENT

The authors thank the Digital Research Alliance of Canada and Denvr for computing resources. SEK is supported by CIFAR and NSERC DG (2021-4086) and UA by NSERC DG (2022-04006).

REFERENCES

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

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.

Adam N Elmachtoub, Vishal Gupta, and Michael L Hamilton. The value of personalized pricing. *Management Science*, 67(10):6055–6070, 2021.

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.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms performance: reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, pp. 33209–33228. PMLR, 2023.

Berk Ustun, Yang Liu, and David C Parkes. Fairness without Harm:Decoupled Classifiers with Preference Guarantees.

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

A 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 sub-group label combining age and sex (e.g., `young_male`). We split this dataset 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 consists of 14 681 samples and the test set of 3236 samples. The **age** groups, **sex** groups, and sub-group labels are defined similarly 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.

B 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 Group 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 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 sub-groups of the test dataset, and the accuracy of each sub-group was reported.

Generic model. To evaluate the gain in personalization of both our model and the baseline model in the various sub-groups, 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 sub-groups of the test dataset.

Decoupling (DCP) Model. We implemented the decoupling (DPC) method as a baseline for comparison. DCP implements personalization by training a separate model for each sub-group. More specifically, we train each model using only the data of the corresponding subgroup (Ustun et al.; Suriyakumar et al., 2023). At test time, for a given data point from subgroup a g, the model corresponding model is used for predictions. To ensure a fair comparison, we used the same neural network architecture and hyperparameters of the label predictor in our model.